Generalized Data Thinning Using Sufficient Statistics

ST 793 Project Report

Netra Prabhu, Ananya Roy

North Carolina State University

Date: 11^{th} **December, 2024**

1 Motivation of the Problem

The motivation of the problem of generalized data thinning comes from the drawbacks of typical procedures such as sample splitting and selective inference when using a single data source for model calibration and evaluation. The procedure of sample splitting in traditional fashion entails partitioning of the available data into two parts, which may either be used for hypothesis testing or hypothesis generation, or for model fitting and model validation. However, this approach is disadvantageous when observation are not independent, such as in time series data, or where the sample size is small after splitting, so that the model cannot make meaningful inferences. Selective inference, on the other hand, assumes that the procedure for hypothesis generation must be outlining precisely and many times depends only on strong assumptions such as Gaussianity which restricts it from being used for more general uses. In order to solve these problems, the notion of generalized data thinning using sufficient statistics was offered up as a flexible approach that analyzes a random variable into independent components but at the same time keep enough amount of information pertaining to unknown parameters. This approach extends the applicability of the data-splitting techniques, especially in the case of the dependent data or small samples, and gives a common view point of different thinning techniques.

2 Literature Review

As mentioned earlier, data thinning has in recent years become an important strategy for model validation, hypothesis testing, and indeed utilization of limited data resources. Other approaches including sample splitting and selective inference, have their weaknesses especially with dependent data or a small sample size. Some recent advancements for generalized thinning score the above difficulties realizing a theory to divide a random variable into various parts all of which are independent, but without sacrificing the information about unknown parameters.

2.1 Traditional Approaches to Data Thinning

- **Sample Splitting:** Sample splitting is one of the easiest partitioning methods that partition a dataset into two sets. More often one part of the dataset is employed to adjust parameters of the model while the other part is used for evaluating performance of the model. This method was introduced by Cox (1975) [2] and it was expanded on form the basis of cross validation methods (Hastie et al. 2009) [3]. However, sample splitting suffers from several problems as discussed below. For example, if data contain outliers these are assigned to only one subset may be such outliers might have minimal effects on the model fitting than when distributed among the subsets. Moreover, where observations are dependent, as in time series data, sample splitting does not meet the requirement of independent training and test samples. To overcome this problem, splitting of data into two parts is not suitable for small size data sets since the data available for the actual model fitting decreases considerably.

  There are more drawbacks of the sample-splitting approach.

  1. Sample splitting does not support **observation-level validation**. For example, suppose the data set is derived from the fifty states of the United States. In that case, inference can only be made about the states that were not included in the training set thus decreasing statistical leverage.
  2. For **unsupervised learning** problems such as clustering, sample splitting may also be inadequate because clusters found in the training sample might not exist in the validation sample.

- **Selective Inference:** Sample splitting can be contrasted with selective inference which allows both hypotheses testing and generation in the same sample. The main concept lies in the conditioning on the event that some hypothesis has been chosen and, thus, eliminating selection bias. Selecting post-RAM inference Rearranging the analysis to provide a coherent and easily applicable procedure is the other recommendation of Taylor and Tibshirani (2015) [5]. This newly developed procedure has been applied in areas such as changepoint detection (Jewell et al., 2022) [4] and clustering (Chen and Witten, 22) [1]. However, this approach has it's defects. Selective inference entails full specification of the procedure used to come up with the hypothesis and in most cases, the observations were assumed to originate from a Gaussian distribution. These restrictions reduce its versatility, especially as a first-step dimensionality reduction technique in non-Gaussian or with large outliers datasets.

2.2 A New Perspective: Generalized Data Thinning

In contrast to the sample splitting, this approach provides each observation both for training and validation phases with the positive impact on the statistic efficacy. It is recognised that the notion of convolution-closed data thinning has already existed for well-known families that include the Gaussian,Poisson family which is demonstrated in the subsequent section. By using this method the random variable can be broken down into a succession of sums of independent components yet keeping the information of unknown parameters of the distribution.
The motivation for generalized data thinning is therefore to develop a unified and flexible alternative to sample splitting that overcomes its limitations while still supporting hypothesis testing, model validation, and prediction error estimation in a broader range of settings. The primary goal is to construct methods that support thinning for a large family of distributions, including those outside the classical convolution-closed families.

