Qualifying Paper Report for Data Fission: Splitting A Single Data Point

Naitong Chen

January 31, 2023

1 Problem Definition

In many of the modern statistical learning problems, we are interested in first using a data-driven approach to posit a statistical model for the problem at hand, and then performing inference using this selected model. However, if we use the same dataset twice for both model selection and inference, classical inference methods would yield unreliable, and typically overly optimistic results (Hong et al., 2018). Selective inference or post-selection inference deals with exactly this problem by developing inference procedues for after model selection that are statistically sound. Assuming that the data at hand are independently and identically distributed (i.i.d.), one of the most straightforward approaches for selective inference is data splitting. The approach of data splitting, as its name suggests, randomly partitions the dataset into two subsets, with one of them used to select a model, and the other used to perform inference. Since the two datasets are independent by assumption, we can then apply any standard model selection and inference methods to conduct the analysis. Despite its simplicity, this approach has its drawbacks in a number of different settings. As an example, in the case where there are high influential points, having these points assigned to one subset may result in a vastly different inference output than if it were assigned to the other. Furthermore, in the case where we do not have many observations available, data splitting, which further reduces the dataset size used for both the selection and inference steps, may increase the level of uncertainty in our analysis.

To address these shortcomings of data splitting, Leiner et al. (2022) introduces data fission, which splits each observation in the dataset into two pieces through external randomness, each containing some information about the original observation. Data fission works as follows. Given a dataset $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \pi_\theta$, with π_θ some distribution whose parameter θ is of interest. With the parameter τ controlling the amount of information allocated to either part, data fission perturbs each X_i to form $f_\tau(X_i)$ and $g_\tau(X_i)$ such that both parts contain information about θ, and that there exists some function h such that $X_i = h(f_\tau(X_i), g_\tau(X_i))$. Data fission also requires one of the following two conditions to hold. One can then use $(f_\tau(X_i))_{i=1}^n$ for selection and $(g_\tau(X_i))_{i=1}^n$ for inference.

- (P1): $f_\tau(X_i)$ and $g_\tau(X_i)$ are independent with known distributions (up to unknown θ);
- (P2): $f_\tau(X_i)$ has a known marginal distribution and $g_\tau(X_i)$ has a known conditional distribution given $f_\tau(X_i)$ (up to unknown θ).

Here we provide an instance of data fission satisfying (P1). Suppose $X_i \sim \mathcal{N}(\mu, \Sigma)$. Draw $Z_i \sim \mathcal{N}(0, \Sigma)$. Then $f_\tau(X_i) = X_i + \tau Z_i \sim \mathcal{N}(\mu, (1+\tau^2)\Sigma)$ and $g_\tau(X_i) = X_i - \tau^{-1} Z_i \sim \mathcal{N}(\mu, (1+\tau^{-2})\Sigma)$, and $f_\tau(X_i) \perp\!\!\!\perp g_\tau(X_i)$. $\tau \in (0, \infty)$ controls the level of information allocated to $f_\tau(X_i)$, with a larger τ indicating a less informative $f_\tau(X_i)$. In this particular case, a clear connection between data fission and data splitting can be drawn by looking at the amount of Fisher information allocated to the selection and inference stage. Specifically, by letting $a = \{\frac{1}{n}, \frac{2}{n}, \ldots, 1\}$ be the proportion of observations allocated to the selection stage under data splitting, setting $a = \frac{1}{1+\tau^2}$ ensures that the amount of Fisher information allocated to each of the two stages are the same between the two methods. See Example A.1 for a more detailed derivation. Therefore, data fission in the Gaussian case can be viewed as a continuous analog of data splitting, as τ is not restricted to the finite number of possible values of a.

Moving beyond the Gaussian case, Leiner et al. (2022) also provides a general data fission procedure that satisfies (P2) for distributions in the exponential family via conjugate prior and likelihood pairs. Specifically, by viewing the distribution of X, denoted p_x, as the prior distribution in the Bayesian framework, we can pick the distribution of Z (our source of external randomness) such that p_x is a conjugate prior for the likelihood function of Z. Note that in this case, both X and τ would be a parameter for the density of Z. Then by setting $f_\tau(X) = Z$ and $g_\tau(X) = X$, both the marginal distribution of $f_\tau(X)$ and the conditional distribution of $g_\tau(X)$ given $f_\tau(X)$ would be tractable, thus satisfying (P2). The general form of the distributions of $f_\tau(X)$ and $g_\tau(X) \mid f_\tau(X)$ can be found in Theorem A.2. We also provide an example where we apply Theorem A.2 to Gamma distributed random variables in Example A.3.

Leiner et al. (2022) applies data fission to constructing selective confidence intervals (CIs) in fixed-design (generalized) linear models where the selection steps performs variable selection. Exact CIs for the inference step are derived for Gaussian linear regression models, and asymptotic CIs are derived for generalized linear models. This report focuses on selective inference in fixed-design Gaussian linear regression, which we will have a more in-depth discussion on in Section 4.

2 Significance

As mentioned in the previous section, data splitting may not always be an ideal framework for selective inference. As a result, there has been a variety of alternative procedures developed. For example, the idea of introducing external randomness to derive valid inference procedures rather than randomly partitioning the data is explored in Tian and Taylor (2018) and Rasines and Young (2021). In the Gaussian example discussed in Section 1, their approach is equivalent to setting $f_\tau(X) = X + \tau Z$ and $g_\tau(X) = X$. The finite sample distribution of $g_\tau(X) \mid f_\tau(X)$ is also known when the data are Gaussian. However, if we move beyond Gaussianity, the distribution of $g_\tau(X) \mid f_\tau(X)$ is only known asymptotically. In these cases, using this asymptotic result may hinder the performance of selective inference, particularly in the small sample setting, which is one situation where data splitting struggles that we would like to address. Another approach developed in Fithian et al. (2014) is data carving. Data carving performs model selection using some fixed model selection procedure S on the original dataset. By denoting the selected model as $S(X)$, we then conduct inference on X conditioned on $S(X)$. Instead of injecting external randomness, data carving performs inference using the "leftover information" from the selection step obtained through conditioning. However, since the distribution of $X \mid S(X)$ depends on the selection procedure S, practitioners are limited to choices of S such that $X \mid S(X)$ can be obtained either analytically or numerically. Examples include LASSO (Lee et al., 2016) and step-wise regression (Tibshirani et al., 2016).

