Journal of the
American
Statistical
Association

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

# Comments on “Data Fission: Splitting a Single Data Point”

Sanat K. Sarkar

**To cite this article:** Sanat K. Sarkar (2025) Comments on “Data Fission: Splitting a Single Data Point”, Journal of the American Statistical Association, 120:549, 170-171, DOI: 10.1080/01621459.2024.2414865

**To link to this article:** https://doi.org/10.1080/01621459.2024.2414865

Published online: 14 Apr 2025.

Submit your article to this journal

Article views: 19

View related articles

View Crossmark data

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 120, NO. 549, 170–171: Theory and Methods Discussion
https://doi.org/10.1080/01621459.2024.2414865

Taylor & Francis
Taylor & Francis Group

# Comments on “Data Fission: Splitting a Single Data Point”

Sanat K. Sarkar

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

I commend the authors for presenting a groundbreaking work that transcends traditional data-splitting by introducing the concept of data fission and its application in selective inference. They have shown that data fission can be effectively achieved through carefully chosen external randomization, depending on the underlying distributional assumptions. This novel idea holds great promise for advancing research into innovative, statistically sound, and computationally efficient data-driven strategies for selective inference.

In my comments, I will outline several potential pathways for future research on testing Gaussian means against two-sided alternatives in line with this article.

First, I will build upon the work of Sarkar and Tang (2022) by adapting their selective inference framework using knockoffs for testing regression coefficients in a fixed-design model, applying similar principles to testing Gaussian means through data fission. Additionally, I will propose a new research direction focused on testing or interval estimation of Gaussian means within a selective inference framework, incorporating recent developments from Sarkar and Zhang (in press).

Second, for testing Gaussian means with unknown variances in a selective inference framework, I will present a method for fissioning the two-sided $t$-test or the corresponding $\beta$-test statistics.

Let $X \sim N_d(\mu, \Sigma)$, where $\mu = (\mu_1, \ldots, \mu_d)$ is unknown and $\Sigma$ is a known pd matrix. For any $\Delta = \text{diag}\,\{\delta_i\}$, for which $\Sigma^{-1} - \Delta$ is positive definite, let us decompose $X$ as follows using an independently drawn $Z \sim N_d(0, I_d)$:

$$\begin{aligned} f(X) &= X + (\Sigma^{-1} - \Delta)^{-1}(\Delta - \Delta\Sigma\Delta)^{\frac{1}{2}} Z \\ g(X) &= X - \Delta^{-1}(\Delta - \Delta\Sigma\Delta)^{\frac{1}{2}} Z, \end{aligned}$$

Then, $f(X)$ and $g(X)$ are jointly distributed as Gaussian with

$$\begin{aligned} E(f(X)) &= E(g(X)) = \mu, \\ \text{cov}(f(X)) &= \Sigma + (\Sigma^{-1} - \Delta)^{-1}(\Delta - \Delta\Sigma\Delta) \\ &\quad\ (\Sigma^{-1} - \Delta)^{-1} \\ &= \Sigma + \Sigma\Delta(\Sigma^{-1} - \Delta)^{-1} \\ &= [\Sigma(\Sigma^{-1} - \Delta) + \Sigma\Delta](\Sigma^{-1} - \Delta)^{-1} \\ &= (\Sigma^{-1} - \Delta)^{-1} \\ \text{cov}(g(X)) &= \Sigma + \Delta^{-1}(\Delta - \Delta\Sigma\Delta)\Delta^{-1} \\ &= \Sigma + \Delta^{-1} - \Sigma = \Delta^{-1} \\ \text{cov}(f(X), g(X)) &= \Sigma - (\Sigma^{-1} - \Delta)^{-1}(\Delta - \Delta\Sigma\Delta)\Delta^{-1} \\ &= \Sigma - \Sigma\Delta\Delta^{-1} = O, \end{aligned}$$

The decomposition discussed here extends one of the Gaussian examples given in the paper under discussion, but it now ensures that $g(X)$ exhibits internal independence. This feature is crucial as it allows for the implementation of theoretically valid false discovery rate (FDR) controlled multiple testing, as outlined in Sarkar and Tang (2022), and enables false coverage rate (FCR) controlled interval estimation for the $\mu_i$'s (as in the paper under discussion) after their selection using $f(X)$.

With the introduction of a theoretically valid FDR controlling method, such as the dependence-adjusted Benjamini-Hochberg (DBH) procedure mentioned in Fithian and Lei (2022), one can apply this approach during the selection phase using $f(X)$. Alternatively, one could reverse the roles of $f(X)$ and $g(X)$, using $f(X)$ to test hypotheses while ensuring FDR control after their selection using $g(X)$. This strategy may provide a more nuanced exploitation of the underlying correlation structure, potentially increasing efficiency in the selective inference framework.

However, it is important to note that the DBH procedure involves a complex algorithm that produces a calibrated threshold for each hypothesis, which may not be straightforward for users compared to the original Benjamini-Hochberg method. To address this, the recently introduced Shifted Benjamini-Hochberg (SBH) method, as detailed in Sarkar and Zhang (in press), offers a simpler alternative. The SBH method (see below) leverages partial information from the correlation matrix, such as multiple correlations or eigenvalues, trading off some statistical power for the sake of maintaining the user-friendly nature of the original BH procedure. This method could lead to more computationally efficient strategies for selective inference concerning Gaussian means when the correlation structure is known.

