
Graph Fission and Cross-Validation

James Leiner
Carnegie Mellon University
jleiner@stat.cmu.edu

Aaditya Ramdas
Carnegie Mellon University
aramdas@stat.cmu.edu

Abstract

We introduce a technique called graph fission which takes in a graph which potentially contains only one observation per node (whose distribution lies in a known class) and produces two (or more) independent graphs with the same node/edge set in a way that splits the original graph’s information amongst them in any desired proportion. Our proposal builds on data fission/thinning, a method that uses external randomization to create independent copies of an unstructured dataset. We extend this idea to the graph setting where there may be latent structure between observations. We demonstrate the utility of this framework via two applications: inference after structural trend estimation on graphs and a model selection procedure we term “graph cross-validation”.

1 INTRODUCTION

Sample splitting, where an analyst divides a portion of data to train a model and the remaining portion of data to validate it is a ubiquitously used tool by statisticians. Unfortunately, this approach is typically appropriate only in settings with repeated *i.i.d.* observations, with few exceptions. In cases where the available data is dependent or not identically distributed, for instance time series or fixed-design regression, sample splitting is generally not practical or easily interpretable. Nonetheless, sample splitting strategies are often still used due to a lack of alternatives. This often leads to an analysis pipeline that is heuristic and does not come with theoretical guarantees.

Building on an idea for post-selection inference in

linear models (Tian and Taylor, 2018; Rasines and Young, 2023), recent work by Leiner et al. (2023) proposed a general solution to this problem through use of external randomization to create synthetic copies of the data that are *i.i.d.* by construction and each of which contain a portion of the information contained in the original dataset. Such a construction does not always work: it is only feasible if the distribution of the data lies in certain known classes. While Leiner et al. (2023) applied to varied contexts like estimation after multiple testing and post-selection inference for trend filtering, follow-up work (Neufeld et al., 2022, 2023; Dharamshi et al., 2023) has provided improved constructions in several settings and applied it to new settings like latent variable estimation.

In this paper, we will apply these procedures to the graph setting and demonstrate their efficacy through two practical applications: cross-validation for graph-valued data and inference after structural trend estimation over graphs. Graph-valued data presents unique challenges because in some settings, only a single graph is observed and there is a learning or inference problem on the graph, where one believes that the graph structure captures the structure in either the signal or the noise (i.e. either the signal is smooth over the graph, or the graph captures the correlation structure of the noise). By extending the aforementioned “fissioning” idea to graphs, an analyst can gain access to techniques that are not generally available in settings without repeated *i.i.d.* observations, including cross-validation to estimate out-of-sample risk and the creation of independent training, validation and test graphs (or selection and inference graphs) so that the analyst may explore different modeling techniques without violating error control on downstream inferential procedures via “double dipping”.

Contributions. Our main contributions are:

- When the distribution of errors is known and falls into the “convolution-closed” class (Joe, 1996), we demonstrate how data fissioning (Leiner et al., 2023) and thinning (Neufeld et al., 2023) methods

can enable cross-validation on graphs for hyperparameter tuning. We demonstrate the utility of this procedure on trend estimation problems, but note that it applies generically to procedures that requires hyperparameter tuning over graphs, such as the training of graph neural networks.

- We extend the results of [Leiner et al. \(2023\)](#); [Neufeld et al. \(2023\)](#) to enable inference for structural trend estimation on graphs in the Gaussian setting, in the presence of increasing dimensions and unknown error variance. Existing results all rely heavily on known specification of errors or a low-dimensional setting that ensures consistent estimators of the error variance are available. We provide a result ([Theorem 1](#)) that enables graph fission in the presence of unknown error variance in the Gaussian case. This allows for the application of these techniques in real-world settings where the variance is not known a priori.

Paper outline. In [Section 2](#), we review decompositions for fissioning a single dataset into m independent copies and provide examples of how this method can be used in the context of graph-valued data. We also discuss common methods for estimating a structural trend over a graph. In [Section 3](#), we build on these techniques by introducing graph cross-validation, which we then use to tune hyperparameters used in trend estimation. In [Section 4](#), we demonstrate a second application of graph fission: the creation of valid confidence intervals after model selection on a graph. In [Section 5](#), we apply these methods to real data by constructing confidence intervals on taxicab usage in New York City. We conclude in [Section 6](#).

2 METHODOLOGY

Let $\mathcal{G} = (V, E, Y)$ be a graph with a known vertex (V) and edge (E) set and a set of observations $Y = (y_1, \dots, y_n) \in \mathbb{R}^n$ over the vertices. We use a standard nonparametric regression framework, where

$$y_i = \mu_i + \epsilon_i,$$

$\mu_i = E[y_i]$, and ϵ_i is a 0 mean random variable. Denote $\mu = (\mu_1, \dots, \mu_n)$ and $\epsilon = (\epsilon_1, \dots, \epsilon_n)$.

2.1 Decomposition Rules

We aim to create m new synthetic copies of \mathcal{G} , which we denote as $\mathcal{G}_1, \dots, \mathcal{G}_m$ with corresponding observations labeled as $Y^{\mathcal{G}_1}, \dots, Y^{\mathcal{G}_m}$. We require the synthetic graphs to have the following properties:

1. \mathcal{G}_i has the same non-random structure (i.e. V and

E). Furthermore, $\mathbb{E}[Y^{\mathcal{G}_j}] = h_1(\mu)$ for all $j \in [m]$ and some known deterministic function h_1 .

2. Taken together, the individual datasets recover the original data Y in the sense that there exists a known deterministic function h_2 such that $\mathcal{G} = h_2(\mathcal{G}_1, \dots, \mathcal{G}_m)$, and you can not recover \mathcal{G} from any strict subset of graphs $\mathcal{G}_1, \dots, \mathcal{G}_m$
3. The information contained in Y is divided across $\mathcal{G}_1, \dots, \mathcal{G}_m$ in any proportion desired.

Remark 1. *At this stage, we also note the existence of an alternative less stringent set of requirements. Following the terminology of [Leiner et al. \(2023\)](#), we call this the **P2** regime. Here, we only require the creation of two synthetic copies of these graphs $\mathcal{G}_1, \mathcal{G}_2$. The properties that they must fulfill are:*

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

Although we will not focus on this idea throughout much of the paper, we will revisit it in [Section 4](#), as it is a key ingredient for [Theorem 1](#).

Following the terminology of “data fission” ([Leiner et al., 2023](#)), we call the above task “graph fission”. Instead of using the preceding paper’s techniques, we instead employ the decomposition rules of [Neufeld et al. \(2023\)](#) which provide an algorithm for splitting a random variable into m independent copies when the distribution of that variable is *convolution-closed*.

Definition 1 ([Joe \(1996\)](#)). *Let F_θ be a distribution indexed by a parameter θ in parameter space Θ . Drawing $X' \sim F_{\theta_1}$ and $X'' \sim F_{\theta_2}$ independently, if $X' + X'' \sim F_{\theta_1 + \theta_2}$ whenever $\theta_1 + \theta_2 \in \Theta$ then F_θ is convolution-closed in the parameter θ .*

Note that many distributions encountered in data analysis including Gaussian, Poisson, binomial, and negative binomial are convolution-closed.

A property of many convolution-closed distributions is that when X_1, \dots, X_m are each drawn from the same family of distributions, then the joint density of $(X_1, \dots, X_m) | \sum_{i=1}^m X_i = x$ is tractable. Hence, m synthetic samples can then be generated from a single sample x by drawing from this joint distribution, conditioning on their total sum adding up to x . This intuition is formalized below.

Fact 1 ([Theorem 2 of Neufeld et al. \(2023\)](#)). *Assume $X \sim F_\theta$ for a convolution-closed distribution in parameter $\theta \in \Theta$. Choose τ_1, \dots, τ_m such that $\sum_{j=1}^m \tau_j = 1$ and $\tau_j \theta \in \Theta$. Let $G_{\theta_1, \dots, \theta_m}$ be the joint distribution of $(X_1, \dots, X_m) | \sum_{i=1}^m X_i = X$ when X_i are drawn independently from F_{θ_i} . If $X_1, \dots, X_m \sim G_{\tau_1 \theta, \dots, \tau_m \theta}$, then*

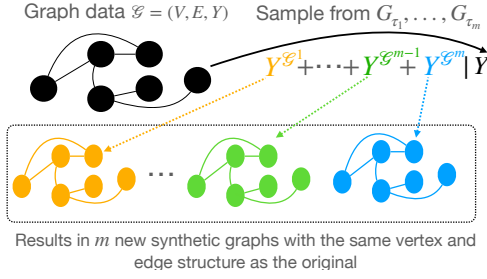


Figure 1: Graphical illustration of Fact 1.

the following holds: (i) $X_i \sim F_{\tau_i, \theta}$; (ii) X_1, \dots, X_m are mutually independent; (iii) $\sum_{i=1}^m X_i = X$; (iv) whenever F_θ has finite first moment, $\mathbb{E}[X_j] = \theta_j \mathbb{E}[X]$.

