

Variational Inference for Spatial Correlated Failure Time Data under Bayesian Framework

Yueyao Wang¹

The 2024 JRC
June 19th, 2024

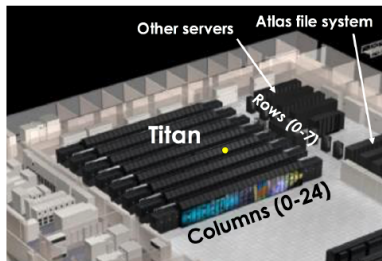
Joint work with Yili Hong², Xinwei Deng², and Laura Freeman²

¹School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, PRC

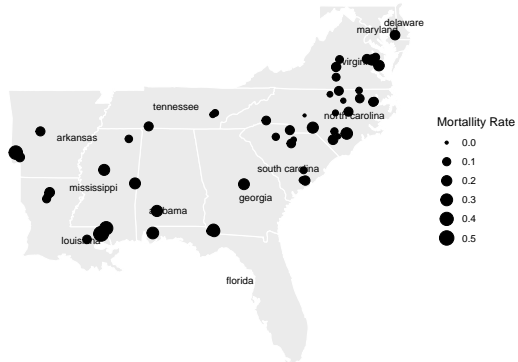
²Department of Statistics, Virginia Tech, Blacksburg, VA

- In the modern survival analysis, the geo-graphically referenced time-to-event data are often collected for analysis.
- The **spatial dependence** among survival times needs to be properly accounted using **spatial survival models**.
- The Markov Chain Monte Carlo (MCMC) methods for Bayesian framework can be time-consuming when the number of spatial locations is large.
- We investigate the capability of an approximate approach, **variational inference** (VI) for the inference of spatial survival models.
- We focus on two models, the **proportional hazards model** and the **cumulative exposure model**.

The Motivating Examples



(a) The Titan GPU Failure Time Data



(b) Pine Tree Survival Data

Suppose there are m distinct locations s_1, \dots, s_m .

- Let t_{ij} be the observed event time for the j th unit in the i th location s_i , where $i = 1, \dots, m, j = 1, \dots, n_i$.
- Let δ_{ij} be the corresponding censoring indicator.
- Denote $\mathbf{x}_{ij}(t)$ to be the p -dimensional vector of related covariates at time t .
- $\mathcal{D} = \{t_{ij}, \delta_{ij}, \mathbf{x}_{ij}(t) : t \leq t_{ij}, i = 1, \dots, m, j = 1, \dots, n_i\}$.

Spatial Cumulative Exposure Model (CEM)

- The **cumulative damage level** by a certain time t given time-varying covariates $\mathbf{x}(t)$:

$$u_{ij}(t) = \int_0^t \exp \left[-\mathbf{x}_{ij}(s)^\top \boldsymbol{\beta} \right] ds.$$

- The $\log[u(T_{ij})]$ is assumed to follow a location-scale distribution.

$$\log [u_{ij}(T_{ij})] = \mu + \gamma_i + \sigma \epsilon_{ij}.$$

- γ_i : spatial random parameter; ϵ_{ij} follows the standard location-scale distribution.
- Note that when the covariates are constant, the CEM can be simplified to an accelerated failure time (AFT) model: $\log(T_{ij}) = \mu + \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \gamma_i + \sigma \epsilon_{ij}$.

Proportional Hazards Model (PH) with Spatial Effect

- The hazard function of the j th unit in i th location is modeled as

$$h_{ij}(t) = h_0(t) \exp \left[\mathbf{x}_{ij}(t)^\top \boldsymbol{\beta} + \gamma_i \right],$$

- $h_0(t)$: the baseline hazard function; $\boldsymbol{\beta}$: the coefficient vector for covariates; γ_i : the spatial random effect at location s_i .
- A parametric baseline hazard function $h_0(t; \boldsymbol{\theta}_h)$ is used, where $\boldsymbol{\theta}_h$ is the parameter vector. E.g., Weibull hazard function $h_0(t) = at^b$ with $\boldsymbol{\theta}_h = (a, b)^\top$.

