Journal of the American Statistical Association

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

Discussion of "Data Fission: Splitting a Single Data Point" – Some Asymptotic Results for Data Fission

Lihua Lei

To cite this article: Lihua Lei (2025) Discussion of "Data Fission: Splitting a Single Data Point" – Some Asymptotic Results for Data Fission, Journal of the American Statistical Association, 120:549, 147-150, DOI: 10.1080/01621459.2024.2441416

To link to this article: https://doi.org/10.1080/01621459.2024.2441416

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 120, NO. 549, 147–150: Theory and Methods Discussion
https://doi.org/10.1080/01621459.2024.2441416

Discussion of "Data Fission: Splitting a Single Data Point" – Some Asymptotic Results for Data Fission

Lihua Lei

Graduate School of Business and Department of Statistics, Stanford University, Stanford, CA

This comment was presented as part of the JSM 2024 *JASA* Theory and Methods invited session Data Fission: Splitting a Single Data Point.

1. Introduction

Leiner et al. (2025) introduce the data fission technique, which unifies and generalizes multiple methods in selective inference. We congratulate the authors on this impactful work and elegant idea. For a broad class of parametric distributions, it offers a general recipe by exploiting a clever analogy to Bayesian inference to split a single data point W into two components $f(W)$ and $g(W)$, such that the marginal distribution of $f(W)$ and the conditional distribution of $g(W)$ given $f(W)$ are both known distributions, up to the knowledge of the parameter. This property allows the researcher to use $f(W)$ for model selection and $g(W)$ for statistical inference, without the need to adjust for selection.

Is the data fission procedure sensitive to parametric assumptions? In this discussion, we make an attempt in answering this question for fixed-design regressions with non-Gaussian errors and unknown error variance. To set up the notation, we let $X \in \mathbb{R}^{n\times p}$ be a fixed matrix and

$$Y = \mu + \epsilon \in \mathbb{R}^n, \quad \epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} F, \quad \text{var}(\epsilon_i) = \sigma^2 > 0, \quad (1)$$

where the error variance σ^2 is unknown. Let $\hat{\sigma}^2$ be any estimate of σ^2. For instance, we can set $\hat{\sigma}^2$ to be the residual sum of squares from the regression of Y on X, that is,

$$\hat{\sigma}^2 = \frac{1}{n-p} Y^T (I - X(X^T X)^{-1} X^T) Y. \quad (2)$$

We define the data fission procedure as follows, where $\tau > 0$ and $(X, Y, \hat{\sigma})$ are user inputs.

Step 1. Sample $Z \sim N(0, I_n)$;
Step 2. Compute $f(Y) = Y + \tau\hat{\sigma}Z, g(Y) = Y - \tau^{-1}\hat{\sigma}Z$;
Step 3. Select a subset $M \subset \{1, \ldots, p\}$ based on $f(Y)$;
Step 4. Choose a vector η_M that depends on M in the column span of X;
Step 5. Construct a confidence interval for $\eta_M^T \mu$ as $[\eta_M^T g(Y) \pm z_{1-\alpha/2}\sqrt{1+\tau^{-2}}\hat{\sigma}\|\eta_M\|]$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution.

In Step 4, a common option is $\eta_M = X_M(X_M^T X_M)^{-1} e_j$ where X_M is the submatrix of X formed by columns in M and e_j is the jth canonical basis in $\mathbb{R}^{|M|}$. Our result can be easily generalized to multidimensional parameters but we focus on unidimensional parameters for ease of exposition.