Fact 1 can be applied whenever the joint distribution of Y is convolution-closed (a special case of this being when y_i are individually convolution-closed and mutually independent), and the analyst is able to draw from the distribution $G_{\tau_1, \dots, \tau_m}$. See Figure 1 for an illustration of this procedure. As a first example, we will use Fact 1 to fission a graph with Gaussian errors.

Example 1 (Gaussian graph). Assume $y_i \sim N(\mu_i, \sigma^2)$. We draw $y_i^{\mathcal{G}^1}, \dots, y_i^{\mathcal{G}^m}$ from the distribution

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

Marginally, $y_i^{\mathcal{G}^j} \sim N(\mu_i, m\sigma^2)$, $j \in [m]$, all mutually independent. Note that this procedure splits the information evenly across all graphs, because the Fisher information $\mathcal{I}_{\mathcal{G}^j}(\mu_i) = \frac{1}{m\sigma^2}$ for all j .

Example 2 (Gaussian graph with correlated errors). The preceding example can be generalized to the correlated Gaussian case, albeit with a more complicated decomposition strategy. Assume $Y \sim N(\mu, \Sigma)$, where Σ is any (known) covariance matrix. Then draw $Y^{\mathcal{G}^1} | Y \sim N(Y, (m-1)\Sigma)$ and $Y^{\mathcal{G}^{-1}} := \frac{m}{m-1}Y - \frac{1}{m-1}Y^{(1)}$. Marginally, $Y^{\mathcal{G}^1} \sim N(\mu, m\Sigma)$ and $Y^{\mathcal{G}^{-1}} \sim N(\mu, \frac{m}{m-1}\Sigma)$. Continuing to draw $Y^{\mathcal{G}^{j+1}} | Y^{\mathcal{G}^{-j}} \sim N(Y^{\mathcal{G}^{-j}}, (m-j-1)\frac{m}{m-j}\Sigma)$ with $Y^{\mathcal{G}^{-j-1}} := \frac{m}{m-j}Y^{\mathcal{G}^{-j}} - \frac{1}{m-j}Y^{\mathcal{G}^{j+1}}$ and proceeding m times will result in $Y^{\mathcal{G}^j} \stackrel{i.i.d.}{\sim} N(\mu, m\Sigma)$.

One drawback is that the covariance matrix must be known prior to fissioning. We defer discussion of the case of unknown σ to Section 4. Not all distributions will have this issue—eg: Poisson errors can be fissioned without needing to estimate an unknown parameter.

Example 3 (Poisson graph). Assume $y_i \sim$

$\text{Pois}(\mu_i)$. We can draw a new vector $y_i^{\mathcal{G}^1}, \dots, y_i^{\mathcal{G}^m}$ as $\text{Multinomial}(y_i, (\frac{1}{m}, \dots, \frac{1}{m}))$. Then each component $y_i^{\mathcal{G}^j} \sim \text{Pois}(\frac{\mu_i}{m})$. We again note that this divides the Fisher information evenly across each graph.

Note that in the preceding examples, we index over both the nodes ($i \in [n]$) and synthetic samples at each node ($j \in [m]$). For clarity, we keep this notation consistent throughout the paper, with j always indexing over synthetic samples and i indexing over nodes.

Both of these examples, as well as several more, can be found in Leiner et al. (2023); Neufeld et al. (2023), but the latter paper has a more unified treatment.

2.2 Structural Trend Estimation on Graphs

As a unifying example to work with across the paper, we consider estimating a structural trend over μ by solving an optimization problem of the form,

$$\hat{\beta} := \underset{\beta \in \mathbb{R}^n}{\text{argmin}} \underbrace{\ell(Y, \beta)}_{\text{Loss}} + \underbrace{D(\beta)}_{\text{Penalty}}. \quad (1)$$

The loss can be any convex function, but we will focus on square loss $\ell(Y, \beta) := \frac{1}{n} \|Y - \beta\|_2^2$ for continuous-valued graph data and Poisson loss $\ell(Y, \beta) := \frac{1}{n} \sum_{i=1}^n (-y_i \beta_i + \exp(\beta_i))$ for count data.

In order to estimate a smooth trend that aligns with the graph structure given by the vertex and edge set, most approaches will add a regularization term that encourages smoothing over adjacent nodes (Kondor and Lafferty, 2002; Sharpnack et al., 2013; Wang et al., 2014). Graph trend filtering (Wang et al., 2014) accomplishes this by introducing a graph difference operator defined as $\Delta^{(1)} \in \{-1, 0, 1\}^{n \times p}$, which contain a single row for each of the p edges in the graph. The row corresponding to a particular edge $e_v = (i, j)$ is

$$\Delta_v^{(1)} = (0, \dots, \underset{\uparrow}{-1}, \dots, \underset{\uparrow}{1}, \dots, 0),$$

where the position of 1 and -1 is arbitrary. Setting the penalty $D(\beta) := \lambda \|\Delta^{(1)}\beta\|_1$ penalizes any differences between values of $\hat{\beta}$ at adjacent nodes, leading to a piecewise constant solution across connected components. Recursively applying this operator yields,

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

where L denotes the graph Laplacian. The corresponding penalty term $\|\Delta^{(k+1)}\beta\|_1$ penalizes higher order graphs differences in an analogous fashion. For instance, $k = 1$ enforces a piecewise linear structure across connected components, $k = 2$ enforces a piece-

wise quadratic structure, and so on.

Other ways of constructing the penalty term include using an L_2 penalty term (i.e. $D(\beta) := \lambda \|\Delta^{(k+1)}\beta\|_2^2$) instead of L_1 , which is equivalent to graph Laplacian smoothing (Smola and Kondor, 2003), and combining the two to create an elastic net penalty term by setting $D(\beta) := \lambda_1 \|\Delta^{(k+1)}\beta\|_1 + \lambda_2 \|\Delta^{(k+1)}\beta\|_2^2$.

In the case where L_2 regularization is used alongside square loss, the solution can be computed in closed form as $\hat{\beta} = \left(I + n\lambda (\Delta^{(k+1)})^T \Delta^{(k+1)}\right)^{-1} Y$. When L_1 regularization is used alongside square loss, Arnold and Tibshirani (2016) present efficient algorithms to compute $\hat{\beta}$ along any solution path of λ which are currently implemented in the R package GENLASSO.

The solutions corresponding to other convex loss functions can be computed through standard optimization techniques such as gradient descent, though designing algorithms that can compute these solution paths most efficiently is an open area of investigation.

As an illustration, we generate data on a 10×10 grid. Nodes at Manhattan distance 1 share an edge, resulting in a graph with 100 nodes and 180 edges. We then estimate the structural trend using square loss and a variety of different penalties in Figure 2. In the case of L_1 penalties, the fitted solutions become piecewise polynomials. For the L_2 and elastic net penalties, fewer components are chosen to be exactly 0 so the structural trend tends to pick up more local variation.

3 GRAPH CROSS-VALIDATION

Across all methods for estimating the structural trends over graphs, the training of hyperparameters is a consistent challenge. When *i.i.d.* data is observed, for instance in random-design linear regression, the most common method for tuning λ is via cross-validation which does not have a direct analog in the graph setting where data is not identically distributed. Some existing methods for performing cross-validation on structured data (Ghosh et al., 2020; Celeux and Durand, 2008) can be applied to the graph setting, but they rely on an assumption that the structural trend does not vary substantially within the held-out set which may not be reasonable for datasets with substantial variation across nodes. We will not need any such assumption, as discussed below.

Another alternative to cross-validation is to choose λ is to minimize Stein’s Unbiased Risk Estimate (SURE) (Stein, 1981), but this approach comes with its own set of drawbacks. SURE is an unbiased estimate of the risk only in the case of Gaussian data (though data fission was recently used to extend SURE to the Pois-

son case (Oliveira et al., 2022)). Moreover, it requires knowledge of the effective number of degrees of freedom of the estimator which may not be readily available in all instances. Lastly, although SURE and cross-validation are asymptotically equivalent when the data is Gaussian, cross-validation often has superior finite sample performance.

Motivated by these concerns, we apply the methodology of Section 2.1 to construct an analog of cross-validation in the graph setting.

Assumption 1. *Assume that the distribution of Y is convolution-closed and follows distribution F_θ .*

Under Assumption 1, we are able to generate m new independent graphs $\mathcal{G}_1, \dots, \mathcal{G}_m$ such that $\mathbb{E}(Y^{\mathcal{G}_j}) = \frac{1}{m}\mu$ and $Y^{\mathcal{G}_j} \sim F_{\theta}^{\frac{1}{m}}$ for all $j \in [m]$ using Fact 1. To perform cross-validation, we average $m-1$ of these graphs together and leave the remaining held out graph for testing. Denote $Y^{\mathcal{G}_{-j}} := \sum_{j \neq i} Y^{\mathcal{G}_i}$. By construction,

$$Y^{\mathcal{G}_{-j}} \sim F_{\theta}^{\frac{m-1}{m}} \text{ and } Y^{\mathcal{G}_j} \sim F_{\theta}^{\frac{1}{m}}.$$

We can therefore use \mathcal{G}_{-j} to estimate θ and evaluate this estimate using the held out graph \mathcal{G}_j to get an unbiased estimate of the risk of the procedure. Repeating this process m times for each $j \in [m]$ mimics the process of m -fold cross-validation.