3 Problem Statement

The motivation for generalized data thinning, is therefore to come up with a more versatile approach than sample splitting while still allowing hypothesis testing, model validation and prediction error estimation in diverse scenarios. The primary goal is to build tools and frameworks for thinning that apply to a vast number of distribution, convolution-closed cases. Here in this report our focus is mainly on simulation studies and data analysis. Before jumping into it,we will look at the generalized thinning procedure.

4 Generalized Thinning using Sufficient Statistics

- **Definition 1: Generalized Data Thinning:** Consider a family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$. Suppose that there exists a distribution G_t, not depending on θ, and a deterministic function $T(\cdot)$ such that when we sample $(X^{(1)}, \dots, X^{(K)}) \mid X$ from G_X, for $X \sim P_\theta$, the following properties hold:

  1. $X^{(1)}, \dots, X^{(K)}$ are mutually independent (with distributions depending on θ), and
  2. $X = T(X^{(1)}, \dots, X^{(K)})$.

  Then we say that $\mathcal{P}$ is thinned by the function $T(\cdot)$.

This can be thought of as breaking **X** into K independent pieces ensuring that no information about θ is lost. Sample splitting can be seen as a special case of generalized thinning. Thinning of convolution-closed families of distributions such as Binomial,Gamma,Poisson and Gaussian using $T(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}) = \sum_{k=1}^{K} \mathbf{x}^{(k)}$. To ensure that there exists a distribution G_t as in Definition 1 which does not depend on θ. It turns out that sufficiency is the key principle required for this assurance.

- **Theorem 1:** Suppose P is thinned by a function $T(\cdot)$ and, for $X \sim P_\theta$, let $Q_\theta^{(1)} \times \cdots \times Q_\theta^{(K)}$ denote the distribution of the mutually independent random variables $(X^{(1)}, \ldots, X^{(K)})$ sampled as in Definition 1. Then, the following hold:

  (a) $T(X^{(1)}, \ldots, X^{(K)})$ is a sufficient statistic for θ based on $(X^{(1)}, \ldots, X^{(K)})$.

  (b) The distribution G_t in Definition 1 is the conditional distribution

  $$(X^{(1)}, \ldots, X^{(K)}) \mid T(X^{(1)}, \ldots, X^{(K)}) = t,$$

  where $(X^{(1)}, \ldots, X^{(K)}) \sim Q_\theta^{(1)} \times \cdots \times Q_\theta^{(K)}$.

Following is the simple algorithm to find families of distributions $\mathcal{P}$ and functions $T(.)$ such that $\mathcal{P}$ can be thinned by $T(.)$.

Algorithm 1 Finding distributions that can be thinned

Choose K families of distributions, $Q^{(k)} = \{Q_\theta^{(k)} : \theta \in \Omega\}$ for $k = 1, \ldots, K$. Let $(X^{(1)}, \ldots, X^{(K)}) \sim Q_\theta^{(1)} \times \cdots \times Q_\theta^{(K)}$, and let $T(X^{(1)}, \ldots, X^{(K)})$ denote a sufficient statistic for θ. Let P_θ denote the distribution of $T(X^{(1)}, \ldots, X^{(K)})$. By construction, the family $P = \{P_\theta : \theta \in \Omega\}$ is thinned by $T(\cdot)$.

Now, let's look at how the above regime can be utilized in different settings.

4.1 Thinning Natural Exponential Families

According to (Lehmann & Romano, 2005) a natural exponential family starts with a known probability distribution H, and then forms a family of distributions $\mathcal{P}_H = \{P_H^\theta : \theta \in \Omega\}$ based on H, as follows:

$$dP_H^\theta(x) = e^{x^\top \theta - \psi_H(\theta)} \, dH(x). \tag{1}$$

where $e^{-\psi_H(\theta)}$ is a normalizing constant and Ω is the set of θ for which this normalization is possible (i.e., for which $\psi_H(\theta) < \infty$).