The Spatial Random Effect

- Spatial random vector: $\gamma = (\gamma_1, \dots, \gamma_m)^\top$

$$\gamma \sim \text{MVN}(\mathbf{0}, \Sigma), \text{ where } \Sigma = \sigma_\gamma^2 \Omega.$$

- σ_γ^2 : the overall spatial variability.
- $\Omega = (\rho_{i,i'})_{m \times m}$: the correlation matrix.
- $\rho_{i,i'}$: the spatial correlation between the random effect of location s_i and $s_{i'}$.
- The exponential correlation function is used

$$\rho_{i,i'} = \exp[-d(s_i, s_{i'})/\nu], \nu > 0,$$

- $d(s_i, s_{i'})$: the Euclidean distance between locations s_i and $s_{i'}$.
- ν : the length scale parameter that describes the rate of decay of correlations.

Bayesian Framework for Spatial Survival Model

$$\begin{aligned}\gamma &\propto \text{MVN}(\mathbf{0}, \sigma_\gamma^2 \Omega), \\ \sigma_\gamma^2 &\propto \text{IGAM}(a_\sigma, b_\sigma), \\ \nu &\propto \text{IGAM}(a_\nu, b_\nu), \\ \beta_p &\propto \mathbf{1}_p.\end{aligned}$$

For spatial CEM:

$$\begin{aligned}\log [u_{ij}(T_{ij})] &= \mu + \gamma_i + \sigma \epsilon_{ij}, \\ \mu &\propto \mathbf{1}, \quad \sigma_l \propto \mathbf{1}.\end{aligned}$$

For spatial PH:

$$\begin{aligned}h_{ij}(t) &= h_0(t; \theta_h) \exp \left[\mathbf{x}_{ij}(t)^\top \boldsymbol{\beta} + \gamma_i \right], \\ \theta_h &\propto \mathbf{1}.\end{aligned}$$

- The key idea of VI is to use a relative simple distribution, variational distribution: $q(\theta|\eta)$ to approximate the exact posterior $p(\theta|\mathcal{D})$.
- Here η is the parameter vector in the variational probability distribution.
- Then a metric that evaluates the distance between two distributions $p(\theta|\mathcal{D})$ and $q(\theta|\eta)$ is optimized to obtain the estimate of η .

KL Divergence

$$\text{KL}[q(\theta|\eta)||p(\theta|\mathcal{D})] = \int q(\theta|\eta) \log \left[\frac{q(\theta|\eta)}{p(\theta|\mathcal{D})} \right] d\theta$$

The Evidence Lower Bound (ELBO):

$$\mathcal{L}_{\text{VI}} = \log[p(\mathcal{D})] - \text{KL}(q(\theta|\eta)||p(\theta|\mathcal{D})) = \mathbb{E}_{q(\theta|\eta)} \left\{ \log \left[\frac{p(\theta, \mathcal{D})}{q(\theta|\eta)} \right] \right\}.$$

α -Divergence: $\alpha \rightarrow 1$, α -divergence \rightarrow KL divergence.

$$D[q(\theta|\eta)||p(\theta|\mathcal{D})] = \frac{1}{\alpha - 1} \log \left[\int q(\theta|\eta)^\alpha p(\theta|\mathcal{D})^{1-\alpha} d\theta \right].$$

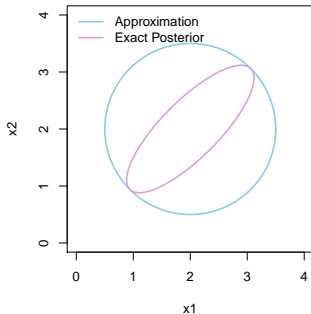
The variational R nyi (VR) bound:

$$\mathcal{L}_\alpha = \log[p(\mathcal{D})] - D[q(\theta|\eta)||p(\theta|\mathcal{D})] = \frac{1}{\alpha - 1} \log \left\{ \mathbb{E}_{q(\theta|\eta)} \left[\left(\frac{p(\theta, \mathcal{D})}{q(\theta|\eta)} \right)^{1-\alpha} \right] \right\}.$$

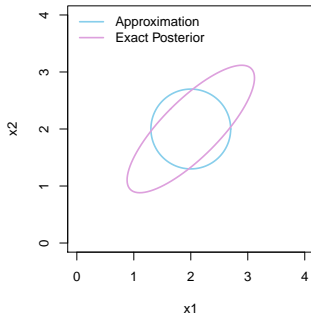
α —Divergence Characteristics

Mass Covering

$$\alpha \rightarrow -\infty$$

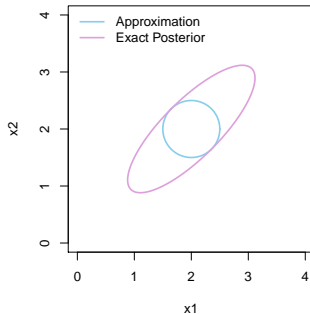


$$\alpha \rightarrow 1$$



Zero Forcing

$$\alpha \rightarrow \infty$$



Require: $q(\theta|\eta)$: the variational distribution.