Remark 2. *If F is a location-scale distribution, we can rescale $Y^{\mathcal{G}_{-j}}$ and $Y^{\mathcal{G}_j}$ so their distributions are functions of the same parameter (e.g., Example 1). Otherwise, the analyst will need to be careful to scale $\hat{\theta}$ appropriately when out-of-the-box estimation procedures are used so evaluation on the test set is comparable with the training data.*

Gaussian Data with Unknown Variance. If $y_i \sim N(\mu_i, \sigma^2)$ and σ^2 is known, then Fact 1 can be applied directly. When σ^2 is unknown, it will need to be estimated. Estimating the error variance in high dimensional regression problems is difficult. Most estimators are only provably consistent under assumptions that are not verifiable in practice (Fan et al., 2011) and rely on techniques like cross-validation (Reid et al., 2016) which we do not have access to. Nonetheless, a heuristic that works empirically is to estimate

$$\hat{\sigma}^2 := \frac{1}{n - \text{df}(\hat{\beta}_\lambda)} \left\| Y - \hat{\beta}_\lambda \right\|_2^2, \quad (2)$$

where $\hat{\beta}_\lambda$ is fit using a fixed λ lasso penalty. Since λ is typically chosen either through cross-validation or by minimizing SURE (which pre-supposes knowledge of σ^2), we unfortunately have to use a pre-determined version of λ to make this selection. We choose $\lambda =$

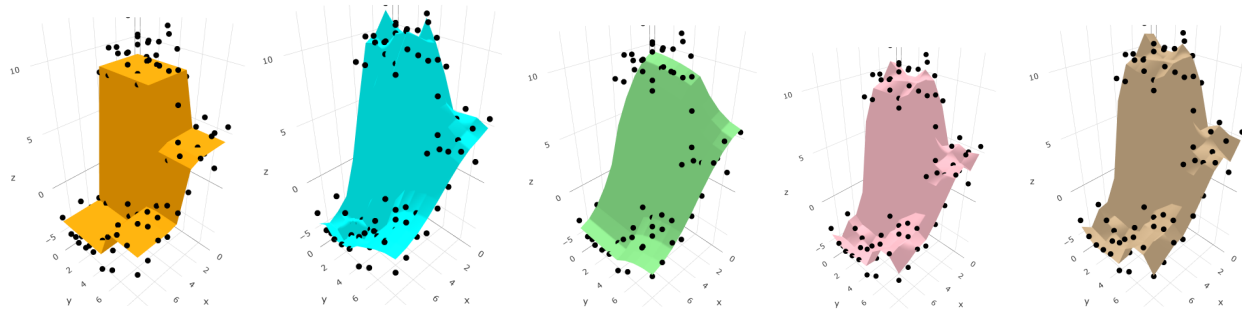


Figure 2: Example of synthetic data points and corresponding structural trend solution $\hat{\beta}$ when fit using square loss and a variety of penalties. From left to right: piecewise constant ($\|\Delta^{(1)}\beta\|_1$), linear ($\|\Delta^{(2)}\beta\|_1$), quadratic ($\|\Delta^{(3)}\beta\|_1$), ridge ($\|\Delta^{(1)}\beta\|_2^2$), and elastic net ($\alpha\|\Delta^{(1)}\beta\|_1 + (1-\alpha)\|\Delta^{(1)}\beta\|_2^2$).

$\sqrt{\frac{\log n}{n}}$ since many results require that λ grow at this rate to ensure consistency (Yu and Bien, 2017).

Simulation. We apply this method to the problem of estimating an optimal λ for structural smoothing problems as described in Section 2. We again use a node set aligned in a grid such that each vertex $v \in [1, 10] \times [1, 10]$ where nodes at Manhattan distance 1 are connected. A ground truth trend μ is constructed by randomly choosing a percentage of nodes to be active nodes that permit a structural change relative to adjacent nodes. We then generate piecewise polynomials over the connected components of the inactive set of nodes and draw observations as $y_i \sim N(\mu_i, 1)$.

The structural trend $\hat{\beta}$ is estimated using square loss and a penalty term as described in Section 2.2. When using graph fission to select λ , we consider a rule which picks the λ that minimizes the average test error (over each of the m folds) as well as the so called “one-standard error” rule which picks the largest value of λ (i.e. the most parsimonious model) which falls within one standard deviation of the minimum error.

As a point of comparison, we compare graph cross-validation with “ordinary” cross-validation in the form of structured cross-validation proposals (Ghosh et al., 2020; Celeux and Durand, 2008), adjusted to the graph setting. These amount to selecting a subset of nodes $\mathcal{I} \subseteq V$ to hold out for evaluation. We then estimate a structural trend $\hat{\beta}_{-\mathcal{I}}$ by ignoring the data in \mathcal{I} , and define the trend $\hat{\beta}_{\mathcal{I}}$ in the holdout set to be the average of adjacent nodes in the training set for each point. This amounts to assuming that all the nodes in the held-out set are inactive so the structural trend can be interpolated from the training set. This approximation makes the most sense when there is minimal structural change within \mathcal{I} , which corresponds to a smooth trend over the graph with few breakpoints.

To illustrate this disadvantage, we vary two parameters during simulation. The first is the percentage of nodes that are allowed to be active. The second is the overall size of discontinuities that are allowed at each of the active nodes. Results are shown in Figure 3. When the structural trend is smooth (i.e. fewer active nodes and smaller jump sizes), there is little to no difference between the approaches, but trends that are volatile benefit significantly from using graph fission.

In the Appendix, we repeat this experiment with misspecified errors (drawn from t , skewed normal, and Laplace distributions), and the results are nearly identical compared to the correctly specified case. These results suggest that a central limit theorem in the vein of Austern and Zhou (2020) may hold for graph cross-validation, but we leave theoretical guarantees an area of future investigation.

4 INFERENCE AFTER TREND ESTIMATION

After selecting λ via graph cross-validation, an analyst may wish to perform inference on the structural trend in addition to having a single point estimate. Unfortunately, even for fixed λ , out-of-the-box inferential procedures are not available because the active set of variables are chosen adaptively based on the data.

Leiner et al. (2023) provides a method for confidence interval construction in the case of univariate trend filtering, and when Assumption 1 holds, this general framework applies here with slight modifications. To summarize, an analyst can apply Fact 1 to generate two synthetic graphs \mathcal{G}^{sel} and \mathcal{G}^{inf} . Then, $Y^{\mathcal{G}^{\text{sel}}}$ can be used to select a *basis* that the analyst constrains $\hat{\beta}$ to fall into the span of. $Y^{\mathcal{G}^{\text{inf}}}$ is then used to calculate $\hat{\beta}$ by projecting the held-out graph onto the chosen basis. Since choosing a basis was done independently

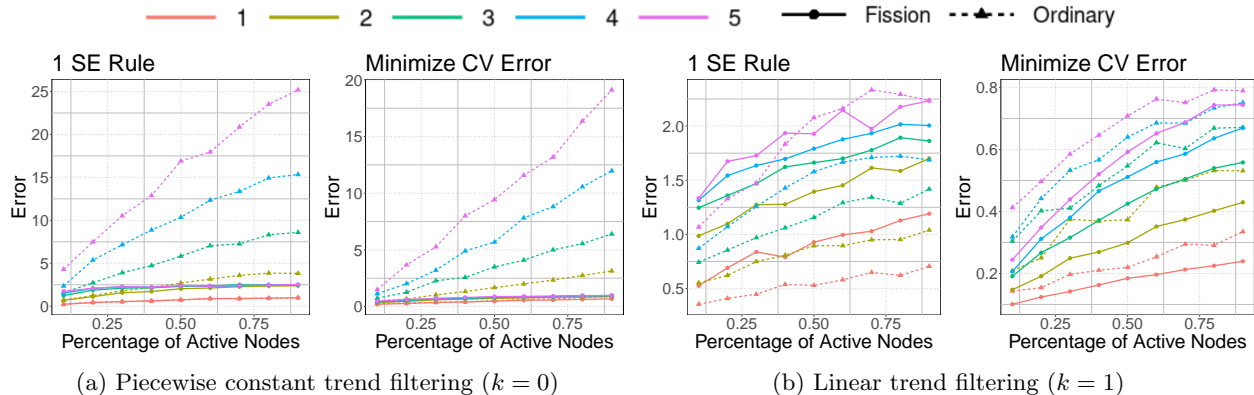


Figure 3: We vary the size of jumps at breakpoints (colors) along with the percentage of active nodes in the graph, and compare graph cross-validation against ordinary cross-validation (with L_1 penalty, $\sigma^2 = 1$, 10 folds). The relative performance of graph cross-validation (dotted) compared to 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 .

of the projection step, standard methods for producing confidence intervals will have proper coverage.