Data fission combines the advantages of both approaches. Similar to the approach in Tian and Taylor (2018) and Rasines and Young (2021), we create two slightly perturbed datasets, $f_\tau(X)$ and $g_\tau(X)$, both of which are of the same size of the original data. This is done by injecting external randomness (Z) to the original data at hand. Here the distribution of Z and the perturbations are carefully constructed so that both the finite sample marginal distribution of $f_\tau(X)$ and conditional distribution $g_\tau(X) \mid f_\tau(X)$ are known analytically. Subsequently, we conduct inference under $g_\tau(X) \mid f_\tau(X)$, which is similar in spirit to data carving in terms of taking advantage of the "leftover information". It is worth noting that, by cleverly leveraging the mechanism of conjugate distributions, which Leiner et al. (2022) terms "conjugate prior reversal", we can derive data fission procedures for many distributions in the rich exponential family. The conditional distribution used for inference is then available regardless of the choice of the selection algorithm, thus making data fission applicable to a much wider-range of problems than both of the other approaches discussed.

3 Limitations and challenges

Closer inspection of the two data fission examples given in Section 1 reveals that both the marginal distribution of $f_\tau(X)$ and conditional distribution $f_\tau(X) \mid g_\tau(X)$ depend on parameters of the distribution of the original data X. In fact, a general assumption of data fission is that the distribution of X needs to be known. This may not be a realstic assumption for many cases in practice. Although Leiner et al. (2022) provides an asymptotic CI for Gaussian linear regression when only a consistent estimator of the variance rather than the true value is available, much of the questions regarding the robustness of data fission to the distribution assumption beyond the Gaussian case remain unexplored. Furthermore, data fission may lead to $f_\tau(X)$ and $g_\tau(X) \mid f_\tau(X)$ being under different distribution families than that of X. As an example, in Example A.3, with gamma distributed X, we have $f_\tau(X)$ following a negative binomial distribution. At the same time, the density of $g_\tau(X) \mid f_\tau(X)$ as a function of the original parameter of interest θ may be highly complex and potentially non-convex. This poses challenges in the inference stage, as discussed in Appendix A.5 of Leiner et al. (2022). (We provide an instance of this problem in Example A.5.) Finally, we note that many data fission procedures will also result in cases where the distributions of $g_\tau(X) \mid f_\tau(X)$ depend explicitly on the realized values of Z. We more closely explore the effect this has on selective inference in Section 4.

Before concluding this section, we note an instance of imprecise notation as well as a possible mistake in Leiner et al. (2022). Theorem 1 in Leiner et al. (2022) misses the base measure term $m(\cdot)$ in the density function of X and mistakenly uses the raw value of x rather than its natural parameter $\phi(x)$ in both the densities of X and Z. Here ϕ is the natural parameter transformation function for the distribution of Z. For example, for $\mathsf{Poiss}(\lambda)$, $\phi(\lambda) = \log \lambda$. We present a corrected version of the proof in Theorem A.2. Although the final statement of the theorem remains the same, this notation impreciseness would likely result in incorrect applications of the theorem that may fail to lead us to the desired data fission procedure. Additionally,

there seems to be an incorrect data fission procedure derived for gamma distributed data. In particular, the marginal density for Z does not match that of Theorem A.2. We present the details in Example A.4. It is worth noting that even if the result presented in Leiner et al. (2022) were indeed correct, it is still unclear how we can modify the inference step to accommodate this data fission procedure, as the dimension of Z is no longer necessarily the same as X.

4 Paper-specific project

In this section, we compare two of the data fission procedures introduced in Leiner et al. (2022) for Gaussian distributed data against data splitting, in the context of constructing selective CIs in fixed-design linear regression models. Suppose we are given a set of n observations such that for $i \in \{1, \ldots, n\}$, $y_i = x_i^\top \beta + \epsilon_i$, where for some $p \in \mathbb{N}$ and known $\sigma > 0$, $x_i \in \mathbb{R}^p$, $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, $\beta \in \mathbb{R}^p$. Equivalently, we can write $Y_i \sim \mathcal{N}(x_i^\top \beta, \sigma^2)$. Note here we assume that the covariates x_i's are fixed. The goal is to first perform variable selection and then construct CIs for the regression coefficients corresponding to the selected variables. We would like to compare variable selection accuracy as well as inference quality obtained using the following three procedures:

- Data fission (P1): for each $i \in \{1, \ldots, n\}$, and some fixed $\tau \in (0, \infty)$, draw $Z_i \sim \mathcal{N}(0, \sigma^2)$. Let $f_\tau(Y_i) = Y_i + \tau Z_i$, $g_\tau(Y_i) = Y_i - \tau^{-1} Z_i$. Then $f_\tau(Y_i) \perp\!\!\!\perp g_\tau(Y_i)$ and $f_\tau(Y_i) \sim \mathcal{N}(x_i^\top \beta, (1+\tau^2)\sigma^2)$, $g_\tau(Y_i)\mathcal{N}(x_i^\top \beta, (1+\tau^{-2})\sigma^2)$. We use $(f_\tau(Y_i))_{i=1}^n$ for selection and $(g_\tau(Y_i))_{i=1}^n$ for inference.
- Data fission (P2): for each $i \in \{1, \ldots, n\}$, and some fixed $\tau \in (0, \infty)$, draw $Z_i \sim \mathcal{N}(Y_i, \tau\sigma^2)$. Let $f_\tau(Y_i) = Z_i$, $g_\tau(Y_i) = Y_i$. Then $f_\tau(Y_i) \sim \mathcal{N}(x_i^\top \beta, (1+\tau)\sigma^2)$, $g_\tau(Y_i) \mid f_\tau(Y_i) \sim \mathcal{N}\left(\frac{\tau}{\tau+1} x_i^\top \beta + \frac{1}{\tau+1} f_\tau(Y_i), \frac{\tau}{\tau+1}\sigma^2\right)$. We use $(f_\tau(Y_i))_{i=1}^n$ for selection and $(g_\tau(Y_i))_{i=1}^n$ for inference.
- Data splitting: for some fixed $a \in \{\frac{1}{n}, \frac{2}{n}, \ldots, 1\}$, randomly draw an observations from $(Y_i)_{i=1}^n$ without replacement. Without loss of generality, denote the first an observations as those that are selected. Use $(Y_i)_{i=1}^{an}$ for selection and $(Y_i)_{i=an+1}^n$ for inference.

We begin by observing the three datasets used for the selection step. Both data fission procedures inflate the variance of each observation without changing their underlying means. Data splitting, on the other hand, directly reduces the number of observations without perturbing their underlying distributions. This also introduces uncertainty to the selection step. It is then of interest to compare variable selection accuracies under different sample sizes and the amount of variance inflated. However, recall that from the Fisher information perspective, to make the comparison fair, we set $a = \frac{1}{1+\tau^2}$. As a result, the comparison reduces to be about varying the sample sizes.

