Generalized Data Thinning using Sufficient Statistics

Netra Prabhu, Ananya Roy

NC State University

December 3,2024

What is Data Thinning?

- A statistical technique where a single data point is split into multiple independent parts.
- **Key Principles:**
  - **Independence:** The thinned random variables $X^{(1)}, \ldots, X^{(K)}$ must be mutually independent.
  - **Sufficiency:** The function $T(X^{(1)}, \ldots, X^{(K)})$ must retain all the information about the unknown parameter(s).
  - **Flexibility:** Works with a broad set of families including exponential and non-exponential distributions.

Need of Data Thinning

Applications:

- **Hypothesis Testing:** Use the data both to generate and to test a hypothesis.
- **Model Validation:** Use the data both to fit a complicated model, and to obtain an accurate estimate of the expected prediction error.
- **Bias/Prediction Error Reduction:** Mitigates biases in scenarios like cross-validation and error estimation.

Existing Methods:

- **Sample Splitting (Cox, 1975):[1]** Divides data into subsets for model fitting and validation but lacks flexibility for complex dependencies.
- **Convolution-Closed Thinning (Neufeld et al., 2023)[3]:** They consider splitting, or thinning, a random variable X drawn from a convolution-closed family into K independent random variables $X^{(1)}, \ldots, X^{(K)}$ such that:

$$X = \sum_{k=1}^{K} X^{(k)},$$

and $X^{(1)}, \ldots, X^{(K)}$ come from the same family of distributions as X.

Generalized Thinning

- Splits a random variable X into K independent random variables $X^{(1)}, \ldots, X^{(K)}$.
- Ensures the following two properties:
  1. $X = T(X^{(1)}, \ldots, X^{(K)})$
  2. $X^{(1)}, \ldots, X^{(K)}$ are mutually independent.
- Simultaneously encompass both convolution-closed data thinning and sample splitting.

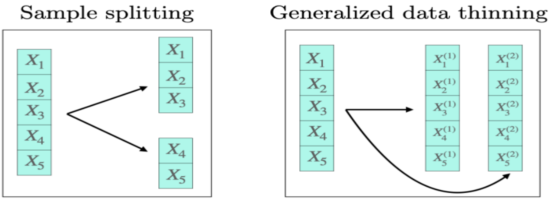


Figure: Left: Sample splitting assigns each observation to either a training or a test set. Right: Generalized data thinning splits each observation into two parts that are independent and that can be used to recover the original observation $T(X^{(1)}, X^{(2)}) = X$. Source: pg 5, paper [2]

Key Contributions:

- Sufficiency is the key property underlying the choice of the function $T()$.
- Generalizes thinning beyond convolution-closed families to broader distribution classes.
- Extends applicability to non-exponential families like Beta and Uniform distributions.
- Demonstrates use cases in scenarios unsuitable for traditional sample splitting.
- Preserves independence and sufficiency, enabling robust model validation and inference.

Generalized Thinning Procedure

Definition 1:

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

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

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

Theorem : 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:

1. $T(X^{(1)}, \ldots, X^{(K)})$ is a sufficient statistic for θ based on $(X^{(1)}, \ldots, X^{(K)})$.
2. 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 $(X^{(1)}, \ldots, X^{(K)}) \sim Q_\theta^{(1)} \times \cdots \times Q_\theta^{(K)}$.

Algorithm for Finding distributions that can be thinned:

1. Choose K families of distributions, $Q^{(k)} = \{Q_\theta^{(k)} : \theta \in \Omega\}$ for $k = 1, \ldots, K$.
2. 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.
3. 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 $T(\cdot)$.

Thinning Natural Exponential Families(NEF)

- NEF starts with a known probability distribution H
- Forms a family of distributions $P_H = \{P_H^\theta : \theta \in \Omega\}$ based on H, as follows:

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

- $\psi_H(\theta)$: Normalizing constant ensuring P_θ is a valid probability distribution.