Unique complications arise in the graph setting when nuisance parameters need to be estimated from the data prior to applying [Fact 1](#). Focusing on the Gaussian setting, prior work ([Leiner et al., 2023](#); [Neufeld et al., 2023](#)) only provides inference procedures with rigorous guarantees when the error variance is known a priori or when the dimension of the problem is fixed. Since nodes and edges come online in tandem, the dimension of the penalty matrix will increase with n , invalidating the validity of these procedures. Some asymptotic guarantees in a high dimensional setting are explored in [Rasines and Young \(2023\)](#), but this work still relies on having access to consistent estimates of the error variance which is an assumption that tends not to hold in practice.

We first discuss how the framework of [Leiner et al. \(2023\)](#) can be applied to this problem for the case of known nuisance parameters in [Section 4.1](#). We then extend this methodology to the Gaussian case when the error variance needs to be estimated in [Section 4.2](#).

4.1 Inference Under [Assumption 1](#)

Let $Y \sim P_\theta$ for some convolution closed distribution in parameter space Θ with corresponding density function $p(y, \theta)$. At the selection stage, we recommend using L_1 penalties to select a model because the basis structure of these estimates is well understood — see [Algorithm 1](#) for an explicit formula for extracting a basis from a trend fit using an L_1 penalty. That said, because \mathcal{G}^{sel} and \mathcal{G}^{inf} are independent, the analyst is free to choose a basis in a completely arbitrary way if

Algorithm 1 Basis construction for L_1 penalties

Require: Fitted trend $\hat{\beta}$, order of penalty matrix (k), graph Laplacian matrix (L)

if k is even **then**

$C \leftarrow L^{\frac{k}{2}} \hat{\beta}$

Identify unique values of C , denoted c_1, \dots, c_ℓ .

for $t = 1, 2, \dots, \ell$ **do**

Let c_t be a vector with an entry of 1 whenever a row of C is equal to c_t and 0 otherwise.

end for

$B \leftarrow (L^\dagger)^{\frac{k}{2}} [c_1^T \ \dots \ c_\ell^T]$

else

$C \leftarrow L^{\frac{k+1}{2}} \hat{\beta}$

Identify $A \subseteq \{1, \dots, n\}$ corresponding to the non-zero rows of C . Let B be $(L^\dagger)^{\frac{k+1}{2}}$ with only the columns corresponding to A included.

end if

$B \leftarrow [1 \ B]$

desired. For instance, one can use an L_1 penalty tuned with graph cross-validation as an initial step, but then manually adjust the basis via visual inspection or expert judgement. This contrasts with other methods for post-selection inference on generalized lasso problems ([Chen et al., 2022](#); [Hyun et al., 2018](#)) that are only valid when the analyst commits to using an L_1 penalty in a deterministic way.

Lemma 1. *For a graph \mathcal{G} and corresponding Laplacian matrix L , let $\hat{\beta}$ be the solution of [\(1\)](#) with $k \in \mathbb{N}$ and $D(\beta) := \lambda \|\Delta^{(k)} \beta\|_1$. Let B be the output of [Algorithm 1](#) using $\hat{\beta}$, k , and L as inputs. Then $\hat{\beta}$ is contained in the column span of B .*

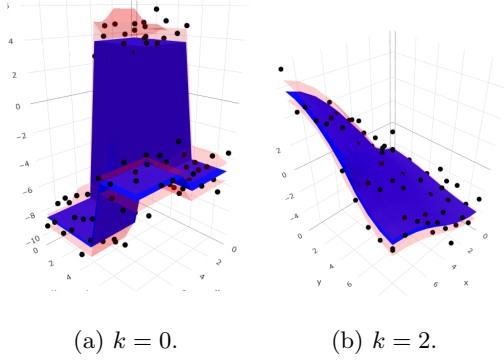


Figure 4: Examples of confidence intervals (red) constructed from [Theorem 1](#) for two example runs. Ground truth (blue) generated from piecewise constant (left) or quadratic (right) bases.

After selecting a basis $B \in \mathbb{R}^m$ using $Y^{\mathcal{G}^{\text{sel}}}$, solve for

$$\hat{\gamma} := \operatorname{argmin}_{\gamma \in \mathbb{R}^m} \left(\sum_{i=1}^n p(y_i^{\mathcal{G}^{\text{inf}}}, (1-\tau)\gamma^T b_i) \right),$$

where b_i denotes the i -th row of B . Our estimate for θ in this context is simply $B\hat{\gamma}$. The target for inference then becomes $B\gamma$ where $\gamma := \operatorname{argmin}_{\gamma} D_{\text{KL}}(P_{\theta} \| P_{B\gamma})$, is the projection parameter which minimizes the KL distance between the true distribution and the working model. Inference can then be performed on this parameter using sandwich estimators for variance, as described in [Theorem 4](#) of [Leiner et al. \(2023\)](#).

4.2 Inference with Nuisance Parameters

A key assumption that we relax is that the distribution of errors is known *exactly*. In particular, [Fact 1](#) requires knowledge of σ^2 for Gaussian data which is unlikely in application. Instead of using [Fact 1](#), we instead work within the **P2** regime defined in [Remark 1](#) which weakens the requirement that $Y^{\mathcal{G}^{\text{inf}}} \perp\!\!\!\perp Y^{\mathcal{G}^{\text{sel}}}$ and only requires that the law of $Y^{\mathcal{G}^{\text{inf}}}|Y^{\mathcal{G}^{\text{sel}}}$ is known. This allows us to consider errors of the form $\epsilon_i \sim N(0, \sigma^2)$ where σ is unknown.

To this end, we construct $Y^{\mathcal{G}^{\text{sel}}} = Y + Z$ by adding user-generated noise $Z \sim N(0, \sigma_0^2 I_n)$ for an arbitrary choice of σ_0 and let $Y^{\mathcal{G}^{\text{inf}}} := Y$. After selecting a basis using $Y^{\mathcal{G}^{\text{sel}}}$, we base inference on the conditional distribution of $Y|Y^{\mathcal{G}^{\text{sel}}}$. Let $\tau := \frac{\mathcal{I}_{Y^{\mathcal{G}^{\text{sel}}}(\mu)}}{\mathcal{I}_Y(\mu)} = \frac{\sigma^2}{\sigma^2 + \sigma_0^2}$ be the proportion of the total Fisher information that is allocated to the selection step. Then,

$$Y|Y^{\mathcal{G}^{\text{sel}}} \sim N\left(\mu(1-\tau) + \tau Y^{\mathcal{G}^{\text{sel}}}, \sigma^2(1-\tau)I_n\right). \quad (3)$$

Given a selected basis $B \in \mathbb{R}^{n \times m}$, we define the target for inference as the projection of the structural trend onto the chosen basis $B(B^T B)^{-1} B^T \mu$. In what follows, we aim to construct confidence intervals that cover any component j of this projection which we label $\eta^T \mu := e_j^T B(B^T B)^{-1} B^T \mu$ for brevity. Since the selection of a direction to project onto is based only on $Y^{\mathcal{G}^{\text{sel}}}$, we assume $\eta = h(Y^{\mathcal{G}^{\text{sel}}})$, where h is some unknown deterministic function of the data.

Under these assumptions, we can create a pivot to construct confidence intervals using [Proposition 1](#) below.

Proposition 1. *Assume $Y \sim N(\mu, \sigma^2 I_n)$ and $Y^{\mathcal{G}^{\text{sel}}} = Y + Z$ where $Z \sim N(0, \sigma_0^2)$ is independent of Y . If $\eta = h(Y^{\mathcal{G}^{\text{sel}}})$ for some deterministic function h , then*

$$\frac{\sqrt{1-\tau}}{\sigma \|\eta\|_2} \left(\frac{\eta^T Y - \tau \eta^T Y^{\mathcal{G}^{\text{sel}}}}{1-\tau} - \eta^T \mu \right) | Y^{\mathcal{G}^{\text{sel}}} \sim N(0, 1).$$

Furthermore, if $\hat{\sigma}$ is a consistent estimator of σ conditional on $Y^{\mathcal{G}^{\text{sel}}}$ and $\hat{\tau} := \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \sigma_0^2}$, then

$$\frac{\sqrt{1-\hat{\tau}}}{\hat{\sigma} \|\eta\|_2} \left(\frac{\eta^T Y - \hat{\tau} \eta^T Y^{\mathcal{G}^{\text{sel}}}}{1-\hat{\tau}} - \eta^T \mu \right) | Y^{\mathcal{G}^{\text{sel}}} \xrightarrow{d} N(0, 1).$$

The challenge in using this pivot is to find a consistent estimator for σ . When the graph-fused lasso is used to estimate μ , [Padilla \(2024\)](#) provides a consistent estimate of the error variance under the assumption that $\|\mu\|_{\infty}$ is bounded and $\frac{\|\Delta^{(1)}\mu\|_1}{n} \rightarrow 0$. However, in the more general case, empirical studies ([Reid et al., 2016](#)) confirm that common estimators of variance are downward biased in high-dimensional settings. In these cases, we can still bound σ^2 between a conservative estimate and anti-conservative estimate and then use these bounds to construct a confidence interval.

