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

Naitong Chen

January 26, 2023

1 Problem Definition

Given a dataset $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \pi_\theta$, with π_θ being a distribution from the exponential family whose parameter θ is of interest. We decompose each X_i to $f(X_i)$ and $g(X_i)$ such that both parts contain information about θ, and there exists some function h such that $X_i = h(f(X_i), g(X_i))$ satisfying either of the two properties:

- (P1): $f(X_i)$ and $g(X_i)$ are independent with known distributions (up to unknown θ);
- (P2): $f(X_i)$ has a known marginal distribution and $g(X_i)$ has a known conditional distribution given $f(X_i)$ (up to unknown θ).
  Typos and missing terms in the proof of the original paper. See Theorem A.2 for a modified version.

2 Significance

3 Limitations and challenges

- paper does not discuss robustness of the method to the distribution assumptions
- experiments do not cover cases where fissioned data get transformed to follow a different distribution
- paper does not discuss how to choose tuning parameter (controlling amount of information split between $f(X)$ and $g(X)$)
- following the previous point, this paper does not discuss the relations between having a discrete vs. continuous tuning parameter (e.g. the two different ways of fissioning exponentially distributed data in Appendix B of Leiner et al. (2022))
- Proofs have missing terms, inaccurate notations.
- The discrete version of data fission for Gamma data may be incorrect.

4 Paper-specific project

In this section we compare two data fission procesures for Gaussian random variables, one P1 and one P2, against data splitting in the context of selective CIs in fixed-design linear regression. In particular, we compare and contrast these methods at both the selection step and the inference step.

We work under the setting of ... (specify linear regression form)

List three fisison methods and their corresponding distributions for regression parameters.

From the above, note that the difference comes from tau and the size of the data (which affects convergence of $(X_M^\top X_M)^{-1}$), and so for simplicity, we fix $\Sigma = \sigma^2 I$ with $\sigma = 1$.

Data splitting: halving data, inflating variance (through $X^\top X$), but targets right beta star.

Data fission p1: targets right beta star, but variance inflated by external randomness. ¡- infernce. selection -¿ inflated variance.

Data fission p2': variance deflated, but mean now no longer targeting beta star. ¡- infernce. selection -¿ inflated variance.

Overall, trading off the number of observations and either a) inflated variance or b) having a biased target instead of beta star. We show their consequences in a simulation study below.

To make the comparison fair, we choose tau and a so that the Fisher information allocated to f(Y) is the same.

As a result, we compare how these three methods perform across different dataset sizes.

We follow procedure of paper, do variable selection using LASSO with default setting in glmnet. We run 200 trials and report the median for each metric. Plots with a measure of uncertainty are in the appendix.

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.

In addition, we look at the L2 error between ideal target of inference and the actual targeted parameters for data fission P2.

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

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

A Proofs and examples

Example A.1. *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 $X := \left[X_1, \dots, X_n\right]^\top$. Recall that data splitting defines $f(X)$ and $g(X)$ as

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

for $a \in \left\{\frac{1}{n}, \frac{2}{n}, \dots, 1\right\}$. 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 $\tau \in (0, \infty)$,

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

Note that for all $i \in \{1, \dots, n\}$, $X_i + \tau Z_i \sim \mathcal{N}\left(\theta, (1+\tau^2)\sigma^2\right)$, $X_i - \frac{1}{\tau}Z_i \sim \mathcal{N}\left(\theta, (1+\frac{1}{\tau^2})\sigma^2\right)$. 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 $Z = (Z_1, \cdots, Z_B)$ 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 $B = 1$,

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

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.4. *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,*

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

◁

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, with* $B = 1$, *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*

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

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

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

which may be convex in β *but without an argmin, or non-convex (depending on the value of* α*). Therfore, we can instead use the working model*

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

For this problem, we have that

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

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

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

then the plug-in estimator for variance becomes

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

$\triangleleft$

Example A.6. *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.4, 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*

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

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

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

which may be convex in β but without an argmin, or non-convex (depending on the value of α). Therefore, we can instead use the working model

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

For this problem, we have that

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

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

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

then the plug-in estimator for variance becomes

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

$\triangleleft$

No need to ensure $\exp(\beta^{\top} x_{i,\mathcal{M}})$ stays above 1 or τ, since mean already have same support?