We now consider the inference stage. Suppose that the selected model is $M \subset \{1, \ldots, p\}$ and that

$$X_M = \left[x_{M,1}, \ldots, x_{M,n}\right]^\top, \quad Y = \left[Y_1, \ldots, Y_n\right]^\top,$$

where $x_{M,i}$ is a vector consisting of only the covariates corresponding to the selected features. We then have our ideal target parameter given M as

$$\beta^\star(M) = \arg\min_{\tilde{\beta}} \mathbb{E}_Y \|Y - X_M \tilde{\beta}\|^2 = (X_M^\top X_M)^{-1}(X_M^\top \mu)$$

where $\mu = \left[x_1^\top \beta, \ldots, x_n^\top \beta\right]^\top$. However, since we are using slightly perturbed datasets for inference, our estimator $\hat{\beta}(M)$ given the model M for each of the three procedures are

- Data fission (P1): $\hat{\beta}(M) = (X_M^\top X_M)^{-1} X_M^\top g_\tau(Y) \sim \mathcal{N}(\beta^\star(M), \sigma^2(1+\tau^{-2})(X_M^\top X_M)^{-1})$;
- Data fission (P2):
  $\hat{\beta}(M) = (X_M^\top X_M)^{-1} X_M^\top g_\tau(Y) \mid f_\tau(Y) \sim \mathcal{N}(\frac{\tau}{\tau+1}\beta^\star(M) + \frac{1}{\tau+1}(X_M^\top X_M)^{-1} X_M^\top f_\tau(Y), \sigma^2 \frac{\tau}{\tau+1}(X_M^\top X_M)^{-1})$;
- Data splitting: $\hat{\beta}(M) = (X_M^\top X_M)^{-1} X_M^\top Y \sim \mathcal{N}(\beta^\star(M), \sigma^2(X_M^\top X_M)^{-1})$.

Note that here f and g are applied to each entry of Y and that X_M and Y in data splitting only contains a subset set of na observations.

Looking at the means of each estimator above, both the data fission (P1) and data splitting procedures target $\beta^\star(M)$. On the other hand, the mean of the data fission (P2) estimator depends on the realized values of the external random variable Z. Since $\mathbb{E}[f_\tau(Y)] = \mu$, if we marginalize $f_\tau(Y)$ over the distribution of $\hat{\beta}(M)$, data fission (P2) would also target the ideal parameter $\beta^\star(M)$. However, it is reasonable to suspect that this randomness might affect the quality of inference.

We now look at the variances of the three estimators. Similar to the selection stage, data fission (P1) inflates the variance of $\hat{\beta}(M)$ by a function of τ, and data splitting introduces additional uncertainty by reducing the sample size. However, in data fission (P2), the variance is deflated. Suppose for each i, x_i is generated (i.i.d.) by some distribution π, and assume that for all selected model M, $\mathbb{E}[x_{M,1}x_{M,1}^\top]$ is finite and strictly positive. Then we have

$$\frac{1}{n}X_M^\top X_M \xrightarrow{p} \mathbb{E}[x_{M,1}x_{M,1}^\top] \implies \left(\frac{1}{n}X_M^\top X_M\right)^{-1} \xrightarrow{p} \left(\mathbb{E}[x_{M,1}x_{M,1}^\top]\right)^{-1} \implies X_M^\top X_M \xrightarrow{p} \frac{1}{n}\left(\mathbb{E}[x_{M,1}x_{M,1}^\top]\right)^{-1}.$$

Under this assumption, the difference in inference between data splitting and data fission once again comes down to the sample size and the amount of variance inflated or deflated by τ. We therefore set up our simulations as following. Note that this is mostly in line with Section 4 of Leiner et al. (2022).

We set $a = \frac{1}{2}$ and $\tau = 1$ to ensure the amount of Fisher information allocated to the selection stage is the same across all three methods. Let $p = 20, \beta_1 = \beta_{19} = 1$, $\beta_2 = \beta_{20} = -1$, and the rest of the entries in β be 0. We also generate the covariates from the standard multivariate Guassian distribution. Note that there is no intercept in our regression model. We then conduct selective inference with varying sample sizes $10, 20, 50, 100$ to evaluate how the performances across all three methods change under different sample sizes. To compare variable selection accuracy, we use as our metrics

$$\text{power} = \frac{|j \in M : \beta_j \neq 0|}{|j \in [p] : \beta_j \neq 0|}, \quad \text{precision} = \frac{|j \in M : \beta_j \neq 0|}{|M|}.$$

To compare the quality of inference, we use false coverage rate (FCR), avg. CI length, and avg. L2 error:

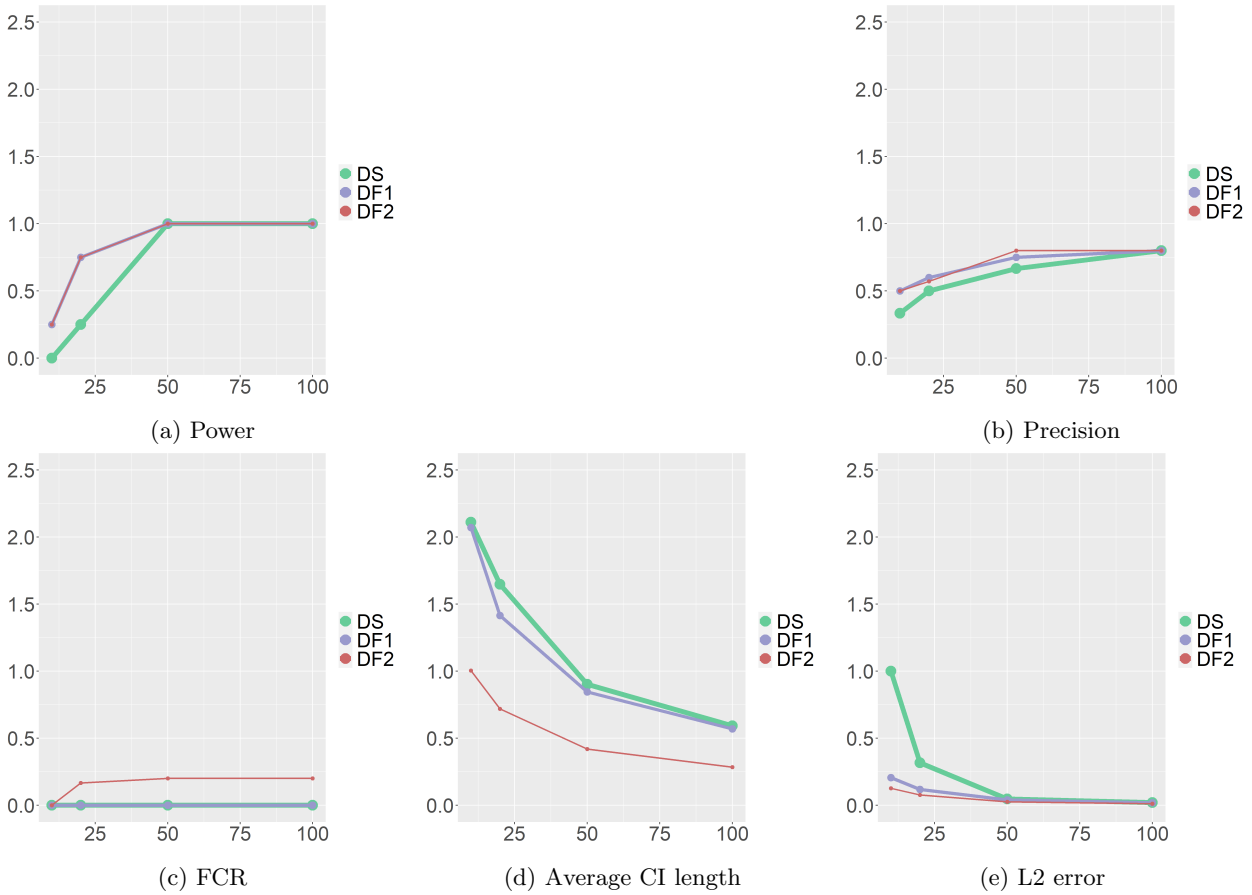
$$\text{FCR} = \frac{|k \in M : [\beta^\star(M)]_k \notin CI_k|}{\max\{|M|, 1\}}, \quad \overline{\text{CI len.}} = \frac{\sum_{k\in M} |\text{CI}_k(2) - \text{CI}_k(1)|}{|M|}, \quad \text{L2 err.} = \frac{\|\beta^\star(M) - \hat{\beta}(M)\|_2^2}{|M|}.$$