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

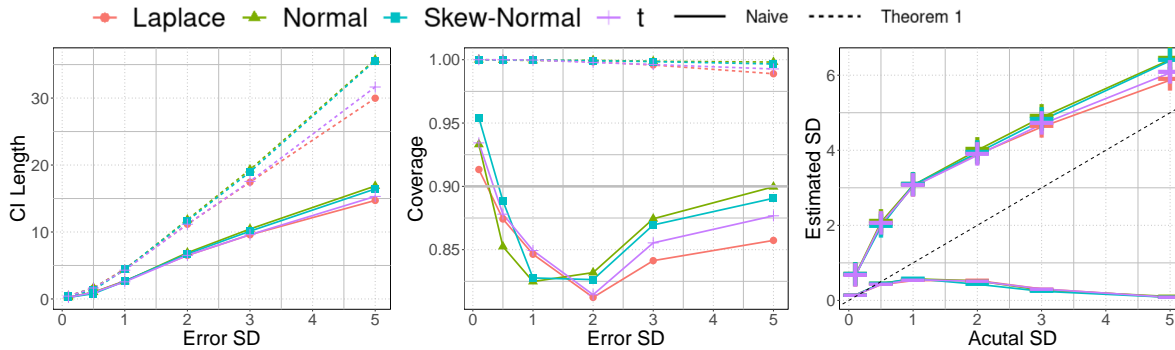
$$\hat{\tau}_{\text{low}} = \frac{\hat{\sigma}_{\text{low}}^2}{\hat{\sigma}_{\text{low}}^2 + \sigma_0^2}, \quad \hat{\tau}_{\text{high}} = \frac{\hat{\sigma}_{\text{high}}^2}{\hat{\sigma}_{\text{high}}^2 + \sigma_0^2},$$

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

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

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

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



(a) Confidence interval length for those designed using [Theorem 1](#) (dashed), compared with naive intervals (solid) constructed under the assumption that $\hat{\sigma}$ is consistent. (b) Coverage rates for intervals designed using [Theorem 1](#) (dashed), compared with naive intervals (solid) for target $1 - \alpha = 0.9$. (c) $\hat{\sigma}_{\text{low}}$ (–) and $\hat{\sigma}_{\text{high}}$ (+) as a function of the true σ .

Figure 5: CI length (left) and coverage (middle) for linear trend filtering ($k = 1$). Intervals designed using [Theorem 1](#) ensure adequate coverage, while naively constructed intervals undercover for target $1 - \alpha = 0.9$. The conservatism of the intervals from [Theorem 1](#) are driven by the difference between conservative and anti-conservative estimates of σ (right).

That is, $\lim_{n \rightarrow \infty} \mathbb{P}(\eta^T \mu \in C_{1-\alpha} \mid Y^{\mathcal{G}^{\text{sel}}}) \geq 1 - \alpha$.

The price of not having consistent estimators of $\hat{\sigma}$ is that these CIs will have an irreducible length, even as $n \rightarrow \infty$. We quantify this gap explicitly in [Corollary 1](#).

Corollary 1. *Assume the conditions of [Theorem 1](#) hold and $\|\eta\|_2 \rightarrow 0$ as $n \rightarrow \infty$, $\hat{\sigma}_{\text{low}} \xrightarrow{p} \sigma_{\text{low}}$, $\hat{\sigma}_{\text{high}} \xrightarrow{p} \sigma_{\text{high}}$ for some $\sigma_{\text{high}} \geq \sigma$, and $\sigma_{\text{low}} \leq \sigma$. Then, the length of $C_{1-\alpha}$ converges in probability to*

$$\eta^T \mu \left(\frac{\sigma_{\text{high}}^2 - \sigma_{\text{low}}^2}{\sigma^2} + \frac{\sigma_{\text{high}}^2 - \sigma_{\text{low}}^2}{\sigma_0^2} - \frac{\sigma_{\text{high}}^2 - \sigma_{\text{low}}^2}{\sigma^2 + \sigma_0^2} \right).$$

The upshot of this result is that the CI length is reducible by constructing more precise estimates of σ . In the case that the analyst manages to construct estimators that are indeed consistent, the CI length will converge to 0, and the interval will cover at an approximately $1 - \alpha$ level instead of being conservative.

In the worst case, we can select $\hat{\sigma}_{\text{low}}^2 := 0$ and use the sample variance as an overestimate, letting $\hat{\sigma}_{\text{high}}^2 := \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$, which [Tibshirani et al. \(2018\)](#) demonstrate is a provable overestimate under mild regularity conditions. However, the length of intervals constructed using such loose bounds may be undesirably wide. More intelligent estimates can be chosen using empirical studies such as [Reid et al. \(2016\)](#). For an overestimate, estimators in the form of (2) where λ is chosen via graph cross-validation and a “one-standard error rule” have a tendency to be conservative. For an underestimate, first choose a basis through a lasso

penalty and graph cross-validation. An estimate of the variance using the residual standard error from an OLS regression fit on the same data using this choice of basis will be downward biased due to not adjusting for the selection step. We leave rigorous guarantees for estimators of this type an open line of inquiry.

4.3 Simulations

We construct synthetic datasets with a ground truth mean μ in exactly the same manner as [Section 3](#). We choose σ_0 using the estimator defined in (2). After creating \mathcal{G}^{sel} and \mathcal{G}^{inf} , \mathcal{G}^{sel} is further split into multiple graphs to select λ via graph cross-validation and a one-standard error rule as described in [Section 3](#). CIs are constructed using [Theorem 1](#) (note examples for a single trial with $k = 0$ and $k = 2$ in [Figure 4](#)). For comparison, we also consider the naive approach that assumes σ was estimated correctly using (2) and constructs \mathcal{G}^{sel} and \mathcal{G}^{inf} using [Fact 1](#) without adjustment. Although confidence intervals are constructed under the assumption of normally distributed errors, we also investigate the effect of misspecified errors (drawn from t , skewed normal, and Laplace distributions). We repeat these experiments over 500 trials for $k = 1$ and report the results in [Figure 5](#). Results demonstrate that confidence intervals constructed from [Theorem 1](#) have proper coverage and the high and low estimators indeed bound the true σ . Naive confidence intervals constructed assuming consistent estimates of $\hat{\sigma}$ severely undercover. Additional results detailing construction of confidence intervals for Poisson distributed count data are contained in the Appendix.

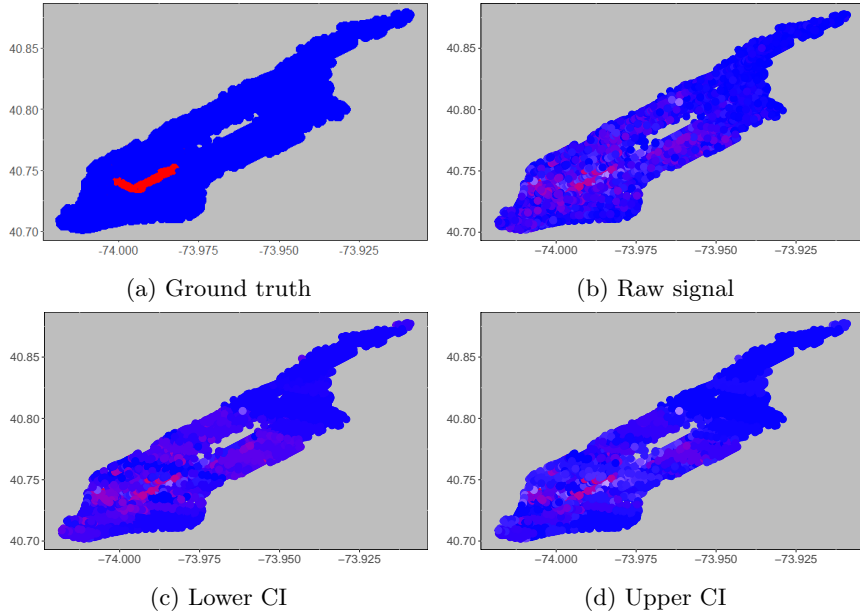


Figure 6: Confidence intervals around dropoffs (red means fewer and white means more than average) at each intersection. Confidence intervals correctly cover the known ground truth signal occurring due to road closures.

5 APPLICATION

We conclude with a real-world application, building on an example given in Wang et al. (2014) that uses grand trend filtering to smooth a trend on the number of taxi drop-offs occurring at different intersections in Manhattan. The dataset was provided by Doraiswamy et al. (2014), who obtained the dataset from the NYC Taxicab and Limousine commission, and consists of nodes representing junctions (intersection between two streets) in Manhattan. Edges exist between two junctions if they are connected by a road. In total, there are 3874 nodes and 7070 edges in the dataset.