Let $e(Y, Z, \hat{\sigma})$ be the indicator that the confidence interval in Step 5 covers the true parameter $\eta_M^T \mu$, that is,

$$e(Y, Z, \hat{\sigma}) = I\left(\eta_M^T \mu \in [\eta_M^T g(Y) \pm z_{1-\alpha/2}\sqrt{1+\tau^{-2}}\hat{\sigma}\|\eta_M\|]\right).$$

Our target is to show the unconditional coverage is (asymptotically) at least $1-\alpha$, that is,

$$\mathbb{E}[e(Y, Z, \hat{\sigma})] \geq 1 - \alpha + o(1). \quad (3)$$

Leiner et al. (2025) offer a heuristic argument in Appendix B.5 for the case where μ is linear in X, ϵ_i is Gaussian, and σ^2 is estimated using the method described above. However, it is hard to make the argument rigorous because $f(Y)$ and $g(Y)$ are n-dimensional vectors and a slight estimation error in σ would result in a large deviation in the joint distribution. Rasines and Young (2023) establish a central limit theorem of $\eta_M^T g(Y)$ conditional on $M = S$ for a given model S, provided that the probability of selecting S is not too small and σ^2 can be estimated consistently with data splitting. While it implies the conditional coverage for each given S satisfying the condition, the lack of uniformity makes it difficult to apply their result to prove the unconditional guarantee (3). Moreover, they require the selection rule to be convex, that is, the selection event $\{y : M(X, y) = S\}$ is convex for any $S \subset \{1, \ldots, p\}$. However, this event is often a union of convex sets or "too complicated to be explored analytically"(Rasines and Young 2023).

In this discussion, we study a broad class of selection rules that depends on $f(Y)$ through $X^T f(Y)$, which includes many practical selection rules. We prove the asymptotic coverage guarantee (5) when $\hat{\sigma}^2$ is consistent (Theorem 2.1), and establish a uniform lower bound when $\hat{\sigma}^2$ is asymptotically conservative (Theorem 2.2). Both results allow the dimension to grow with the sample size. Furthermore, we explore how additional restrictions on the selection rule can be leveraged to improve the bounds. In particular, when the selection events are simple convex sets, as defined in Chernozhukov, Chetverikov, and Kato (2017), we can allow the dimension to grow linearly in n in certain cases. By contrast, when the selection events are merely assumed to be convex as in Rasines and Young (2023), we are unable to achieve better dimension dependence than those obtained for selection rules without further constraints.

CONTACT Lihua Lei *lihualei@stanford.edu* Graduate School of Business and Department of Statistics, Stanford University, Stanford, CA.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

All technical proofs are presented in Appendix A of online supplementary material.

2. Main Results

Throughout the commentary we impose the following constraint on the selection rule.

Assumption 2.1. The selected subset M depends on $f(Y)$ through $X^T f(Y)$.

This includes a large class of selection procedures. For example, all regularized regression in the following form, including LASSO, satisfies Assumption 2.1:

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\|f(Y) - X\beta\|^2 + \lambda\rho(\beta)$$
$$\iff \min_{\beta\in\mathbb{R}^p} \beta^T\left(\frac{X^TX}{n}\right)\beta - \frac{2}{n}\beta^T(X^Tf(Y)) + \lambda\rho(\beta). \quad (4)$$

Since the Gram matrix is fixed, the minimization problem only depends on $X^Tf(Y)$. Similarly, the Forward Stepwise selection also satisfies Assumption 2.1. The fixed-X Knockoffs procedure (Barber and Candès 2015), or its uniformly improved variant (Luo, Fithian, and Lei 2022), selects variables based on $X^Tf(Y)$ and $\tilde{X}^Tf(Y)$ where $\tilde{X}$ is the Knockoffs matrix. We can expand the design matrix to include $\tilde{X}$ to derive asymptotics.

Next, we assume that $\hat{\sigma}^2$ is asymptotically upwardly-biased.

Assumption 2.2. There exists $\sigma_+^2 \geq \sigma^2$ such that $\hat{\sigma}^2 = \sigma_+^2 + o_{\mathbb{P}}(h_n)$ for some deterministic sequence $h_n \to 0$.

For the estimator (2), standard calculation implies

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 + \frac{1}{n-p}\|(I - X(X^TX)^{-1}X^T)\mu\|^2 \geq \sigma^2.$$

When μ is in the span of X, as in standard linear models, $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$. With a finite fourth moment, we show in Appendix B.1 that Assumption 2.2 holds with $\sigma_+^2 = \mathbb{E}[\hat{\sigma}^2]$ and any sequence h_n with $\sqrt{n}h_n \to \infty$.