- **Theorem 2: (Thinning natural exponential families by addition)**
  The natural exponential family $\mathbb{P}_H$ can be thinned by $T(x^{(1)}, \ldots, x^{(K)}) = \sum_{k=1}^{K} x^{(k)}$ into K natural exponential families $\mathbb{P}_{H_1}, \ldots, \mathbb{P}_{H_K}$ if and only if H is the K-way convolution of $H_1, \ldots, H_K$. The K-way convolution is defined as follows:

- **Definition 2 (K-way convolution):** A probability distribution H is the K-way convolution of distributions $H_1, \ldots, H_K$ if $\sum_{k=1}^{K} Y_k \sim H \quad \text{for} \quad (Y_1, \ldots, Y_K) \sim H_1 \times \cdots \times H_K$.

4.2 Thinning Natural into General Exponential Families

We will now see how can we use Algorithm 1 in case where $Q^{(K)}$ are (possibly non-natural) exponential families, for which the sufficient statistic need not be the identity.

- **Proposition 1 : (Thinning natural exponential families with more general functions $T()$.)** Let $X^{(1)}, \ldots, X^{(K)}$ be independent random variables with $X^{(k)} \sim Q_\theta^{(k)}$ for $k = 1, \ldots, K$ from any (i.e., possibly non-natural) exponential families $Q^{(k)}$ as in (2).

  Let P_θ denote the distribution of $\sum_{k=1}^K T^{(k)}(X^{(k)})$. Then, $P = \{P_\theta : \theta \in \Omega\}$ is a natural exponential family, and we can thin it into $X^{(1)}, \ldots, X^{(K)}$ using the function

$$T(x^{(1)}, \ldots, x^{(K)}) = \sum_{k=1}^K T^{(k)}(x^{(k)}).$$

- **Theorem 3 (Thinning functions for natural exponential families)**
  Suppose $X \sim P_\theta$, where $P = \{P_\theta : \theta \in \Omega\}$ is a full-rank natural exponential family with density/mass function $p_\theta(x) = \exp\left(\theta^\top x - \psi(\theta)\right) h(x)$. If P can be thinned by $T(\cdot)$ into $X^{(1)}, \ldots, X^{(K)}$, then:

  1. The function $T(x^{(1)}, \ldots, x^{(K)})$ is of the form $T(x^{(1)}, \ldots, x^{(K)}) = \sum_{k=1}^K T^{(k)}(x^{(k)})$.
  2. $X^{(k)} \overset{\text{ind}}{\sim} Q_\theta^{(k)}$, where $Q_\theta^{(k)}$ is an exponential family with sufficient statistic $T^{(k)}(X^{(k)})$.

This provides a more flexible approach. For e.g., in Algorithm 1 suppose we start with $X^{(k)} \sim Gamma(\frac{\alpha}{K}, \theta)$ for $k = 1, 2, ..., K$ and we know that the statistic $T(X^{(1)}, \ldots, X^{(K)}) = \sum_{k=1}^K X^{(k)}$ is sufficient for θ. Hence, we can thin the distribution of $\sum_{k=1}^K X^{(k)}$.. Sampling from G_t as in Theorem 1 corresponds to the multi-fold gamma data thinning procedure of Neufeld et . (2023) where $\epsilon_k = \frac{1}{K}$.

4.3 Indirect Thinning of General Exponential Families

Whenever we choose to thin a function $S(X)$ instead of thinning X; if $S(X)$ is sufficient for θ based on X then t thinning $S(X)$rather than X does not result in a loss of information about θ.
Let $P = \{P_\theta : \theta \in \Omega\}$ be a general exponential family. That is,
$dP_\theta(x) = \exp\left([S(x)]^\top \eta(\theta) - \psi(\theta)\right) dH(x)$, where $e^{-\psi(\theta)}$ is the normalizing constant. Since $S(X)$ is sufficient for θ, we can indirectly thin X through $S(\cdot)$ as follows:

- We now consider $X^{(1)}, \ldots, X^{(K)}$ that belong to a general exponential family, where $T^{(k)}(\cdot)$ in is not necessarily the identity. Suppose further that

$$S(X) \overset{D}{=} \sum_{k=1}^K T^{(k)}(X^{(k)}).$$

  Then, by Proposition 1, we can indirectly thin X through $S(\cdot)$ into $X^{(1)}, \ldots, X^{(K)}$ using the function

$$T(x^{(1)}, \ldots, x^{(K)}) = \sum_{k=1}^K T^{(k)}(x^{(k)}).$$

For e.g.,Suppose $X \sim \mathcal{N}_n\left(\theta_1 \mathbf{1}_n, \theta_2 \mathbf{I}_n\right)$. Then $S(X)$ is sufficient for $\theta = (\theta_1, \theta_2)$, where

$$S(x) = \left(\frac{1}{n}\sum_{i=1}^n x_i, \quad \frac{1}{n-1}\sum_{i=1}^n \left(x_i - \frac{1}{n}\sum_{i'=1}^n x_{i'}\right)^2\right).$$

Indirectly thinning X through $S(\cdot)$ into $K = n$ univariate normals. To do that, start with $X^{(k)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_1, \theta_2)$ for $k = 1, \ldots, K$. A sufficient statistic for θ based on $(X^{(1)}, \ldots, X^{(K)})$ is $T(X^{(1)}, \ldots, X^{(K)})$, where $T(x^{(1)},\dots,x^{(K)}) = S(x^{(1)},\dots,x^{(K)^T})$ Furthermore, $T(X^{(1)},\dots,X^{(K)})$ has the same distribution as $S(X)$, since $(X^{(1)},\dots,X^{(K)})^T$ and X have the same distribution. This establishes that we can indirectly thin X through $S(\cdot)$ by $T(\cdot)$.

4.4 Thinning Outside Exponential Families

For thinning $Uniform(0,\theta)$, we start with $X^{(k)} \overset{\text{i.i.d.}}{\sim} \theta \cdot \text{Beta}\left(\frac{1}{K},1\right)$ for $k = 1,\dots,K$, and we know that $T(X^{(1)},\dots,X^{(K)}) = \max(X^{(1)},\dots,X^{(K)})$ is sufficient for θ. Furthermore, $\max(X^{(1)},\dots,X^{(K)}) \sim \text{Unif}(0,\theta)$. Thus, define G_t to be the conditional distribution of $(X^{(1)},\dots,X^{(K)})$ given $\max(X^{(1)},\dots,X^{(K)}) = t$. Then, by Theorem 1, we can thin $X \sim \text{Unif}(0,\theta)$ by sampling from G_t. To do this, first draw $C \sim \text{Categorical}_K\left(\frac{1}{K},\dots,\frac{1}{K}\right)$. Then, set $X^{(k)} = C_k \cdot X + (1 - C_k) \cdot Z_k$, where $Z_k \overset{\text{i.i.d.}}{\sim} X \cdot \text{Beta}\left(\frac{1}{K},1\right)$.

5 Limitations

There are a few families where thinning strategies do not work. For e.g., a natural exponential family that is based on a distribution that cannot be written as the convolution of two distributions. If $Z^{(1)}$ and $Z^{(2)}$ are independent, nonconstant random variables, then $Z^{(1)} + Z^{(2)}$ cannot be a Bernoulli random variable. Since, Bernoulli is not a convolution of non-constant random variables, it cannot be thinned by any function $T()$. The Cauchy family also cannot be thinned.

6 Simulation

6.1 Simulation Study 1: Optimal ϵ for Data Thinning

Through the first simulation study, we see the impact of data thinning on regression analysis by splitting the response variable, y, which follows a normal distribution. Specifically, the thinning process divides y into two independent components, $y^{(1)}$ and $y^{(2)}$, using thinning proportions ϵ_1 and ϵ_2, where $\epsilon_1 + \epsilon_2 = 1$. Each component follows a normal distribution:

$$y^{(1)} \sim N(\epsilon_1\mu, \epsilon_1\sigma^2), \quad y^{(2)} \sim N(\epsilon_2\mu, \epsilon_2\sigma^2).$$

In this setup, $y^{(1)}$ is used as the training data and $y^{(2)}$ as the test data to evaluate the prediction error, represented by the Mean Squared Error (MSE).

Using the `iris` dataset (excluding the species column) with the simulated response variable y, the study varies ϵ from 0 to 1 in increments of 0.01. For each ϵ, the regression model is trained on $y^{(1)}$ and tested on $y^{(2)}$, with the corresponding MSE recorded. The results, visualized in the attached plot, reveal a U-shaped relationship between ϵ and MSE. Notably, the prediction error is minimized when $\epsilon_1 = \epsilon_2 = 0.5$, achieving a balanced partition between training and test data as shown in Figure 1. This outcome aligns with theoretical expectations, as balanced thinning maximizes the informativeness of both subsets, ensuring optimal model training and evaluation.