The dataset allows us to evaluate whether confidence intervals constructed using graph trend filtering have correct coverage because during certain time periods, we have ground truth knowledge that certain roads are blocked off entirely and the number of drop-offs and pickups will be 0 at specific junctions. One such event that we focus on is the Gay Pride parade, corresponding to a specific time period: 12:00-2:00pm on June 26, 2011. The ground truth is taken from descriptions in the news, which state that the Gay Pride parade started at 36th St. and Fifth Ave. and ended on Christopher St. in Greenwich Village.

A baseline seasonal average was constructed by averaging drop-offs over this specific time period on the same day of each week across the nearest eight weeks. The measurement was then the difference in the amount of drop-offs during this specific time period and the seasonal average. We consider a given trend estimate

as correctly reflecting the ground truth if both the lower and upper bound for the confidence intervals at these intersections are highly negative, reflecting the fact that there are significantly fewer drop-offs at these junctions due to road closures. The results (Figure 6) demonstrate that the ground truth is correctly covered by confidence intervals constructed using Theorem 1.

6 CONCLUSION

We extend a set of techniques for splitting information in non-*i.i.d.* datasets to the graph setting. We also introduce graph cross-validation, which allows for the creation of multiple folds of data with the same graph structure to select hyperparameters in graph learning problems. We apply this method to the problem of structural trend estimation on graphs, and introduce new results which enables inference when variance need to be estimated in the Gaussian case. Empirical studies show that hyperparameters chosen using graph cross-validation (nearly) minimize the risk of the estimator, and confidence intervals constructed using graph fission have correct coverage.

We note several open questions. For applications, graph cross-validation can be used for other estimation problems on graphs, such as hyperparameter tuning for graph neural networks. On the theoretical side, graph fission requires an assumption of correctly specified errors. Empirical results show the procedure is nonetheless robust to modest levels of misspecification, suggesting that asymptotic guarantees are possible.

References

- Arnold, T. B. and R. J. Tibshirani (2016). Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics* 25(1), 1–27.
- Austern, M. and W. Zhou (2020). Asymptotics of cross-validation. *arXiv preprint:2001.11111*.
- Celeux, G. and J.-B. Durand (2008, 02). Selecting hidden markov model state number with cross-validated likelihood. *Computational Statistics* 23, 541–564.
- Chen, Y., S. Jewell, and D. Witten (2022). More powerful selective inference for the graph fused lasso. *Journal of Comp. and Graph. Stat.* 0(0), 1–11.
- Dharamshi, A., A. Neufeld, K. Motwani, L. L. Gao, D. Witten, and J. Bien (2023). Generalized data thinning using sufficient statistics. *arXiv preprint:2303.12931*.
- Doraiswamy, H., N. Ferreira, T. Damoulas, J. Freire, and C. Silva (2014, 12). Using topological analysis to support event-guided exploration in urban data. *Visualization and Computer Graphics, IEEE Transactions on* 20.
- Fan, J., S. Guo, and N. Hao (2011, 10). Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 74(1), 37–65.
- Ghosh, S., W. T. Stephenson, T. D. Nguyen, S. K. Deshpande, and T. Broderick (2020). Approximate cross-validation for structured models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS’20*, Red Hook, NY, USA. Curran Associates Inc.
- 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.
- Joe, H. (1996). Time series models with univariate margins in the convolution-closed infinitely divisible class. *Journal of Applied Probability* 33(3), 664–677.
- Kondor, R. and J. Lafferty (2002). Diffusion kernels on graphs and other discrete input spaces. *International Conference on Machine Learning*.
- Leiner, J., B. Duan, L. Wasserman, and A. Ramdas (2023). Data fission: splitting a single data point. *Journal of the American Statistical Association*.
- Neufeld, A., A. Dharamshi, L. L. Gao, and D. Witten (2023). Data thinning for convolution-closed distributions. *arXiv preprint:2301.07276*.
- Neufeld, A., L. L. Gao, J. Popp, A. Battle, and D. Witten (2022). Inference after latent variable estimation for single-cell RNA sequencing data. *Bio-statistics*.
- Oliveira, N. L., J. Lei, and R. J. Tibshirani (2022). Unbiased test error estimation in the poisson means problem via coupled bootstrap techniques. *arXiv:2212.01943*.
- Padilla, O. H. M. (2024). Variance estimation in graphs with the fused lasso. *arXiv preprint:2207.12638*.
- Rasines, D. G. and G. A. Young (2023). Splitting strategies for post-selection inference. *Biometrika*.
- Reid, S., R. Tibshirani, and J. Friedman (2016). A study of error variance estimation in lasso regression. *Statistica Sinica* 26(1), 35–67.
- Sharpnack, J., A. Singh, and A. Krishnamurthy (2013). Detecting activations over graphs using spanning tree wavelet bases. In *International Conference on Artificial Intelligence and Statistics*, Volume 31, pp. 536–544. PMLR.
- Smola, A. and R. Kondor (2003). Kernels and regularization on graphs. *Conference on Computational Learning Theory* 2777, 144–158.
- Stein, C. M. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics* 9(6), 1135 – 1151.
- Tian, X. and J. Taylor (2018). Selective inference with a randomized response. *The Annals of Stat.* 46(2), 679–710.
- Tibshirani, R. J., A. Rinaldo, R. Tibshirani, and L. Wasserman (2018). Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Stat.* 46(3), 1255–1287.
- Wang, Y.-X., J. Sharpnack, A. Smola, and R. Tibshirani (2014, 10). Trend filtering on graphs. *Journal of Machine Learning Research* 17.
- Yu, G. and J. Bien (2017, 12). Estimating the error variance in a high-dimensional linear model. *Biometrika* 106.

CHECKLIST

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/**Not Applicable**]
 - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**/No/Not Applicable]
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
 - (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
 - (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]
3. For all figures and tables that present empirical results, check if you include:
 - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/No/Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. [**Yes**/No/Not Applicable]
 - (b) The license information of the assets, if applicable. [**Yes**/No/**Not Applicable**]
 - (c) New assets either in the supplemental material or as a URL, if applicable. [**Yes**/No/**Not Applicable**]
 - (d) Information about consent from data providers/curators. [**Yes**/No/Not Applicable]
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Yes**/No/**Not Applicable**]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. [**Yes**/No/**Not Applicable**]
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Yes**/No/**Not Applicable**]
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Yes**/No/**Not Applicable**]

Graph Fission and Cross-Validation: Supplementary Materials

A DEFERRED PROOFS

A.1 Proof of Lemma 1

This follows from a result in Wang et al. (2014), which we recall here.

Lemma 2 (Lemma 1 from Wang et al. (2014)). *Assume without a loss of generality that \mathcal{G} is connected (otherwise the results apply to each connected component of \mathcal{G}). Let D, L be the oriented incidence matrix and Laplacian matrix of \mathcal{G} . For even k , let $A \subseteq \{1, \dots, m\}$, and let \mathcal{G}_{-A} denote the subgraph induced by removing the edges indexed by A (i.e., removing edges $e_\ell, \ell \in A$). Let C_1, \dots, C_s be the connected components of \mathcal{G}_{-A} . Then*

$$\text{null} \left(\Delta_{-A}^{(k+1)} \right) = \text{span}\{\mathbf{1}\} + (L^\dagger)^{\frac{k}{2}} \text{span} \{ \mathbf{1}_{C_1}, \dots, \mathbf{1}_{C_s} \}$$

where $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^n$, and $\mathbf{1}_{C_1}, \dots, \mathbf{1}_{C_s} \in \mathbb{R}^n$ are the indicator vectors over connected components. For odd k , let $A \subseteq \{1, \dots, n\}$. Then

$$\text{null} \left(\Delta_{-A}^{(k+1)} \right) = \text{span}\{\mathbf{1}\} + \left\{ (L^\dagger)^{\frac{k+1}{2}} v : v_{-A} = 0 \right\}.$$

Since the penalty term enforces sparsity in $\Delta^{(k)}\hat{\beta}$, we know that $\text{supp} \left(\Delta^{(k)}\hat{\beta} \right) = A$ for some active set $A \subseteq \{1, \dots, m\}$ which implies that $\hat{\beta} \in \text{null} \left(\Delta_{-A}^{(k)}\hat{\beta} \right)$. Working backwards, this implies that for even k

$$L^{\frac{k}{2}}\hat{\beta} \in L^{\frac{k}{2}} \text{span}\{\mathbf{1}\} + L^{\frac{k}{2}} (L^\dagger)^{\frac{k}{2}} \text{span} \{ \mathbf{1}_{C_1}, \dots, \mathbf{1}_{C_s} \} = \text{span} \{ \mathbf{1}_{C_1}, \dots, \mathbf{1}_{C_s} \}.$$

So, the connected components can be identified by noting the unique values of $L^{\frac{k}{2}}\hat{\beta}$ and seeing which nodes share these values in common. For odd k , we have that

$$L^{\frac{k+1}{2}}\hat{\beta} \in L^{\frac{k+1}{2}} \text{span}\{\mathbf{1}\} + L^{\frac{k+1}{2}} (L^\dagger)^{\frac{k+1}{2}} v = v,$$

where v contains 0 at the non-zero components (A). However, the term $\left\{ (L^\dagger)^{\frac{k+1}{2}} v : v_{-A} = 0 \right\}$ is equivalent to removing the corresponding columns from $(L^\dagger)^{\frac{k+1}{2}}$ and requiring $v \in \mathbb{R}^{|A|}$ instead of \mathbb{R}^n , leading to the construction given in Algorithm 1.