To perform variable selection, we run LASSO using the `glmnet` package in `R` with the default settings and the regularization parameter set to `lambda.1se`. We repeat the above experiments 200 times and report the median of the above metrics (excluding runs that do not end up selecting any variable in the selection step) in Fig. 1. The same set of plots with the IQR of each metric, along with other details of the experiment is included in Appendix B. Note that since the IQRs have a lot of overlaps, the discussion below is only concerned with the average performance rather than individual trials. The code used to run the simulations and generate the plots can be found at https://github.com/NaitongChen/QP-3.

4.1 Discussion of simulation results

We begin by comparing variable selection accuracy through power and precision (Figs. 1a and 1b). Both metrics are similar across different sample sizes between the two data fission procedures, because they have the same $f_\tau(Y)$ marginal distributions with $\tau = 1$. Compared to data splitting, data fission achieves higher power and precision, particularly when sample size is small. When there are few observations, the disadvantage of data splitting working with only half of the observations becomes apparent. In particular, if we look at individual trials (Fig. 3a), data splitting tends to miss true features and pick instead many false features, leading to suboptimal power and precision. This is to be expected, as when there is not a lot of information from the data, halving the number of observations would likely hinder the quality of variable selection. Although data fission inflates the variance of the observations by a factor of two, this seems to be a worthy price to pay in order to not reduce the number of observations used in the selection stage. As we increase the sample size, however, even half of the information becomes sufficient for variable selection, which makes the advantage of data fission less obvious.

We now look at the quality of inference through FCR, avg. CI length, as well as the L2 error between the estimated parameter and the ideal target of inference averaged by the number of variables selected. Fig. 1c shows that the median FCR for data splitting and data fission (P1) are both well below the target level $\alpha = 0.05$ using standard unadjusted 95% CIs. Note that no correction is needed here since the proportion of CIs covering their respective parameters is expected to be 0.95 (Benjamini and Yekutieli, 2005). On the other hand, data fission (P2) has a much higher FCR (above the target level). This is due to the mean of



(a) Power (b) Precision

(c) FCR (d) Average CI length (e) L2 error

Figure 1: Median of each metric across 200 runs. The x-axis shows sample size, and the y-axis shows the value of each metric. DS refers to data splitting, DF1 and DF2 refer to data fission (P1) and data fission (P2).

$\hat{\beta}(M)$ being different than $\beta^\star(M)$ and its variance being further deflated by the fission procedure. This is evident in Fig. 3l. Interestingly, when the sample size is small ($n = 10$), data fission (P2) is also able to control the FCR. One possible explanation is that the small sample size makes $(X^\top X)^{-1}$ highly variable and thus more likely widens the corresponding CI compared to when the sample size is large. It is also worth noting that having a low FCR does not necessarily mean that we recover the true parameters, because there is a discrepancy between β and $\beta^\star(M)$, especially when sample size is small (see for example Fig. 3a).

In terms of the average length of 95% CIs (Fig. 1d), we see that data fission (P1) and data splitting are similar across all sample sizes. This suggests that the effect on the average length of CI is similar between inflating the variance and halving the number of observations. On the other hand, the average CI length for data fission (P2) is much smaller than the other two methods. This can be explained by that the variance of $\hat{\beta}(M)$ is deflated by the fission procedure and that there is no reduction in sample size.

Finally, we compare the L2 error between the estimated parameter and the ideal target of inference averaged by the number of variables selected in Fig. 1e. We see here that the L2 error decays as the sample size increases for all three methods. We see that the L2 error for data splitting is much higher than the two data fission procedures, particularly when the sample size is small. This is again due to data splitting reducing the sample size at the inference stage, which introduces uncertainty. It is somewhat surprising, however, to see that both data fission methods result in very similar L2 errors, despite data fission (P2) being not targeting the ideal parameter ($\mathbb{E}[\hat{\beta}(M)] \neq \beta^\star(M)$). Considering the high FCR of data fission (P2), this implies that the effect of the discrepancy between $\mathbb{E}[\hat{\beta}(M)]$ and $\beta^\star(M)$ does not hinder the inference performance as much as the deflation of $\hat{\beta}(M)$'s variance. This calls for further investigation, and a possible way to explore this issue is to compare the L2 errors across different σ^2 values in order to amplify the effect.

5 Discussion

Overall, in linear regression, P1 and splitting asymptotically similar, but P1 offering better performance when sample size is small, in both selection and inference. P2 is similar to P1 at selection, but inference performance is hindered by bias. To conclude, if we know variance, and have a small sample, use data fission over splitting. P1 better than P2 because no bias. Raises questions on how to deal with this randomness in P2, as it does give a framework for extending data fission to non Gaussian data.

Worth noting that these metrics appear to be highly variable. This implies that given a single instance of data fission vs. data splitting, the performance in terms of these metrics may be pretty similar, but on average, we observe the phenomenons discussed above.