The study underscores the importance of selecting appropriate thinning proportions to minimize prediction error in regression tasks.

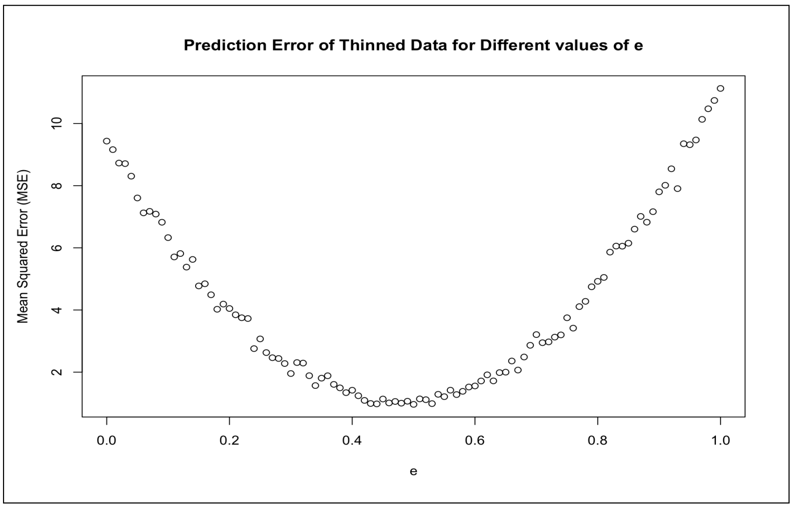
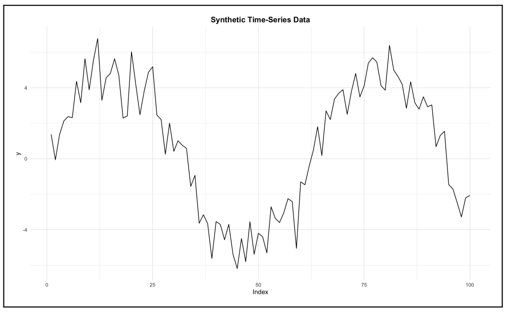


Figure 1: Prediction Error of Thinned Data for Different Values of ϵ.

6.2 Simulation Study 2: Regular vs Data-Thinned CV

This simulation study compares two cross-validation methods—Regular Cross-Validation and Data-Thinned Cross-Validation—using synthetic time-series data. The objective is to evaluate and compare their performance in terms of Mean Squared Error (MSE) for predicting a lagged time-series response variable. First, synthetic time-series data is generated with 100 points modeled as $y = 5\sin(x) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ represents normally distributed noise. Lagged variables (y_{t-1} and y_{t-2}) are computed to form a predictive dataset. The response variable (y) is split into training and test sets, and multiple iterations are performed to assess both methods.