CONTACT Sanat K. Sarkar sanat@temple.edu Department of Statistics, Operations, and Data Science, Fox School of Business, Temple University, 1810 Liacouras Walk, 3rd Floor, Philadelphia, PA 19122.

Shifted BH Method. *Given $X = (X_1, \ldots, X_d)' \sim N_d(\mu, \Sigma)$, where $\Sigma$ is a known correlation matrix, consider testing $H_i : \mu_i = 0$ against $\mu_i \neq 0$ simultaneously for $i = 1, \ldots, d$. Let $\tau_i = 1 - R_i^2$, where $R_i^2$ is the squared multiple correlation between $X_i$ and $X_{-i}$, or $\tau_i = \tau$, for some $\tau \in (0, \eta)$, where $\eta$ is the minimum eigen value of $\Sigma$. With $\Psi = 1 - \bar{\Psi}$ denoting the cdf of $\chi_1^2$ distribution, define $\tilde{P}_i = \bar{\Psi}\left(X_i^2/\tau_i\right)$, for $i = 1, \ldots, d$. Sort the $\tilde{P}_i$'s as $\tilde{P}_{(1)} \leq \cdots \leq \tilde{P}_{(d)}$, and find $R = \max\{i : \tilde{P}_{(i)} \leq i\alpha_1\}$, where $\sum_{i=1}^{d} \bar{\Psi}\{\tau_i \bar{\Psi}^{-1}(\alpha_1)\} = \alpha$. Reject $H_i$ for all i such that $\tilde{P}_i \leq \tilde{P}_{(R)}$.*

Now, suppose $Y = (Y_1, \ldots, Y_d)' \sim N_d((\mu_1, \ldots, \mu_d), \mathrm{diag}\{\sigma_i^2\})$, where $\sigma_i^2$ is unknown and estimated unbiasedly using $V_i \sim \sigma_i^2 \chi_n^2$, which is independent of $Y_i$, for each $i = 1, \ldots, d$. Let $X_i = Y_i^2/(Y_i^2 + V_i)$, one-to-one transformation of the usual $t^2$ test statistic, be the test statistic that is used to marginally test $\mu_i = 0$ against $\mu_i \neq 0$. The following decomposition of each $X_i$ represents fissioned versions of $X$ and can provide selective inference frameworks for the $\mu_i$'s:

Draw $Z_1, \ldots, Z_d$ independently from Beta $((d-1)/2, (n-d+1)/2)$ distribution, for any fixed $1 < d < n+1$. Define $f(X_i) = Z_i + (1 - Z_i)X_i$ and $g(X_i) = X_i/[Z_i + (1 - Z_i)X_i]$. The $f(X_i)$ and $g(X_i)$ are independent, conditional on $(J_1, \ldots, J_d)$, where $J_i \sim \mathrm{Poisson}(\mu_i^2/2\sigma_i^2)$. Moreover, $f(X_i) \sim$ Beta $((d + 2J_i)/2, (n-d+1)/2)$ and $g(X_i) \sim$ Beta $((1 + 2J_i)/2, (d-1)/2)$. This can be checked by noting that

$$f(X_i) = \frac{Y_i^2 + V_i Z_i}{Y_i^2 + V_i Z_i + V_i(1 - Z_i)} \text{ and } g(X_i) = \frac{Y_i^2}{Y_i^2 + V_i Z_i},$$

with $Y_i^2$, $V_i Z_i$ and $V_i(1 - Z_i)$ being independently distributed as $\chi^2_{1+2J_i}$, $\chi^2_{d-1}$, and $\chi^2_{n-d+1}$, respectively, conditionally given $J_i$. Larger (or smaller) $Z_i$ indicates that $f(X_i)$ is more (or less) informative.

Simultaneous inference for the $\mu_i$'s, in terms of testing $\mu_i = 0$ (i.e., $J_i = 0$) against $\mu_i \neq 0$ (i.e., $J_i \neq 0$), or in terms of interval estimation, can be carried out in a selective inference framework with $f(X)$ considered for selection and $g(X)$ for inference.

## Disclosure Statement

No potential conflict of interest was reported by the author(s).

## Funding

The research is supported by NSF grant DMS 2210687.

## References

Fithian, W., and Lei, L. (2022), "Conditional Calibration for False Discovery Rate Control Under Dependence," *Annals of Statistics*, 50, 3091–3118. [170]

Sarkar, S. K., and Tang, C.-Y. (2022), "Adjusting the Benjamini-Hochberg Method for Controlling the False Discovery Rate in Knockoff Assisted Variable Selection," *Biometrika*, 109, 1149–1155. [170]

Sarkar, S. K., and Zhang, S. (in press), "Shifted BH Methods for Controlling False Discovery Rate in Multiple Testing of the Means of Correlated Normals Against Two-Sided Alternatives," *Journal of Statistical Planning & Inference*. [170]