References

Yoav Benjamini and Daniel Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Liang Hong, Todd A Kuffner, and Ryan Martin. On overfitting and post-selection uncertainty assessments. *Biometrika*, 105(1):221–224, 2018.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. 2016.

James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: splitting a single data point. *arXiv preprint arXiv:2112.11079 v4*, 2022.

Daniel G Rasines and G Alastair Young. Splitting strategies for post-selection inference. *arXiv preprint arXiv:2102.02159*, 2021.

Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018.

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

A Proofs and examples

Example A.1. *In the Gaussian case, data fission can be viewed as a continuous analog of data splitting in terms of the allocation of Fisher information.*
Let $\{X_i\}_{i=1}^n \subset \mathbb{R}$ *be i.i.d.* $\mathcal{N}(\theta, \sigma^2)$*. Let* $X := \left[X_1, \ldots, X_n\right]^\top$*. Data splitting defines* $f(X)$ *and* $g(X)$ *as*

$$f^{split}(X) = \left[X_1, \ldots, X_{an}\right], \quad g^{split}(X) = \left[X_{an+1}, \ldots, X_n\right],$$

for $a \in \left\{\frac{1}{n}, \frac{2}{n}, \ldots, 1\right\}$*. Note that the Fisher information for both pieces are*

$$\mathcal{I}_{f^{split}(X)} = an\frac{1}{\sigma^2}, \quad \mathcal{I}_{g^{split}(X)} = (1-a)n\frac{1}{\sigma^2}.$$

On the other hand, data fission first simulates $\{Z_i\}_{i=1}^n$ *distributed as i.i.d.* $\mathcal{N}(0, \sigma^2)$ *and have, for some fixed* $\tau \in (0, \infty)$*,*

$$f^{fisson}(X) = \left[X_1 + \tau Z_1, \ldots, X_n + \tau Z_n\right], \quad g^{fisson}(X) = \left[X_1 - \tfrac{1}{\tau} Z_1, \ldots, X_n - \tfrac{1}{\tau} Z_n\right].$$

Note that for all $i \in \{1, \ldots, n\}$*,* $X_i + \tau Z_i \sim \mathcal{N}\left(\theta, (1+\tau^2)\sigma^2\right)$*,* $X_i - \frac{1}{\tau} Z_i \sim \mathcal{N}\left(\theta, (1+\frac{1}{\tau^2})\sigma^2\right)$*, and for each* i*,* $X_i + \tau Z_i \perp\!\!\!\perp X_i - \frac{1}{\tau} Z_i$*. We then have*

$$\mathcal{I}_{f^{fisson}(X)} = n\frac{1}{(1+\tau^2)\sigma^2}, \quad \mathcal{I}_{g^{fisson}(X)} = n\frac{1}{(1+\frac{1}{\tau^2})\sigma^2}.$$

By setting $a = \frac{1}{1+\tau^2}$*, we have* $\mathcal{I}_{f^{split}(X)} = \mathcal{I}_{f^{fisson}(X)}$ *and* $\mathcal{I}_{g^{split}(X)} = \mathcal{I}_{g^{fisson}(X)}$*.* ◁

Theorem A.2. *Suppose that for some* $A(\cdot), \phi(\cdot), m(\cdot), \theta_1, \theta_2, H(\cdot, \cdot)$*, the density of* X *is given by*

$$p(x \mid \theta_1, \theta_2) = m(x)H(\theta_1, \theta_2)\exp\{\theta_1^\top \phi(x) - \theta_2^\top A(\phi(x))\}.$$

Suppose also that there exists $h(\cdot), T(\cdot), \theta_3$ *such that*

$$p(z \mid x, \theta_3) = h(z)\exp\{\phi(x)^\top T(z) - \theta_3^\top A(\phi(x))\}$$

is a well-defined distribution. First, draw $Z \sim p(z \mid X, \theta_3)$*, and let* $f(X) := Z$ *and* $g(X) := X$*. Then,* $(f(X), g(X))$ *satisfy the data fission property (P2). Specifically, note that* $f(X)$ *has a known marginal distribution*

$$p(z \mid \theta_1, \theta_2, \theta_3) = h(z)\frac{H(\theta_1, \theta_2)}{H(\theta_1 + T(z), \theta_2 + \theta_3)},$$

while $g(X)$ *has a known conditional distribution given* $f(X)$*, which is*

$$p(x \mid z, \theta_1, \theta_2, \theta_3) = p(x \mid \theta_1 + T(z), \theta_2 + \theta_3).$$

Proof. Note that because the density $p(z \mid x, \theta_3)$ must integrate to 1, we can view the function $H(\theta_1, \theta_2)$ as a normalization factor since

$$H(\theta_1, \theta_2) = \frac{1}{\int_{-\infty}^{\infty} m(x)\exp\{\theta_1^\top \phi(x) - \theta_2^\top A(\phi(x))\}dx}.$$

Therefore, to compute the marginal density, we have

$$\begin{aligned} p(z \mid \theta_1, \theta_2, \theta_3) &= \int_{-\infty}^{\infty} m(x)h(z)H(\theta_1, \theta_2)\exp\{(T(z) + \theta_1)^\top \phi(x) - (\theta_2 + \theta_3)^\top A(\phi(x))\}dx \\ &= h(z)\frac{H(\theta_1, \theta_2)}{H(\theta_1 + T(z), \theta_2 + \theta_3)}. \end{aligned}$$

Similarly, the computation of the conditional density is straightforward

$$\begin{aligned} p(x \mid z, \theta_1, \theta_2, \theta_3) &= \frac{m(x)h(z)H(\theta_1,\theta_2)\exp\{(T(z)+\theta_1)^\top \phi(x) - (\theta_2+\theta_3)^\top A(\phi(x))\}}{h(z)\frac{H(\theta_1,\theta_2)}{H(\theta_1+T(z),\theta_2+\theta_3)}} \\ &= m(x)H(\theta_1+T(z),\theta_2+\theta_3)\exp\{\phi(x)^\top(\theta_1+T(z)) - (\theta_2+\theta_3)^\top A(\phi(x))\} \\ &= p(x \mid \theta_1+T(z), \theta_2+\theta_3). \end{aligned}$$