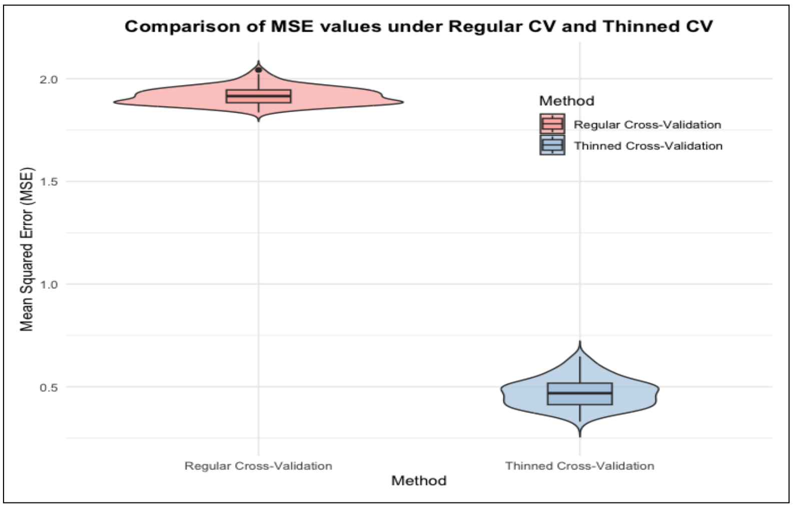


Figure 2: Comparison of Cross-Validation Methods: Distribution of MSE Values

In **Data-Thinned Cross-Validation**, the response variable y is thinned into $K = 3$ subsets with proportions $\epsilon_1 = \epsilon_2 = \epsilon_3 = \frac{1}{3}$. The thinning process ensures that each subset maintains the distributional properties of the original data while being independent. For each subset, lagged variables (y_{t-1}, y_{t-2}) are recomputed, and regression models are trained on 80% of the subset and tested on the remaining 20%. The MSE is calculated for each subset, and the average across subsets represents the MSE for that iteration of thinned cross-validation.

In **Regular Cross-Validation**, the response variable is split into 5 folds using standard k-fold cross-validation. For each fold, the training set consists of data from 4 folds, while the remaining fold is used for testing. A regression model is trained using the lagged predictors, and the MSE is calculated for the test fold. The process is repeated for all 5 folds, and the average MSE for each iteration is computed.

After performing 100 iterations for both methods, the results are summarized and visualized. The following Violin plots compare the distributions of MSE values highlighting that Data-Thinned Cross-Validation demonstrates lower MSE compared to Regular Cross-Validation.

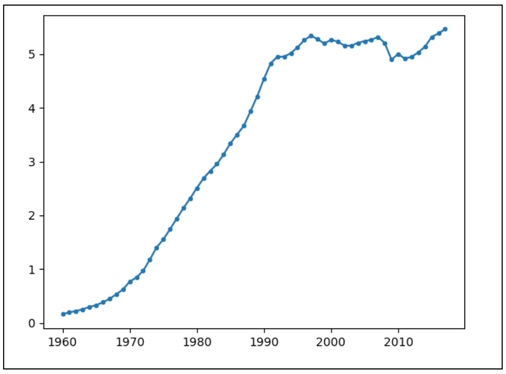
This study underscores the potential of data thinning as an alternative cross-validation approach, especially for time-series or dependent data, where preserving distributional and temporal properties is critical for robust model evaluation. By systematically comparing both methods, the analysis highlights the strengths and weaknesses of each in the context of time-series regression tasks.

7 Data Analysis

Following the paper, we tried to replicate the change-point detection analysis for two separate datasets - **(i) Japan GDP Dataset**, **(ii) LGA Airline Passengers** and see corresponding performances of data thinning procedures in this arena. These two data are consciously considered because of their differing sizes - One (Japan GDP data) representing small sample, while the other (LGA Airline Passenger) being a relatively larger dataset.

7.1 Changepoint Detection Analysis 1: Japan GDP Dataset

This analysis applies generalized data thinning to detect changepoints in Japan's GDP over time, as modeled in the local currency unit (LCU).



The GDP dataset shows a structural break corresponding to the ”lost decade,” a significant period of economic stagnation in Japan. Changepoint detection is used to identify shifts in the underlying variance of GDP growth.

Using the generalized data thinning approach, the time-series GDP data is split into independent components based on a sufficient statistic (e.g., the squared deviation from the mean). This thinning process generates independent subsets of the data, allowing unbiased detection of changepoints by separating the estimation and validation phases of the analysis.

The results (as shown in Figure 3) reveal three changepoints in the GDP time series. The naive method, which does not use data thinning, identified three changepoints, with two showing p-values below the significance threshold ($p < 0.05$). The order-preserved sample splitting method also detected three changepoints, but only one of them was significant. Lastly, the generalized data thinning method identified three changepoints, none of which had p-values below the significance threshold. These results indicate that the naive method likely suffers from inflated Type I error rates, resulting in false positives, while generalized data thinning effectively mitigates this issue. By implementing the generalized data thinning technique, this study confirms its robustness in detecting genuine structural breaks in time-series data, such as GDP growth, without the risk of overfitting or false discoveries. This highlights its utility for applications in econometrics and time-series modeling, even when there are limited data points.