Require: η_0 : initial variational parameter vector.

$r \leftarrow 1$ (initialize iteration number);

$\eta \leftarrow \eta_0$ (initialize variational parameter vector);

1: **while** not converging **do**

2: Take h samples from the variational distribution $\theta_k \sim q(\theta|\eta_{r-1}), k = 1, \dots, h$
and compute a stochastic estimate of $\hat{\mathcal{L}}_\alpha$.

3: Take a gradient descent step in Adam algorithm to update η_r ;

4: $r \leftarrow r + 1$

5: **end while**

6: **return** $\eta^* \leftarrow \eta_r$

GPU Lifetime Dataset

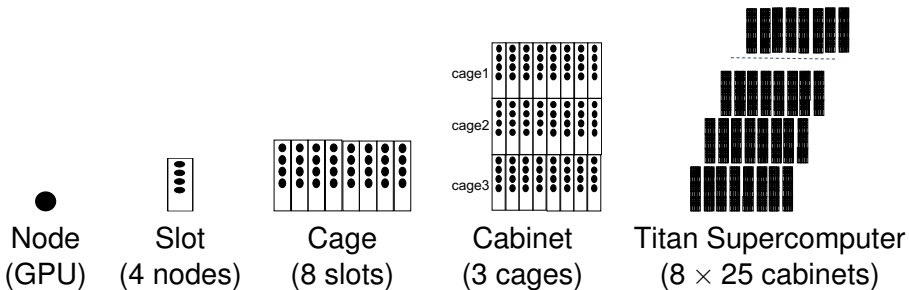


Figure: The physical organization of Titan supercomputer.

- We use a subset of the data, which includes the units in row number 0-7 and column number 1-13.
- The row and column positions of each unit are considered as the location information.
- The node, slot, and cage information are considered as covariates that can affect GPU's lifetime.
- We build a spatial AFT model to study the failures of GPU.
- Three inference methods, Hamiltonian Monte Carlo (HMC), KL-divergence and α -divergence with $\alpha = 0.8$ performance are compared.

$$\log(T_{ij}) = \mu + \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \gamma_i + \sigma \epsilon_{ij}.$$

- The variational distribution assumptions of $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \gamma^\top, \sigma_l, \sigma_\gamma^2, \nu)^\top$ is:

$$\boldsymbol{\beta} \sim \text{MVN}(\boldsymbol{\mu}_\beta, \Sigma_\beta), \text{ where } \Sigma_\beta = \text{Diag}(\boldsymbol{\sigma}_\beta^2),$$

$$\gamma \sim \text{MVN}(\boldsymbol{\mu}_\gamma, \Sigma_\gamma), \text{ where } \Sigma_\gamma = \text{Diag}(\boldsymbol{\sigma}_\gamma^2),$$

$$\sigma_l \sim \text{N}(\mu_\sigma, \sigma_\sigma^2), \text{ where } \sigma_l = \log(\sigma)$$

$$\sigma_\gamma^2 \propto \text{IGAM}(\mathbf{c}_\sigma, \mathbf{d}_\sigma),$$

$$\nu \propto \text{IGAM}(\mathbf{c}_\nu, \mathbf{d}_\nu).$$

- The variational distribution is:

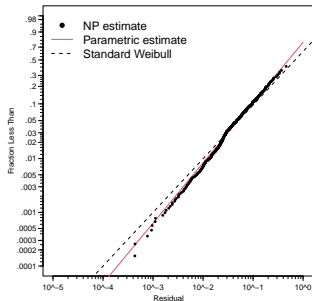
$$q(\boldsymbol{\theta}|\boldsymbol{\eta}) = f_{\text{MVN}}(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \Sigma_\beta) f_{\text{MVN}}(\gamma|\boldsymbol{\mu}_\gamma, \Sigma_\gamma) f_{\text{N}}(\sigma_l|\mu_\sigma, \sigma_\sigma^2) f_{\text{IGAM}}(\sigma_\gamma^2|\mathbf{c}_\sigma, \mathbf{d}_\sigma) f_{\text{IGAM}}(\nu|\mathbf{c}_\nu, \mathbf{d}_\nu).$$