Now we state a generic result that the unconditional coverage $\mathbb{E}[e(Y,Z,\hat{\sigma})]$ changes little if $\hat{\sigma}^2$ is replaced by its limit σ_+^2, up to some minor adjustment. Define $e_{h_n}(Y,Z,\sigma_+)$ as the indicator of coverage when $\hat{\sigma}$ is replaced by σ_+ in Step 2 and by $\sigma_+(1-h_n)$ in Step 5, that is,

$$e_{h_n}(Y,Z,\sigma_+) = I\Big(\eta_{M_+}^T\mu \in [\eta_{M_+}^T g_+(Y) \pm z_{1-\alpha/2}\sqrt{1+\tau^{-2}}\sigma_+(1-h_n)\|\eta_{M_+}\|]\Big), \quad (5)$$

where M_+ is obtained based on $(f(Y),g(Y)) = (f_+(Y),g_+(Y))$ and

$$f_+(Y) = Y + \tau\sigma_+Z, \quad g_+(Y) = Y - \tau^{-1}\sigma_+Z. \quad (6)$$

Lemma 2.1. Under Assumptions 2.1 and 2.2,

$$\mathbb{E}[e(Y,Z;\hat{\sigma})] \geq \mathbb{E}[e_{h_n}(Y,Z;\sigma_+)] - \sqrt{p}h_n + o(1).$$

Under the conditions of Proposition B.1, Lemma 2.1 holds when $p/n = o(1)$. Thus, the effect of unknown variance is negligible if the average sample size per parameter is large.

2.1. Asymptotic Results for General Selection Rules

Consider the Gaussian model $Y^* = \mu + \epsilon^*$ where $\epsilon^* \sim N(0,\sigma^2 I_n)$. Further, let M^* be the selection based on $f(Y) = f_+(Y^*)$ and

$$e_{h_n}(Y^*,Z,\sigma_+) = I\Big(\eta_{M_+^*}^T\mu \in [\eta_{M_+^*}^T g_+(Y^*) \pm z_{1-\alpha/2}\sqrt{1+\tau^{-2}}\sigma_+(1-h_n)\|\eta_{M_+^*}\|]\Big).$$

Importantly, we use the same Z in the Gaussian model. Since e_{h_n} depends on (Y,Z) only through (X^TY, X^TZ), we can show that

$$|\mathbb{E}[e_{h_n}(Y,Z,\sigma_+)] - \mathbb{E}[e_{h_n}(Y^*,Z,\sigma_+)]| \leq d_{\mathrm{TV}}(X^TY, X^TY^*), \quad (7)$$

where $d_{\mathrm{TV}}(W_1,W_2)$ denote the total variation distance between variables W_1 and W_2:

$$d_{\mathrm{TV}}(W_1,W_2) = \sup_A |\mathbb{P}(W_1\in A) - \mathbb{P}(W_2\in A)| = \inf_{(W_1',W_2'):W_1'\overset{d}{=}W_1, W_2'\overset{d}{=}W_2} \mathbb{P}(W_1' \neq W_2').$$

To bound $d_{\mathrm{TV}}(X^TY, X^TY^*)$, we impose assumptions on the error distribution following Bubeck and Ganguly (2018) in deriving the Entropic CLT for linear transforms of iid random variables. In particular, it implies the errors are continuous.

Assumption 2.3. $\epsilon_1,\ldots,\epsilon_n$ are iid with the following conditions satisfied.

(a) $\mathrm{KL}(\epsilon_i/\sigma\|N(0,1)) < \infty$, where KL denotes the Kullback-Leibler divergence.
(b) The distribution of ϵ_i has spectral gap c in the sense that, for any smooth function g, $\mathbb{E}[g(\epsilon_i)^2] \leq (1/c)\mathbb{E}[g'^2(\epsilon_i)]$.

The last assumption is on how the covariate dimension p could grow with n.