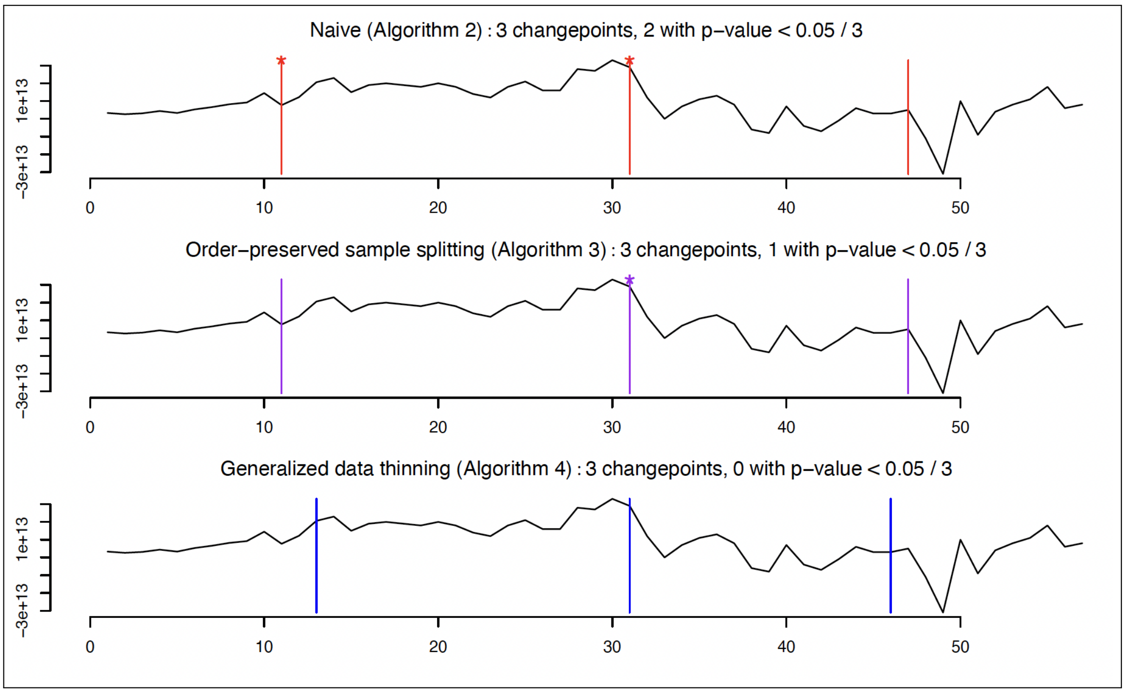
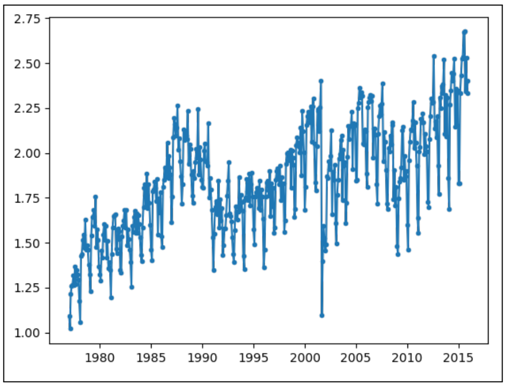


Figure 3: Change-Point Detection for Japan GDP Data

7.2 Changepoint Detection Analysis 2: LGA Airline Passengers

This dataset consists of monthly international passenger counts at LGA (LaGuardia Airport) from 1977 onward, capturing fluctuations in airport activity over several decades. The analysis yielded (Figure 4) multiple changepoints, with distinct methods providing varying results:

- The naive method detected 10 changepoints, of which 7 had p-values below 0.05, indicating statistical significance. However, this method is prone to inflated Type I error rates.
- The order-preserved sample splitting method identified 7 changepoints, with 4 being significant (p ¡ 0.05).
- The generalized data thinning approach detected 9 changepoints, of which 3 were statistically significant. This result demonstrates that while the generalized data thinning method is slightly more conservative, it effectively minimizes false positives compared to the naive approach.



This analysis highlights notable structural changes in passenger activity over time at LGA, reflecting potential external influences such as policy changes, economic factors, or operational adjustments. The results underscore the robustness of generalized data thinning in identifying genuine changepoints in complex time-series data, providing a reliable tool for analyzing fluctuations in transportation metrics.