A.2 Proof of Proposition 1

We start by recalling that $Y|f(Y) \sim N(\mu(1-\tau) + f(Y)\tau, \sigma^2(1-\tau)I_n)$. By assumption η is a deterministic function of $f(Y)$. Therefore $Y|\eta = \hat{\eta}, f(Y)$ is equal in distribution to $Y|f(Y)$ for all $\hat{\eta}$. Thus, we will treat η as fixed conditional on $f(Y)$. Standard properties of the multivariate normal give us that $\eta^T Y|f(Y) \sim N(\eta^T \mu + (1-\tau)\eta^T f(Y), \eta^T \eta \mu(1-\tau))$. We then rearrange terms to arrive at the pivot

$$\frac{\sqrt{1-\tau}}{\sigma \|\eta\|_2} \left(\frac{\eta^T Y - \tau \eta^T f(Y)}{1-\tau} - \eta^T \mu \right) |f(Y) \sim N(0, 1).$$

Finally, if $\hat{\sigma}$ is consistent conditional on $f(Y)$, then so is $\hat{\tau}$ by the continuous mapping theorem. We can therefore apply the continuous mapping theorem once again to conclude that

$$\frac{\sqrt{1-\hat{\tau}}}{\hat{\sigma} \|\eta\|_2} \left(\frac{\eta^T Y - \hat{\tau} \eta^T f(Y)}{1-\hat{\tau}} - \eta^T \mu \right) |f(Y) \xrightarrow{d} N(0, 1).$$

A.3 Proof of Theorem 1

Let $E_1 = \{\hat{\sigma}_{\text{low}} > \sigma\}$ and $E_2 = \{\hat{\sigma}_{\text{high}} < \sigma\}$, and $E_3 = \{\hat{\sigma}_{\text{low}} < \sigma\} \cap \{\sigma < \hat{\sigma}_{\text{high}}\}$. We then have that

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}(E_3 | f(Y)) &= \lim_{n \rightarrow \infty} 1 - \mathbb{P}(E_1 \cup E_2 | f(Y)) \\ &= \lim_{n \rightarrow \infty} 1 - \mathbb{P}(E_1 | f(Y)) - \mathbb{P}(E_2 | f(Y)) \\ &= 1. \end{aligned}$$

We already know from Proposition 1 that

$$C_1 := \left[\frac{\eta^T Y - \tau \eta^T f(Y)}{1-\tau} - z_{\alpha/2} \frac{\sigma \|\eta\|_2}{\sqrt{1-\tau}}, \frac{\eta^T Y - \tau \eta^T f(Y)}{1-\tau} + z_{\alpha/2} \frac{\sigma \|\eta\|_2}{\sqrt{1-\tau}} \right]$$

is a valid $1 - \alpha$ confidence interval. Letting

$$C_2 := \left[A_1 - z_{\alpha/2} \frac{\|\eta\|_2 \hat{\sigma}_{\text{high}}}{\sqrt{1-\hat{\tau}_{\text{high}}}}, A_2 + z_{\alpha/2} \frac{\|\eta\|_2 \hat{\sigma}_{\text{high}}}{\sqrt{1-\hat{\tau}_{\text{high}}}} \right],$$

we note that $f(\tau) = \frac{\eta^T Y - \tau \eta^T f(Y)}{1-\tau}$ is monotonic and continuous with respect to $\tau \in [0, 1]$. If $\hat{\sigma}_{\text{low}} \leq \sigma \leq \hat{\sigma}_{\text{high}}$, then $\hat{\tau}_{\text{high}} \leq \tau \leq \hat{\tau}_{\text{low}}$. We conclude by invoking the intermediate value theorem to see that $\frac{\eta^T Y - \tau \eta^T f(Y)}{1-\tau} \in [A_1, A_2]$ whenever E_3 holds. We also note $\hat{\sigma}_{\text{high}} > \sigma$ implies that $\frac{\hat{\sigma}_{\text{high}}}{\|\eta\|_2 \sqrt{1-\hat{\tau}_{\text{high}}}} \geq \frac{\sigma}{\|\eta\|_2 \sqrt{1-\tau}}$ because τ monotonically increases with σ and is bounded in $[0, 1]$. These two arguments taken together imply that $\{\eta^T \mu \in CI_1\} \cap E_3 \subseteq \{\eta^T \mu \in CI_2\} \cap E_3$.

Therefore, $\lim_{n \rightarrow \infty} \mathbb{P}(\{\eta^T \mu \in CI_2\}) = \lim_{n \rightarrow \infty} \mathbb{P}(\{\eta^T \mu \in CI_2\} \cap E_3) \geq \lim_{n \rightarrow \infty} \mathbb{P}(\{\eta^T \mu \in CI_1\} \cap E_3) = \lim_{n \rightarrow \infty} \mathbb{P}(\{\eta^T \mu \in CI_1\}) = 1 - \alpha$, concluding the proof.

A.4 Proof of Corollary 1

The confidence interval length is given by $|A_1 - A_2| + 2z_{\alpha/2} \frac{\|\eta\|_2 \hat{\sigma}_{\text{high}}}{\sqrt{1-\hat{\tau}_{\text{high}}}}$. By assumption $\|\eta\|_2 \rightarrow 0$ so we only have to concern ourselves with $|A_1 - A_2|$. We further have that $\hat{\tau}_{\text{low}} \xrightarrow{p} \tau_{\text{low}}$, $\hat{\tau}_{\text{high}} \xrightarrow{p} \tau_{\text{high}}$, $\hat{\sigma}_{\text{low}} \xrightarrow{p} \sigma_{\text{low}}$, $\hat{\sigma}_{\text{high}} \xrightarrow{p} \sigma_{\text{high}}$ either by assumption or via the continuous mapping theorem. This implies that $A_1 - A_2$ will

converge in distribution to a variable that we will label \tilde{A} with conditional distribution of

$$\tilde{A}|f(Y) \sim N\left(\eta^T \mu \left(\frac{\sigma_{\text{low}}^2 - \sigma_{\text{high}}^2}{\sigma_0^2}\right) - \eta^T f(Y) \left(\frac{\sigma_{\text{low}}^2 - \sigma_{\text{high}}^2}{\sigma^2 + \sigma_0^2} - \frac{\sigma_{\text{low}}^2 - \sigma_{\text{high}}^2}{\sigma^2}\right), \|\eta\|_2 \sigma^2 (1 - \tau) \left(\frac{\sigma_{\text{low}}^2 - \sigma_{\text{high}}^2}{\sigma^2 + \sigma_0^2} - \frac{\sigma_{\text{low}}^2 - \sigma_{\text{high}}^2}{\sigma^2}\right) I_n\right).$$

Because $\|\eta\|_2 \rightarrow 0$, the conditional and unconditional variance of this variable will converge to 0 and the marginal expectation will be equal to:

$$E[\tilde{A}] = E\left[E[\tilde{A}|f(Y)]\right] = \eta^T \mu \left(\frac{\sigma_{\text{low}}^2 - \sigma_{\text{high}}^2}{\sigma^2} + \frac{\sigma_{\text{low}}^2 - \sigma_{\text{high}}^2}{\sigma_0^2} - \frac{\sigma_{\text{low}}^2 - \sigma_{\text{high}}^2}{\sigma^2 + \sigma_0^2}\right).$$

Taking everything together and then applying the continuous mapping theorem gives us the result that

$$|A_1 - A_2| \xrightarrow{p} \eta^T \mu \left(\frac{\sigma_{\text{high}}^2 - \sigma_{\text{low}}^2}{\sigma^2} + \frac{\sigma_{\text{high}}^2 - \sigma_{\text{low}}^2}{\sigma_0^2} - \frac{\sigma_{\text{high}}^2 - \sigma_{\text{low}}^2}{\sigma^2 + \sigma_0^2}\right).$$

B ADDITIONAL EXPERIMENTAL RESULTS

B.1 Confidence Intervals for Poisson Data

We repeat the experiments in Section 4, but with Poisson distributed data. We note that because no unknown parameters do not need to be estimated in this case, the methodology is more straightforward. Fact 1 can be applied directly, and confidence intervals from standard software packages generally have correct coverage of the structural trend. In particular, we use the decompositions defined in Example 2 to construct \mathcal{G}_{sel} and \mathcal{G}_{inf} . In particular, the procedure then becomes:

1. Using only \mathcal{G}_{sel} , fit $\hat{\beta}$ as the solution to the optimization problem,

$$\hat{\beta} := \operatorname{argmin}_{\beta \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n (-y_i \beta_i + \exp(\beta_i)) + \|\beta\|_1.$$

2. The solutions will be sparse, so construct a basis B using Lemma 1 as before.
3. Fit a new trend $\hat{\gamma}$ using any implementations of GLMs with B as the set of corresponding covariates. To generate confidence intervals, we recommend using sandwich estimators of variance as described in [Leiner et al. \(2023\)](#). For instance, those implemented in the `CLUBSANDWICH` package in R.
4. The confidence intervals will cover the projection parameter $\gamma := \operatorname{argmin}_{\gamma} D_{\text{KL}}(P_{\theta} \| P_{B\gamma})$, where P_{θ} denotes the Poisson distribution with parameter θ .