This completes the proof. □

Example A.3. *Suppose $X \sim \mathsf{Gam}(\alpha,\beta)$. Draw $Z \sim \mathsf{Poiss}(\tau X)$, where $\tau \in (0,\infty)$ is a tuning parameter. Let $f(X) = Z$ and $g(X) = X$.*
By writing

$$\mathsf{Gam}(x \mid \alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\exp\{(\alpha-1)\log x - \beta x\},$$

where $\mathsf{Gam}(\cdot \mid \alpha,\beta)$ denotes the pdf of $\mathsf{Gam}(\alpha,\beta)$, we have that $\theta_2 = \beta, \theta_1 = \alpha-1, \phi(x) = \log x, A(\phi(x)) = \exp(\phi(x)) = \exp(\log x) = x, m(x) = 1$. Therefore, $H(\theta_1,\theta_2) = \frac{\theta_2^{(\theta_1+1)}}{\Gamma(\theta_1+1)}$. Now since

$$\begin{aligned} \mathsf{Poiss}(z \mid \tau x) &= \frac{1}{z!}\exp\{z\log(\tau x) - \tau x\} \\ &= \frac{1}{z!}\exp\{z\log\tau + z\log x - \tau x\} \\ &= \frac{\tau^z}{z!}\exp\{z\log x - \tau x\}, \end{aligned}$$

we have that $h(z) = \frac{\tau^z}{z!}, T(z) = z, \theta_3 = \tau$. Therefore, by Theorem A.2,

$$\begin{aligned} p(z \mid \theta_1,\theta_2,\theta_3) &= \frac{\tau^z}{z!}\frac{\frac{\beta^\alpha}{\Gamma(\alpha)}}{\frac{(\beta+\tau)^{(\alpha+z)}}{\Gamma(\alpha+z)}} \\ &= \frac{(\alpha+z-1)!}{(\alpha-1)!z!}\left(\frac{\beta}{\beta+\tau}\right)^\alpha\left(\frac{\tau}{\beta+\tau}\right)^z \\ &= \mathsf{NB}\left(z \mid \alpha, \frac{\beta}{\beta+\tau}\right); \\ p(x \mid z,\theta_1,\theta_2,\theta_3) &= \frac{(\beta+\tau)^{(\alpha+z)}}{\Gamma(\alpha+z)}\exp\{(\alpha+z-1)\log(x) - (\beta+\tau)x\} \\ &= \mathsf{Gam}(x \mid \alpha+z, \beta+\tau). \end{aligned}$$

◁

Example A.4. *Suppose $X \sim \mathsf{Gam}(\alpha,\beta)$. Draw $Z = (Z_1,\cdots,Z_B)$ where each element is i.i.d. $Z_i \sim \mathsf{Poiss}(X)$ and $B \in \{1,2,\dots\}$ is a tuning parameter. Let $f(X) = Z$ and $g(X) = X$.*
By writing

$$\mathsf{Gam}(x \mid \alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\exp\{(\alpha-1)\log x - \beta x\},$$

where $\mathsf{Gam}(\cdot \mid \alpha,\beta)$ denotes the pdf of $\mathsf{Gam}(\alpha,\beta)$, we have that $\theta_2 = \beta, \theta_1 = \alpha-1, \phi(x) = \log x, A(\phi(x)) = \exp(\phi(x)) = \exp(\log x) = x, m(x) = 1$. Therefore, $H(\theta_1,\theta_2) = \frac{\theta_2^{(\theta_1+1)}}{\Gamma(\theta_1+1)}$. Now since

$$\begin{aligned} \mathsf{Poiss}(z \mid x) &= \prod_{i=1}^B \frac{1}{z_i!}\exp\{z_i\log x - x\} \\ &= \left(\prod_{i=1}^B \frac{1}{z_i!}\right)\exp\left\{\left(\sum_{i=1}^B z_i\right)\log x - Bx\right\}, \end{aligned}$$

we have that $h(z) = \prod_{i=1}^{B} \frac{1}{z_i!}, T(z) = \sum_{i=1}^{B} z_i, \theta_3 = B$. *Therefore, by Theorem A.2, when* $B = 1$,

$$\begin{aligned}
p(z \mid \theta_1, \theta_2, \theta_3) &= \frac{1}{z!} \frac{\frac{\beta^\alpha}{\Gamma(\alpha)}}{\frac{(\beta+1)^{(\alpha+z)}}{\Gamma(\alpha+z)}} \\
&= \frac{(\alpha+z-1)!}{(\alpha-1)!z!} \left(\frac{\beta}{\beta+1}\right)^{\alpha} \left(\frac{1}{\beta+1}\right)^{z} \\
&= \mathsf{NB}\left(z \mid \alpha, \frac{\beta}{\beta+1}\right); \\
p(x \mid z, \theta_1, \theta_2, \theta_3) &= \frac{(\beta+1)^{(\alpha+z)}}{\Gamma(\alpha+z)} \exp\left\{(\alpha+z-1)\log(x) - (\beta+1)x\right\} \\
&= \mathsf{Gam}\left(x \mid \alpha+z, \beta+1\right).
\end{aligned}$$

However, when $B > 1$,

$$\begin{aligned}
p(z \mid \theta_1, \theta_2, \theta_3) &= \left(\prod_{i=1}^{B} \frac{1}{z_i!}\right) \frac{\frac{\beta^\alpha}{\Gamma(\alpha)}}{\frac{(\beta+B)^{\left(\alpha+\sum_{i=1}^{B} z_i\right)}}{\Gamma\left(\alpha+\sum_{i=1}^{B} z_i\right)}} \\
&\neq \prod_{i=1}^{B} \mathsf{NB}\left(z_i \mid \alpha, \frac{\beta}{\beta+B}\right); \\
p(x \mid z, \theta_1, \theta_2, \theta_3) &= \frac{(\beta+B)^{\left(\alpha+\sum_{i=1}^{B} z_i\right)}}{\Gamma\left(\alpha+\sum_{i=1}^{B} z_i\right)} \exp\left\{\left(\alpha - 1 + \sum_{i=1}^{B} z_i\right)\log(x) - (\beta+B)x\right\} \\
&= \mathsf{Gam}\left(x \mid \alpha + \sum_{i=1}^{B} z_i, \beta+B\right).
\end{aligned}$$