8 Conclusion

In this project, we explored the utility of generalized data thinning as a robust alternative to traditional data-splitting methods, addressing the limitations posed by small sample sizes and dependent data. By leveraging

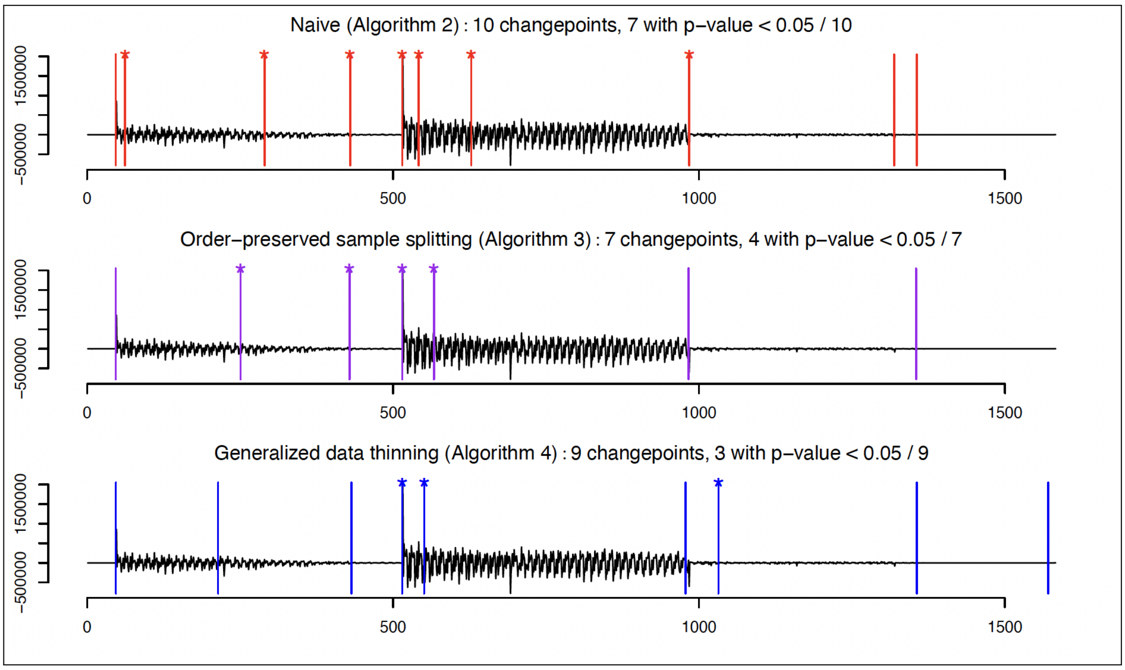


Figure 4: Change-Point Detection for LGA Airport Passengers (International)

sufficient statistics, generalized thinning provides a unified framework for partitioning data into independent components without sacrificing information about the underlying parameters.

Through theoretical exploration and empirical analysis, the study demonstrated the versatility of generalized thinning across multiple settings, including regression analysis, cross-validation, and changepoint detection. In simulation studies, we observed the thinning approach outperformed traditional methods, such as naive data splitting and order-preserved sample splitting, particularly in scenarios with dependent data or limited observations. For instance, in the analysis of Japan's GDP data and LGA Airline Passenger data, generalized thinning identified changepoints more conservatively, reducing false positives compared to the naive approach, while maintaining sufficient sensitivity.

These results underscore the importance of using information-preserving thinning techniques for statistical inference and model validation, particularly in time-series data and other complex settings. Generalized data thinning not only provides a robust method for changepoint detection but also extends to broader applications in econometrics, transportation analysis, and beyond, where maintaining the integrity of data is critical for accurate and reliable conclusions.

References

[1] Yiqun T Chen and Daniela M Witten. Selective inference for k-means clustering. 2022.

[2] D R Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441, August 1975.

[3] Tibshirani R. Friedman J. H. Friedman J. H. Hastie, T. *The Elements of Statistical Learning*. Springer, 2009.

[4] Sean Jewell, Paul Fearnhead, and Daniela Witten. Testing for a change in mean after changepoint detection. 2019.

[5] Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proc. Natl. Acad. Sci. U. S. A.*, 112(25):7629–7634, June 2015.