Thinning by Addition: The natural exponential family P_H can be thinned by $T(x^{(1)}, \dots, x^{(K)}) = \sum_{k=1}^K x^{(k)}$ into K NEFs $P_{H_1}, \dots, P_{H_K}$ if and only if H is the K-way convolution of $H_1, \dots, H_K$.

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

- Example: Gaussian distributions.

$$N(\mu, \sigma^2) \to \text{Split into } N(\epsilon_k \mu, \epsilon_k \sigma^2), \text{ with } \sum \epsilon_k = 1.$$

Thinning Natural into General Exponential Families

- Allows for more flexibility in the sufficient statistic and thinning function.
- Uses the previously defined algorithm.

Proposition:

- 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 $Q^{(k)}$ with sufficient statistic $T^{(k)}(X^{(k)})$.
- P_θ : the distribution of $\sum_{k=1}^{K} T^{(k)}(X^{(k)})$. (sufficient statistic for θ based on $(X^{(1)}, \ldots, X^{(K)}) \sim Q_\theta^{(1)} \times \cdots \times Q_\theta^{(K)}$.)

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:

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

- Implies that many natural exponential families can be thinned by a function of the above form.

Indirect Thinning of General Exponential Families

Key Concept: Consider $X \sim P_\theta \in P$. Suppose we thin a sufficient statistic $S(X)$ for θ by a function $T(\cdot)$. Then, we say that the family P is **indirectly thinned** through $S(\cdot)$ by $T(\cdot)$.

Indirect Thinning of General Exponential Families:

- Let $P = \{P_\theta : \theta \in \Omega\}$ be a general exponential family. That is, $dP_\theta(x) = \exp\{[S(x)]^\top \eta(\theta) - \psi(\theta)\} dH(x)$,
- $S(X)$: sufficient for θ, $S(X)$ belongs to a NEF (Lehmann & Romano 2005).

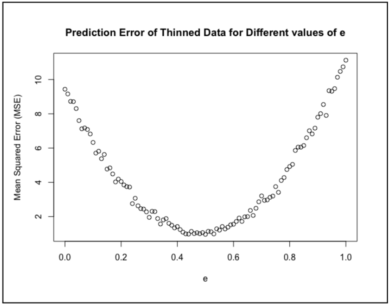
Thus, indirectly thinning X through $S(\cdot)$ as follows:

1. We now consider $X^{(1)}, \dots, X^{(K)}$ that belong to a general exponential family, where $T^{(k)}(\cdot)$ is not necessarily the identity. Suppose further that: $S(X) \overset{D}{=} \sum_{k=1}^K T^{(k)}(X^{(k)})$.
2. Then, indirectly thinning X through $S(\cdot)$ into $X^{(1)}, \dots, X^{(K)}$, by:

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

Simulation

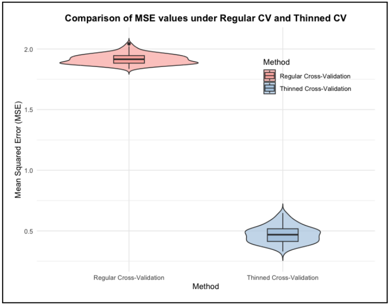
1. Optimal value of ϵ for Normal Distribution Thinning:



$X \sim N(\mu, \sigma^2)$ thinned into -

1. $X^{(1)} \sim N(\epsilon\mu, \epsilon\sigma^2)$
2. $X^{(2)} \sim N((1-\epsilon)\mu, (1-\epsilon)\sigma^2)$