Experimental results are shown in [Figure 7](#) and are broadly consistent with the case of unknown Gaussian errors.

B.2 Additional Simulations for Graph Cross Validation

We repeat the cross-validation experiments, but investigate how sensitive the methodology is to a misspecified model for the error distribution. In particular, we experiment with three different choices for the error term: Laplace, skew normal distribution with scale parameter equal to 1 and shape parameter equal to 5, and a t -distribution with 5 degrees of freedom. In all cases, the error terms are also rescaled to have 0 mean and unit variance. Results are shown for $k = 0$ in [Figure 8](#) and $k = 1$ in [Figure 9](#). We note the trends are nearly identical across all permutations.

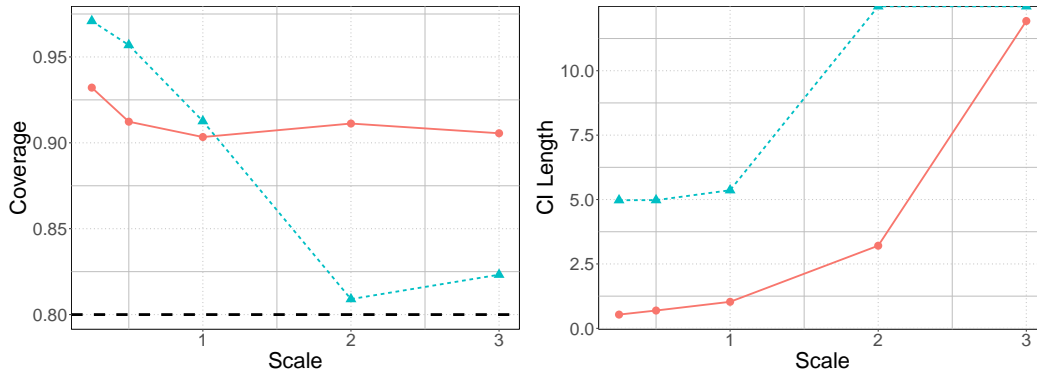


Figure 7: CI length and coverage (against target $1 - \alpha = 0.8$) as magnitude of underlying trend θ increases for Poisson-distributed data with $k = 0$ (solid) and $k = 1$ (dashed) graph penalties. The procedure guarantees correct coverage empirically, with confidence intervals increasing with larger θ .

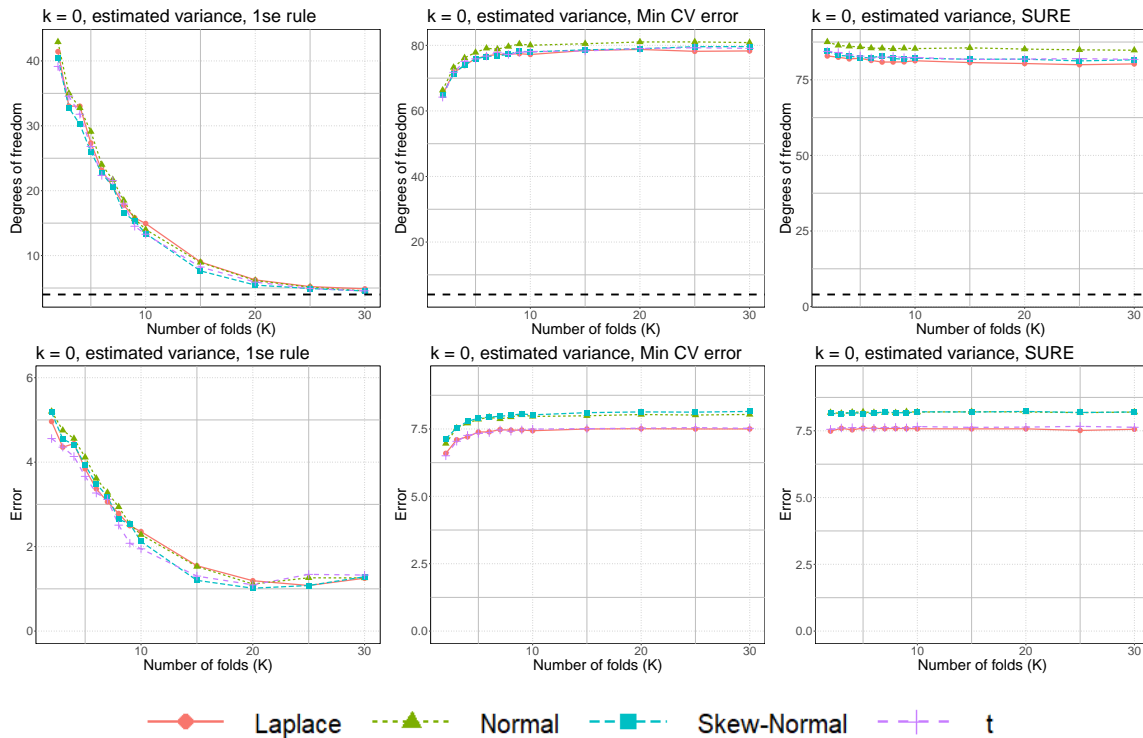


Figure 8: Degrees of freedom and error when constructing $\hat{\beta}$ using piecewise constant trend filtering ($k = 0$) with graph cross-validation. Results are shown across possibly different choices for misspecification of the errors. We note that results are nearly identical across all of these permutations. This suggests that the procedure is robust to moderate levels of misspecification, and a Central Limit Theorem may apply in this case.

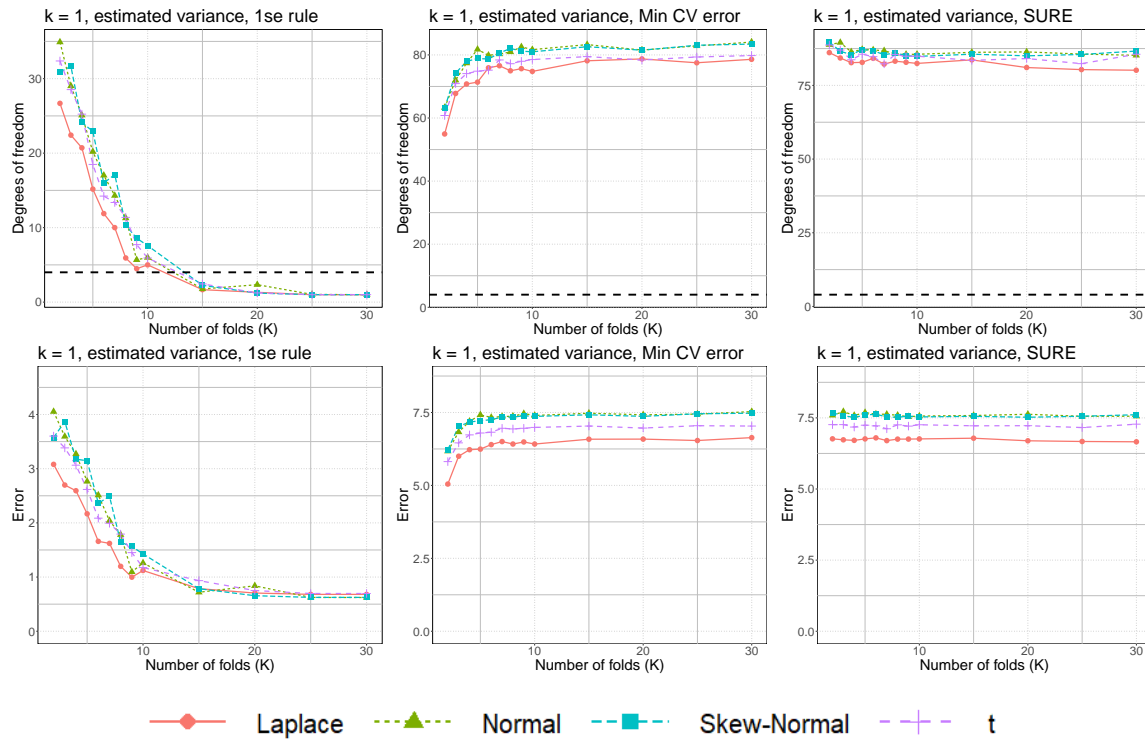


Figure 9: Same view as Figure 8, but for linear trend filtering with $k = 1$. We again see similar behaviors across all error types. The spread between quantitative results across distributions is relatively larger than the case where $k = 0$, though interestingly misspecified errors lead to lower risk when using cross-validation. This is an interesting area for future investigation.