Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

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

Discussion on "Data Fission: Splitting a Single Data Point"

Zhigen Zhao

To cite this article: Zhigen Zhao (2025) Discussion on "Data Fission: Splitting a Single Data Point", Journal of the American Statistical Association, 120:549, 178-179, DOI: 10.1080/01621459.2024.2416913

To link to this article: https://doi.org/10.1080/01621459.2024.2416913

Published online: 14 Apr 2025.

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 120, NO. 549, 178–179: Theory and Methods Discussion
https://doi.org/10.1080/01621459.2024.2416913

Discussion on "Data Fission: Splitting a Single Data Point"

Zhigen Zhao

Department of Statistics, Operations, and Data Science, Temple University, Philadelphia, PA

I would like to commend the authors for this insightful and thought-provoking paper on "Data Fission: Splitting a Single Data Point." The idea of splitting a data point using random noise to create two parts with specific distributional properties is both innovative and widely applicable, particularly in post-selection inference. The purpose of post-selection inference is to draw conclusions about the parameters selected through a statistical procedure, denoted $S(\mathbf{y})$. The selection set can be chosen based on the full dataset, data splitting, or data fission. In this discussion, we would like to comment on the interplay between various selection rules and post-selection inference methods from the perspective of the Bayes/empirical Bayes framework.

In post-inference problems, the main challenge is dealing with "selection bias," or the "winner's curse." For example, consider the normal mean problem discussed in Section 3 of this article,

$$y_i \sim N(\mu_i, \sigma^2), i = 1, 2, \ldots, p. \quad (1)$$

The parameter corresponding to the largest statistic, $y_{(n)}$, is of particular interest and is selected for further inference. Directly using the largest observed value, $y_{(n)}$, tends to overestimate this selected parameter, leading to low coverage probabilities when constructing confidence intervals of the form $y_{(n)} \pm z_{\alpha/2}\sigma$. Benjamini and Yekutieli (2005) addressed this issue by adjusting the confidence interval to account for the number of selected parameters, $R = |S(\mathbf{y})|$, and using a modified confidence coefficient, $\alpha R/p$, which tends to be conservative, especially when R is small. In contrast, Dawid (1994) argued that Bayesian inference is immune to selection bias, as conditioning on the selection becomes redundant when the conditioning already applies to the full dataset. In particular, when assuming the prior distribution

$$\mu_i \sim N(\theta, \delta^2), \quad (2)$$

it is seen that

$$\mu_i|\mathbf{y}, S(\mathbf{y}) = \mu_i|\mathbf{y} \sim N(My_i + (1-M)\theta, M\sigma^2),$$
$$M = \frac{\delta^2}{\sigma^2 + \delta^2}.$$

The interval $My_i + (1-M)\theta \pm z_{\alpha/2}\sqrt{M\sigma^2}$ ensures a good coverage probability for μ_i for any $i \in S(\mathbf{y})$, given this specific choice of θ and δ^2. When these two hyperparameters are unknown, Zhao and Hwang (2012) and Hwang and Zhao (2013) estimated them from the data and derived empirical Bayes confidence intervals such that

$$\mathrm{EP}(\mu_i \in CI_{EB}|\mathbf{y}) \le 1-\alpha, \text{ for any normal prior } N(\theta, \delta^2).$$

In this article, the authors introduce the concept of data fission, where the data is split into $f(y_i)$ and $g(y_i)$, with $f(y_i)$ used for variable selection and $g(y_i)$ for further inference. For the normal mean model discussed above, the data is split as $f(y_i) = y_i + \tau z_i$ and $g(y_i) = y_i - \frac{1}{\tau}z_i$, where $z \sim N(0, \sigma^2)$ is used for inference. The constructed confidence interval is given as

$$g(y_i) \pm z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{\tau^2}}.$$

As demonstrated in this article, these intervals provide good coverage probabilities when $S(\mathbf{y})$ depends solely on $f(y_i)$, which is independent of $g(y_i)$. However, it remains unclear how the method performs under arbitrary selection rules.

I would like to propose an empirical Bayes interpretation of this approach. Assuming a prior distribution of

$$\mu_i \sim N(\theta, \delta^2), \pi(\theta) \propto 1. \quad (3)$$

Then it is seen that

$$\mathrm{E}(\mu_i|y_i, \theta) = y_i + (1-M)(\theta - y_i),$$

Note that $y_i|\theta \sim N(\theta, \sigma^2 + \delta^2)$. If we replace θ with $y_i - \frac{\sigma^2+\delta^2}{\sigma^2}z_i$ where $z \sim N(0, \sigma^2)$, then

$$\mathrm{E}(\mu_i|y_i, \theta) \approx y_i - \frac{\sigma^2}{\delta^2 + \sigma^2}z_i,$$

CONTACT Zhigen Zhao zhaozhg@temple.edu Department of Statistics, Operations, and Data Science, Temple University, Philadelphia, PA 19121.

which can be viewed as $g(y)$ in the data fission approach when setting $\tau = \sqrt{1 + \frac{\delta^2}{\sigma^2}}$. Based on this connection and the argument of Dawid (1994), it seems that the intervals based on $g(y)$ would provide a valid method for an arbitrary selection rule when assuming the prior distribution (3). In summary,

- When $S(\mathbf{y})$ is chosen arbitrarily, Benjamini and Yekutieli (2005) (BY) method works for any prior distribution, though it requires a longer interval. A much shorter empirical Bayes confidence interval (CI) proposed in Zhao and Hwang (2012) and Hwang and Zhao (2013) guarantees good coverage probabilities for the normal prior (2).
- When $S(\mathbf{y})$ is chosen arbitrarily, the above discussion suggests that the interval centered around $g(y)$ and its empirical Bayesian counterpart could be valid for the class of prior distributions (3) that is broader than (2).
- If $S(\mathbf{Y})$ depends on $\mathbf{y}$ through the $f(y_i)$'s which are independent of $g(y_i)$'s, then inference based on the $g(y_i)$'s is valid for any prior distribution.

The connection between the proposed approach and the Bayes/empirical Bayes framework offers valuable insights into its generalization and potential extensions. I would like to once again congratulate the authors on this exciting paper. I look forward to the authors' rejoinder and to seeing further developments in this area.

Funding

Zhigen Zhao is partially supported by the NSF Grant DMS-2311216.

References

Benjamini, Y., and Yekutieli, D. (2005), "False Discovery Rate-Adjusted Multiple Confidence Intervals for Selected Parameters," *Journal of the American Statistical Association*, 100, 71–93. [178,179]

Dawid, A. P. (1994), "Selection Paradoxes of Bayesian Inference," *Multivariate Analysis and Its Applications, Lecture Notes-Monograph Series*, 24, 211–220. [178,179]

Hwang, J. T., and Zhao, Z. (2013), "Empirical Bayes Confidence Intervals for Selected Parameters in High Dimension with Application to Microarray Data Analysis," *Journal of the American Statistical Association*, 108, 607–618. [178,179]

Zhao, Z., and Hwang, J. T. (2012), "Empirical Bayes False Coverate Rate Controlling Confidence Interval," *Journal of the Royal Statistical Society*, Series B, 74, 871–891. [178,179]