2. Comparison of Thinned Cross-Validation and Regular Cross-Validation methods:

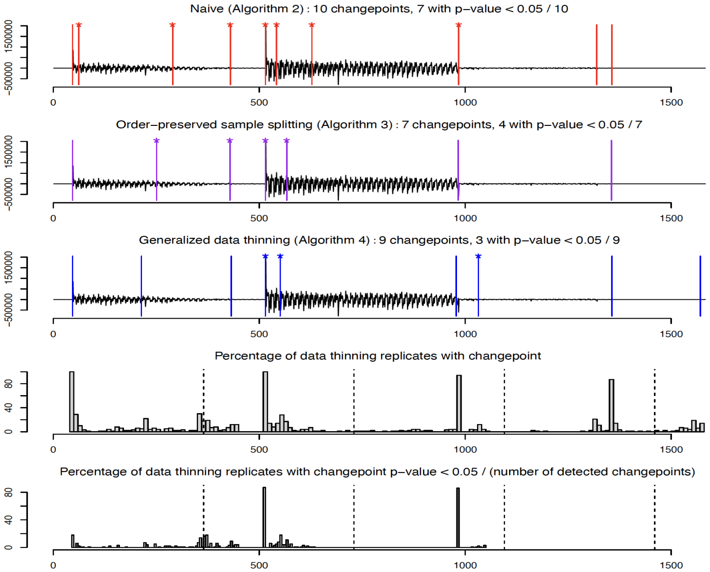


Response variable $y = sin(x) + E$,
$E \sim N(0, 1)$(noisy sinusoidal time-series).

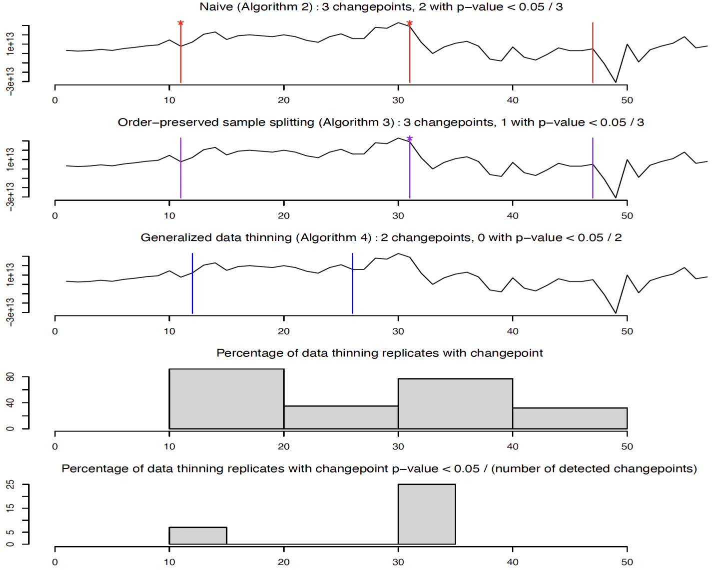
- **Regular CV:** with k=5 folds
- **Thinned CV:** Thin data into $K = 3$ subsets using the `datathin` package, ensuring independence. Perform an 80-20 train-test split within each subset and average MSE across all subsets.

Data Analysis - Change-point detection

LGA Airline Passenger Dataset : Gives the number of passengers arriving and departing at LGA.



GDP of Japan Dataset : Historic GDP of Japan in the Local Currency Unit (LCU)



Limitations

- NEF that is based on a distribution that cannot be written as the convolution of two distributions.
- Convolution-closed family outside of the NEF in which addition is not sufficient.

Examples:

- Bernoulli family cannot be thinned by any function $T()$.
- Cauchy family cannot be thinned by addition.

Conclusions

- **Generalized Framework:**
  - Thins random variables X into independent components $X^{(1)}, \ldots, X^{(K)}$.
  - Preserves all information about unknown parameters using sufficiency.
- **Unified Perspective:**
  - Links sample splitting and data thinning as special cases of a broader framework.
  - Expands to new distributions, including beta, uniform, and shifted exponential.
- **Applications:**
  - Improves model validation, selective inference, and changepoint detection.
  - Provides robust solutions for dependent or small datasets.

References I

[1] D R Cox. "A note on data-splitting for the evaluation of significance levels". In: *Biometrika* 62.2 (Aug. 1975), p. 441.

[2] Ameer Dharamshi et al. "Generalized data thinning using sufficient statistics". In: *Journal of the American Statistical Association* just-accepted (2024), pp. 1–26.

[3] Anna Neufeld et al. "Data thinning for convolution-closed distributions". In: *Journal of Machine Learning Research* 25.57 (2024), pp. 1–35.

Thank you!