Assumption 2.4. As $n\to\infty$, $\sqrt{p}h_n = o(1)$ and $pL(X) = o(1)$ where

$$L(X) = \max_i H(X)_{ii} + n\max_{i\neq j}(H(X)_{ij})^2,$$
$$H(X) = X(X^TX)^{-1}X^T.$$

We show in Appendix B.2 that, if $p = o(\sqrt{n/\log n})$, Assumption 2.4 is satisfied with high probability when X is a realization of a random matrix with iid sub-Gaussian entries, based on a similar proof strategy as in Appendix F of Lei and Ding (2021).

With these assumptions, we prove that the data fission procedure described in Section 1 produces asymptotically valid confidence intervals for all selection rules satisfying Assumption 2.1 if the estimated variance is consistent.

Theorem 2.1. Under Assumptions 2.1–2.4, if $\sigma_+^2 = \sigma^2$,

$$\mathbb{E}[e(Y,Z,\hat{\sigma})] \geq 1 - \alpha + o(1). \quad (8)$$

When σ_+ is a conservative estimate of σ, we can derive lower a bound on the coverage that is uniform in σ_+.

Theorem 2.2. Under Assumptions 2.1–2.4,

$$\mathbb{E}[e(Y,Z,\hat{\sigma})] \geq 1-\alpha-\mathbb{P}\left(2|\eta_M^T Z| \geq \sqrt{\tau^2+1}(1-h_n)\|\eta_M\|\right)+o(1) \quad (9)$$

Theorem 2.2 suggests that the slackness of coverage is small if the dependence between η_M and Z is limited. For example, if M can include at most s variables, then

$$\begin{aligned}&\mathbb{P}\left(2|\eta_M^T Z| \geq \sqrt{\tau^2+1}(1-h_n)\|\eta_M\|\right)\\ &\leq \sum_{M_0:|M_0|\leq s}\mathbb{P}\left(2|\eta_{M_0}^T Z| \geq \sqrt{\tau^2+1}(1-h_n)\|\eta_{M_0}\|\right)\\ &\leq 2p^s\left\{1-\Phi\left(\sqrt{\tau^2+1}(1-h_n)/2\right)\right\}\\ &\leq 2p^s\exp\left\{-(\tau^2+1)(1-h_n)^2/8\right\},\end{aligned}$$

whenever $\sqrt{\tau^2+1}(1-h_n)/2 \geq 1$ since $1-\Phi(z) \leq (1/z)\exp\{-z^2/2\}$. Thus, the slackness is negligible if τ is chosen to be $C\sqrt{s\log p}$ for some universal constant $C>0$.

2.2. Asymptotic Results for Restricted Selection Rules

In (7), we bound the coverage difference between the raw model and Gaussian model using the most conservative total variation distance which works for all selection rules. This could be tightened under further restrictions on the selection rule.

Note that we can reformulate the unconditional coverage $\mathbb{E}[e_{h_n}(Y,Z,\sigma_+)]$ as

$$\mathbb{E}[e_{h_n}(Y,Z,\sigma_+)] = \sum_{S\subset\{1,\dots p\}}\mathbb{P}\left(M_+=S, \eta_S^T g_+(Y)\in[a_S,b_S]\right). \quad (10)$$

where $a_S = \eta_S^T\mu - z_{1-\alpha/2}\sqrt{1+\tau^{-2}}\sigma_+(1-h_n)\|\eta_S\|$ and $b_S = \eta_S^T\mu + z_{1-\alpha/2}\sqrt{1+\tau^{-2}}\sigma_+(1-h_n)\|\eta_S\|$. Since M_+ is a function of $X^T f_+(Y)$ and η_S is in the span of X, there are nonrandom sets A_S and B_S such that $M_+=S \iff X^T f_+(Y)\in A_S, \eta_S^T g_+(Y)\in[a_S,b_S] \iff X^T g_+(Y)\in B_S$. Note that B_S is the intersection of two half-spaces. Since $(X^T f_+(Y), X^T g_+(Y))$ is a linear transform of $(X^T Y, X^T Z)$, we can define C_S such that

$$X^T f_+(Y)\in A_S, X^T g_+(Y)\in B_S \iff (X^T Y, X^T Z)\in C_S. \quad (11)$$

We consider the following high-level assumption on the selection rule.

