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

Naitong Chen

January 30, 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 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 splits the dataset into two pieces, 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 decomposes each X_i to $f_\tau(X_i)$ and $g_\tau(X_i)$ such that both parts contain information about θ, and there exists some function h such that $X_i = h(f_\tau(X_i), g_\tau(X_i))$ satisfies either of the following two properties. 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 both stages are the same between the two methods. See Example A.1 for a more detailed derivation. Here data fission can be viewed as a continuous analog of data splitting because 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 for distributions in the exponential family that satisfies (P2) 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 $f(X) = Z$ and $g(X) = X$, both the marginal distribution of $f(X)$ and the conditional distribution of $g(X)$ given $f(X)$ would be tractable, thus satisfying (P2). The general form of the distributions of $f(X)$ and $g(X) \mid f(X)$ 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 be an ideal framework for selective inference. In fact, there has been a variety of alternative procedures developed. As an example, the idea of introducing external randomness to derive valid inference procedures rather than randomly splitting the data is explore 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 $g(X) \mid f(X)$ is also known when the data are Gaussian. However, if we move beyond Gaussianity, the distribution of $g(X) \mid f(X)$ is only known asymptotically. In these cases, the use of 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, 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 $g(X) \mid f(X)$, which is similar in spirit to data carving in terms of the splitting of 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 approaches discussed.

3 Limitations and challenges

blah

4 Paper-specific project

We begin by comparing variable selection accuracy. In particular we look at power and precision (give definition). Both power and precision are similar between the two data fission procedures, because they have the same $f(X)$ marginal distributions. Compared to data splitting, data fission achieves better power and precision particularly when sample size is small, although this difference wears off as sample size increases. When there are few observations, the consequence that data splitting only works with half of the observations becomes apparent. If we look at individual trials, data splitting tends to miss true features (low power) and pick instead many false features (low precision). When there is not a lot of information from the data, halving the number of observations hinders the quality of variable selection, which is to be expected. Although data fission inflates the variance of the observations (by a factor of 2), this is a worthy sacrifice when the sample size is small. As we increase the sample size, even half of the information becomes sufficient for variable selection, making the advantage of data fission less obvious.

We now look at the quality of inference. Here we look at false coverage rate (FCR), length of CIs, and the L2 error between the estimated parameter and the ideal target of inference.

Median FCR for data splitting and data fission p1 are both well below $\alpha = 0.05$ (No correction needed, cite data fission paper). On the other hand, data fission p2 has a much higher FCR. This is because P2 not targeting the right beta star. In addition, having the CI se decrease does not help. Higher FCR could be that $(X^\top X)^-1$ converges, but randomness in $f(y)$ does not change. Worth noting that having low FCR does not necessarily mean we recover the true parameters, because there is a discrepancy between true and targeted beta's, especially when sample size is small.

Length of CI: P2 smaller than the other two. P1 and splitting are similar. P1 and splitting similar, probably because halving data and inflated variance cancel each other out. P2 has smaller length because variance smaller and sample size doesn't change.

L2 error decays by the variance of beta hat. Data splitting is worse because halving data. Obvious when sample size is small, and less so when sample size is large. Similarities between P1 and P2 somewhat surprising. might need to inflate sigma2 to make things clearer? Need further investigation.

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.

5 Discussion

References

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 $\{X_i\}_{i=1}^n$ *be i.i.d.* $\mathcal{N}(\theta, \sigma^2)$*. Let* $X := \big[X_1, \ldots, X_n\big]^\top$*. Data splitting defines* $f(X)$ *and* $g(X)$ *as*

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

for $a \in \big\{\frac{1}{n}, \frac{2}{n}, \ldots, 1\big\}$*. Note that*

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

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* $\tau \in (0, \infty)$*,*

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

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. ◁