◁

Example A.5. *Assume for all* $i \in \{1, 2, \ldots, n\}$, $Y_i \overset{i.i.d.}{\sim} \mathsf{Gam}(\alpha, \exp(\beta^\top x_i))$, *where each* $x_i \in \mathbb{R}^d$ *is fixed. Following the data fission procedure in Example A.3, we have that for all* i, $f(Y_i) = Z_i, g(Y_i) = Y_i$. *In the selection phase of selective inference, for some fixed* $\lambda > 0$, *we can do model selection via the optimization below*

$$\begin{aligned}
\hat{\beta}_\lambda &= \underset{\beta \in \mathbb{R}^d}{\arg\min} - \sum_{i=1}^{n} \left(\log \mathsf{NB}\left(z_i \mid \alpha, \frac{\exp(\beta^\top x_i)}{\exp(\beta^\top x_i) + \tau}\right)\right) + \lambda \|\beta_1\| \\
&= \underset{\beta \in \mathbb{R}^d}{\arg\min} \left(\sum_{i=1}^{n} -\log \binom{z_i + \alpha - 1}{z_i} - z_i \log\left(\frac{1}{\exp(\beta^\top x_i) + \tau}\right) - \alpha \log\left(\frac{\exp(\beta^\top x_i)}{\exp(\beta^\top x_i) + \tau}\right)\right) + \lambda \|\beta_1\|,
\end{aligned}$$

which is convex in β. *Denote the index set of nonzero entries in* $\hat{\beta}_\lambda$ *to be* $\mathcal{M}$ *and* $|\mathcal{M}| = d' \leq d$.
Using the selected features $\mathcal{M}$, *with* $x_{i,\mathcal{M}}$ *denoting the selected features for the* i*th observation, we can obtain the estimates* $\hat{\beta}_n(\mathcal{M})$ *via*

$$\begin{aligned}
&\hat{\beta}_n(\mathcal{M}) \\
&= \underset{\beta \in \mathbb{R}^{d'}}{\arg\min} - \sum_{i=1}^{n} \left(\log \mathsf{Gam}(y_i \mid \alpha + z_i, \exp(\beta^\top x_{i,\mathcal{M}}) + \tau)\right) \\
&= \underset{\beta \in \mathbb{R}^{d'}}{\arg\min} \sum_{i=1}^{n} -(\alpha + z_i) \log\left(\exp(\beta^\top x_{i,\mathcal{M}}) + \tau\right) + \log\Gamma(\alpha + z_i) - (\alpha + z_i - 1)\log y_i + \left(\exp(\beta^\top x_{i,\mathcal{M}}) + \tau\right) y_i,
\end{aligned}$$

which may be convex in β *but without an argmin, or non-convex (depending on the value of* α). *Therefore,*

we can instead use an alternative working model

$$\begin{aligned}
&\hat{\beta}_n(\mathcal{M}) \\
&= \arg\min_{\beta \in \mathbb{R}^{d'}} -\sum_{i=1}^{n} \left(\log \mathsf{Gam}(y_i \mid \alpha + z_i, \exp(\beta^\top x_{i,\mathcal{M}}))\right) \\
&= \arg\min_{\beta \in \mathbb{R}^{d'}} \sum_{i=1}^{n} -(\alpha + z_i) \log\left(\exp(\beta^\top x_{i,\mathcal{M}})\right) + \log\Gamma(\alpha + z_i) - (\alpha + z_i - 1)\log y_i + \left(\exp(\beta^\top x_{i,\mathcal{M}})\right) y_i.
\end{aligned}$$

For this problem, we have that

$$\frac{\partial m_i}{\partial \eta_i} = \exp(\beta^\top x_{i,\mathcal{M}}), \quad m_i = \frac{\alpha + z_i}{\exp(\beta^\top x_{i,\mathcal{M}})}, \quad v_i = \frac{\alpha + z_i}{\left(\exp(\beta^\top x_{i,\mathcal{M}})\right)^2}.$$

Let D, V, M, G be diagonal matrices with

$$D_{ii} = \frac{\partial m_i}{\partial \eta_i}, \quad V_{ii} = v_i, \quad M_{ii} = m_i, \quad G_{ii} = g(y_i) - m_i.$$

Following the procedure in Appendix A.5 in Leiner et al. (2022), we have the following plug-in estimator for variance

$$\hat{H}_n^{-1}\hat{V}_n\hat{H}_n^{-1} = \left(X_{\mathcal{M}}^\top D^2 V^{-1} X_{\mathcal{M}}\right)^{-1} \left(X_{\mathcal{M}}^\top G^2 V^{-2} D^2 X_{\mathcal{M}}\right) \left(X_{\mathcal{M}}^\top D^2 V^{-1} X_{\mathcal{M}}\right)^{-1}.$$

We can then construct confidence intervals with significance level α, for $k \in \{1, \ldots, p\}$ by

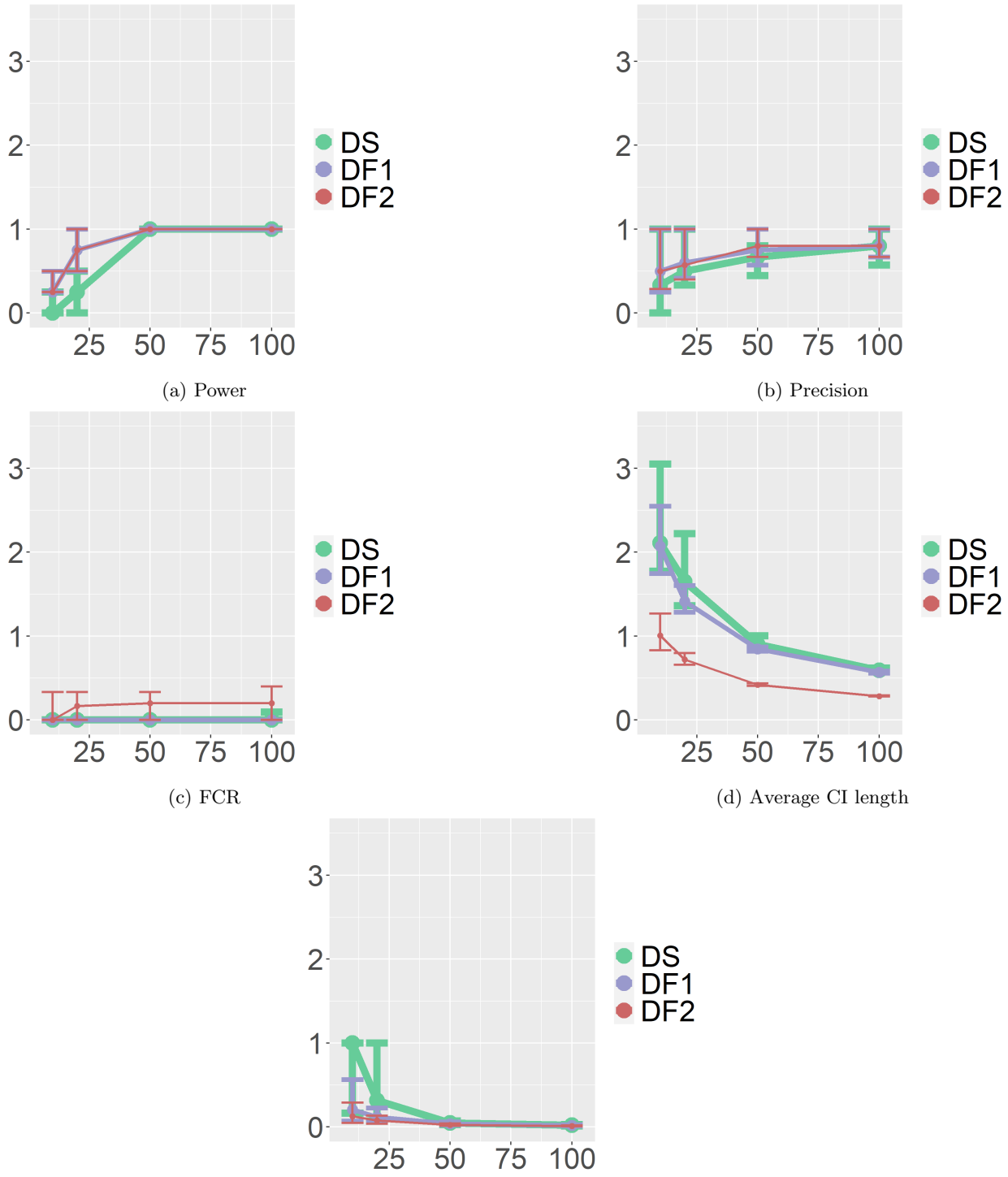
$$[\hat{\beta}_n]_k \pm z_{\frac{\alpha}{2}} \sqrt{[\hat{H}_n^{-1}\hat{V}_n\hat{H}_n^{-1}]_{kk}}.$$

