Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Assumption-Lean Data Fission with Resampled Data

Kevin Fry, Snigdha Panigrahi & Jonathan Taylor

To cite this article: Kevin Fry, Snigdha Panigrahi & Jonathan Taylor (2025) Assumption-Lean Data Fission with Resampled Data, Journal of the American Statistical Association, 120:549, 161-161, DOI: 10.1080/01621459.2024.2412172

To link to this article: https://doi.org/10.1080/01621459.2024.2412172

Published online: 14 Apr 2025.

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 120, NO. 549, 161: Theory and Methods Discussion
https://doi.org/10.1080/01621459.2024.2412172

Assumption-Lean Data Fission with Resampled Data

Kevin Fry[a], Snigdha Panigrahi[b], and Jonathan Taylor[a]

[a]Department of Statistics, Stanford University, Stanford, CA; [b]Department of Statistics, University of Michigan, Ann Arbor, MI

Gaussian additive noise data fission is described in Tian and Taylor (2018) and expanded in Rasines and Young (2022). This current work, Leiner et al. (2025), expands greatly on this idea by creating two data sources from a single source, allowing information from each sample to be used for both selection and inference. Data fission replaces independence with slightly different but still strong conditional "tractability" properties. Due to this "tractable" dependence requirement, many of the examples are parametric in nature.

In Panigrahi, Fry, and Taylor (2024) (see Section A.5), it was shown that when a bootstrap is available, it is quite simple to establish an asymptotic form of fission. The method is simple to describe and is broadly applicable under assumptions in which the resampling method have suitably well-developed CLT properties.

Suppose the analyst has access to a resampling procedure with a concentration parameter κ. Formally, this resampling procedure should satisfy the condition that for any $\kappa > 0$, the analyst can resample a functional $\hat{\theta}^*$ such that

$$\hat{\theta}^* | \hat{\theta} \dot{\sim} N(\hat{\theta}, \kappa \cdot \text{var}(\hat{\theta}))$$

and have a CLT

$$\hat{\theta} \dot{\sim} N(\theta_0, \text{var}(\hat{\theta})).$$

A canonical example is the κn-out-of-n bootstrap, or the parametric bootstrap. In this example, denote bootstrappable functionals by $\hat{\theta}^* = \hat{\theta}(X^*, Y^*)$ based on the bootstrap sample (X^*, Y^*). The analyst selects a collection of target functionals $S^*(X^*, Y^*)$ using some collection $\{\hat{\theta}^*_j : j \in \mathcal{J}\}$, and wishes to mimic sample splitting with a proportion π of the data used for selection and $1 - \pi$ for inference.

Standard bootstrap variance calculations show that

$$\begin{pmatrix} \hat{\theta} \\ \hat{\theta}^* \end{pmatrix} \dot{\sim} N\left(\begin{pmatrix} \theta_0 \\ \theta_0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1+\kappa \end{pmatrix} \otimes \text{var}(\hat{\theta}) \right).$$

A straightforward consequence of this is that for any target set S

$$\hat{\theta}^{\perp}_S = \hat{\theta}_S + \frac{1}{\kappa} \cdot (\hat{\theta}_S - \hat{\theta}^*_S)$$

is (asymptotically) independent *of all* bootstrap functionals of (X^*, Y^*) and has marginal variance $\left(1 + \frac{1}{\kappa}\right)$ times the sampling variance of $\hat{\theta}_S$. In particular, if S^* is measurable with respect to bootstrappable functionals, then for any S in the range of S^*, we have

$$\hat{\theta}^{\perp}_S | S^*(X^*, Y^*) = S \dot{\sim} N\left(\theta_{0,S}, \left(1 + \frac{1}{\kappa}\right) \cdot \text{var}(\hat{\theta}_S)\right).$$

As a concrete example, if we set $\pi = \frac{4}{5}$ and $\kappa = \frac{1-\pi}{\pi} = \frac{1}{4}$, then

$$\text{var}(\hat{\theta}^{\perp}_S) = 5 \cdot \text{var}(\hat{\theta})$$

This is precisely the form of the variance one would expect to see under data splitting with 4/5 of the data used for selection and 1/5 used for inference. The asymptotic variance $\text{var}(\hat{\theta}_S)$ can of course be estimated from several bootstrap draws.

This method is attractive in that it offers asymptotic data fission without any strong model assumptions, and instead relies on the existence of a resampling procedure satisfying a CLT with concentration parameter κ.

References

Leiner, J., Duan, B., Wasserman, L., and Ramdas, A. (2025), "Data Fission: Splitting a Single Data Point," *Journal of the American Statistical Association*, this issue, DOI: 10.1080/01621459.2023.2270748. [161]

Panigrahi, S., Fry, K., and Taylor, J. (2024), "Exact Selective Inference with Randomization," Biometrika, asae019. [161]

Rasines, D. G., and Young, G. A. (2022), "Splitting Strategies for Post-Selection Inference," *Biometrika*, 110, 597–614. [161]

Tian, X., and Taylor, J. (2018), "Selective Inference with a Randomized Response," *The Annals of Statistics*, 46, 679–710. [161]

CONTACT Kevin Fry *kfry@stanford.edu* Department of Statistics, Stanford University, Stanford, CA.