Assumption 2.5. Let $\mathcal{C}$ be a class of subsets in R^{2p} such that $C_S\in S$ for all S, and

$$\rho(\mathcal{C}) = \sup_{C\in\mathcal{C}}|\mathbb{P}((X^T Y, X^T Z)\in C) - \mathbb{P}((X^T Y^*, X^T Z)\in C)|.$$

There exists a set $\mathcal{S}$ of subsets of $\{1,\dots,p\}$ such that

$$\mathbb{P}(M_+^*\notin\mathcal{S}) = o(1), \quad |\mathcal{S}|\cdot\rho(\mathcal{C}) = o(1).$$

Remark 2.1. When $|M_+^*|\leq s$ almost surely, we can choose $\mathcal{S}$ to include all subsets of size no more than s. Then Assumption 2.5 reduces to $p^s\cdot\rho(\mathcal{C}) = o(1)$.

With the new assumption, we can prove the following result.

Corollary 2.1. Under Assumptions 2.1, 2.2, and 2.5, (9) holds. If $\sigma_+=\sigma$, (8) holds.

We discuss the implication of Assumption 2.5 for two choices of $\mathcal{C}$.

Example 1. Rasines and Young (2023) consider the case where A_S is convex for any S. Clearly, C_S is also convex since B_S is convex and linear transformations maintain convexity. Thus, we can choose $\mathcal{C}$ to be the set of convex sets in R^{2p}. Using a similar argument as in Rasines and Young (2023), we prove in Appendix B.3 that

$$\rho(\mathcal{C}) = O\left(p^{1/4}\cdot\sum_i (H(X)_{ii})^{3/2}\right) \quad (12)$$

where $H(X)$ is defined in Assumption 2.4. Since $\sum_i H(X)_{ii} = p$, by Jensen's inequality, $\sum_i (H(X)_{ii})^{3/2} \geq p^{3/2}/\sqrt{n}$. Thus, Assumption 2.5 implies $p = o(n^{2/7})$, which is worse than Assumption 2.4 for the case discussed in Appendix B.2.

Example 2. If A_S is a convex polytope for all S, C_S is a convex polytope with two more facets given by B_S. We can choose $\mathcal{C}$ to be the set of polytopes in R^{2p} with at most m facets. These are "simple convex sets" defined in Chernozhukov, Chetverikov, and Kato (2017) if $m \leq (np)^d$ for some constant d. When d is a constant, we prove in Appendix B.4 that, under regularity conditions on X and ϵ,

$$\rho(\mathcal{C}) = O\left(\frac{(\log np)^{7/6}}{n^{1/6}}\right), \quad (13)$$

As long as $p\leq n$, Assumption 2.5 holds if $|\mathcal{S}| = o(n^{1/6}/(\log n)^{7/6})$. Thus, Assumption 2.5 is weaker than Assumption 2.4 for small $|\mathcal{S}|$.

Acknowledgments

The author would like to thank James Leiner and Aaditya Ramdas for the valuable comments.

Disclosure Statement

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

Funding

The author is grateful for the support of National Science Foundation Grant DMS-2338464.

References

Barber, R. F., and Candès, E. J. (2015), "Controlling the False Discovery Rate via Knockoffs," *The Annals of Statistics*, 43, 2055–2085. [148]

Bubeck, S., and Ganguly, S. (2018), "Entropic CLT and Phase Transition in High-Dimensional Wishart Matrices," *International Mathematics Research Notices*, 2018, 588–606. [148]

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017), "Central Limit Theorems and Bootstrap in High Dimensions," *Annals of Probability: An Official Journal of the Institute of Mathematical Statistics*, 45, 2309–2352. [147,149]

Lei, L., and Ding, P. (2021), "Regression Adjustment in Completely Randomized Experiments with a Diverging Number of Covariates," *Biometrika*, 108, 815–828. [148]

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. [147]

Luo, Y., Fithian, W., and Lei, L. (2022), "Improving Knockoffs with Conditional Clibration," arXiv preprint arXiv:2208.09542. [148]

Rasines, D. G., and Alastair Young, G. (2023), "Splitting Strategies for Post-Selection Inference," *Biometrika*, 110, 597–614. [147,149]