Note now the quality of this confidence interval depends on the robustness of the working model. ◁

B Additional simulation details

Table 1: Number of runs (out of 200) that do not select any variable.

	n=10	n=20	n=50	n=100
Data splitting	103	60	1	0
Data fission (P1)	72	49	0	0
Data fission (P2)	66	41	0	0



(a) Power

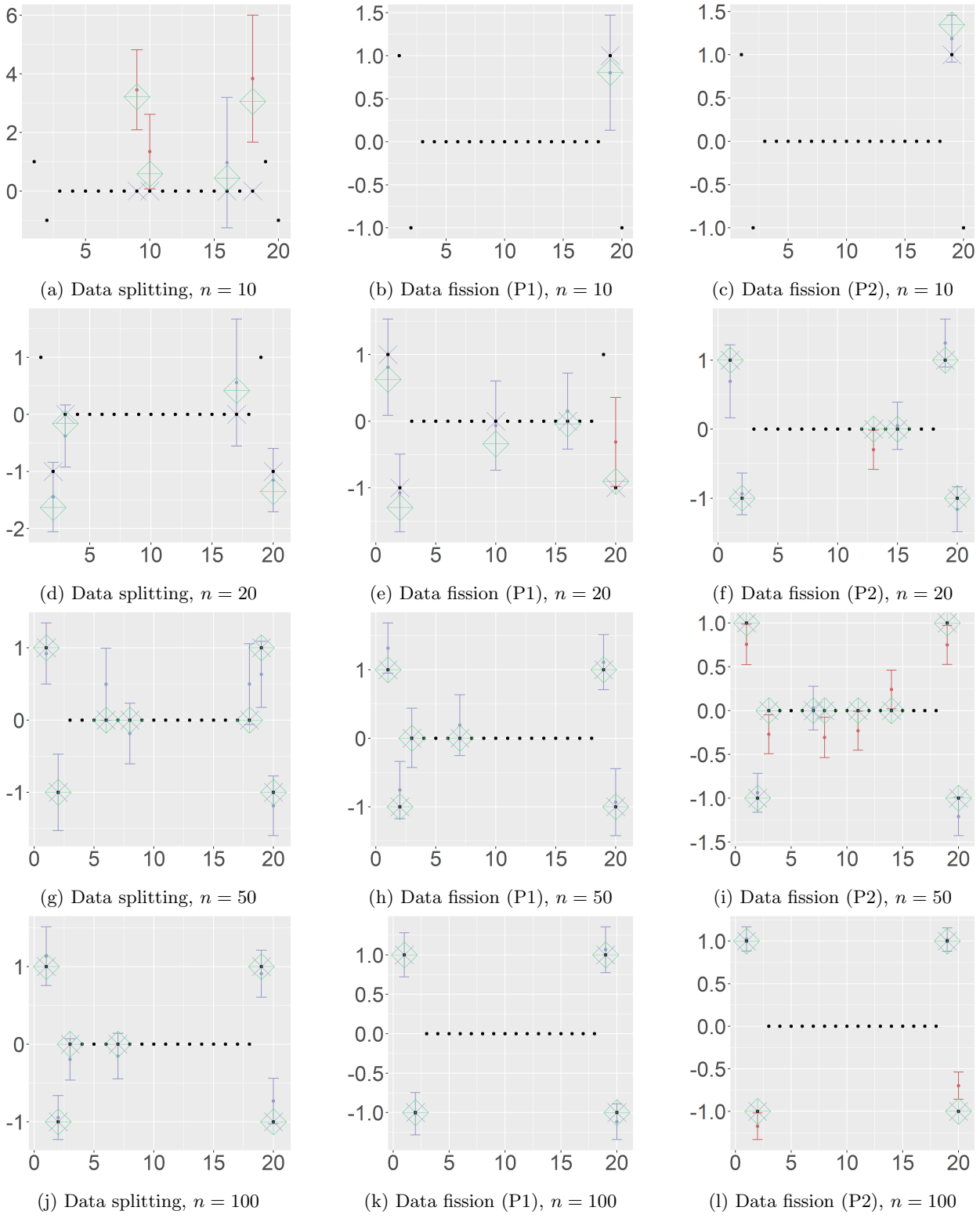
(b) Precision

(c) FCR

(d) Average CI length

(e) L2 error

Figure 2: Median and corresponding IQR of each metric across 200 runs. The x-axis shows sample size, and the y-axis shows the value of each metric. DS refers to data splitting, DF1 and DF2 refer to data fission (P1) and data fission (P2). The high IQRs for when sample size is small is likely due to many runs not selecting any variable (Table 1), all of which we excluded from the calculation.



(a) Data splitting, $n = 10$ (b) Data fission (P1), $n = 10$ (c) Data fission (P2), $n = 10$

(d) Data splitting, $n = 20$ (e) Data fission (P1), $n = 20$ (f) Data fission (P2), $n = 20$

(g) Data splitting, $n = 50$ (h) Data fission (P1), $n = 50$ (i) Data fission (P2), $n = 50$

(j) Data splitting, $n = 100$ (k) Data fission (P1), $n = 100$ (l) Data fission (P2), $n = 100$

Figure 3: An instance of the selected variables (lavender crosses) and the corresponding CIs for each experiment setting. The green diamonds indicate the targeted parameters ($\beta^\star(M)$), and the black dots indicate the true β values. CIs are drawn in red if they fail to cover $\beta^\star(M)$. The x-axis shows the index of the estimated β vector, and the y-axis shows the value of the estimate.