- The censored Cox-Snell residual of spatial AFT model is an extension of the standardized residual, which is defined as

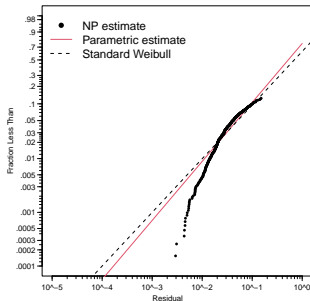
$$\hat{\epsilon}_{ij} = \frac{\log(t_{ij}) - \mathbf{x}_{ij}^{\top} \hat{\boldsymbol{\beta}} - \gamma_i}{\hat{\sigma}}.$$

- With model assumptions, the residuals should approximately follow a Weibull distribution.

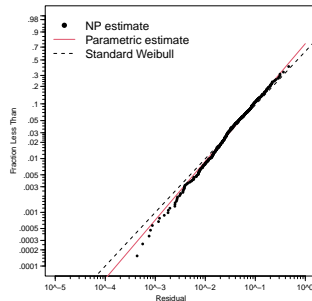
Residual Plot for GPU Lifetime Data



(a) $\alpha = 0.8$



(b) KL



(c) HMC

Figure: Weibull probability plot of residuals for α -divergence, KL divergence and HMC with the GPU data.

Table: The negative log likelihood and computing time of α -divergence, KL divergence and HMC inferences with the GPU data.

	$\alpha = 0.8$	KL	HMC
NLL	2034.93	2405.40	2040.68
Time (minutes)	7.28	10.07	20.89

The Computing Time Versus Number of Locations

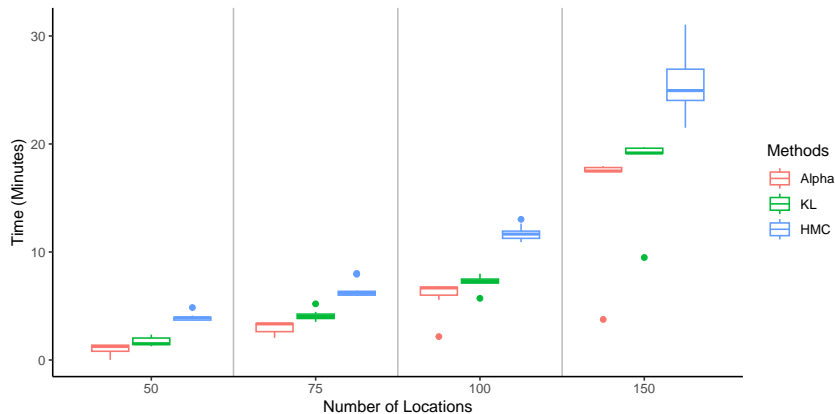


Figure: The computing time of three inference methods versus the number of locations with the GPU data.

Pine Tree Survival Data

- The survival and growth of trees with different living conditions and the thinning treatments are of interest.
- During each tree's lifetime, variables such as total height (TH), diameter at breast height (DBH), and crown class are recorded every three years up to 7 times.
- The event of interest is the death of a tree. A tree is recorded as censored if it survives till the 7th follow-up period.
- Due to the computational limitation, we randomly select 60 sites from the original dataset with 13,911 trees.
- A PH model is considered to model the survival rate of pine trees with explanatory variables.

- We assume the following variational distribution assumptions:

$$\beta \sim \text{MVN}(\mu_\beta, \Sigma_\beta), \text{ where } \Sigma_\beta = \text{Diag}(\sigma_\beta^2),$$

$$\gamma \sim \text{MVN}(\mu_\gamma, \Sigma_\gamma), \text{ where } \Sigma_\gamma = \text{Diag}(\sigma_\gamma^2),$$

$$a_l \sim \text{N}(\mu_a, \sigma_a^2),$$

$$b_l \sim \text{N}(\mu_b, \sigma_b^2),$$

$$\sigma_\gamma^2 \propto \text{IGAM}(c_\sigma, d_\sigma),$$

$$\nu \propto \text{IGAM}(c_\nu, d_\nu),$$

- The variational distribution is:

$$q(\theta|\eta) = f_{\text{MVN}}(\beta|\mu_\beta, \Sigma_\beta) f_{\text{MVN}}(\gamma|\mu_\gamma, \Sigma_\gamma) f_{\text{N}}(a_l|\mu_a, \sigma_a^2) f_{\text{N}}(b_l|\mu_b, \sigma_b^2) f_{\text{IGAM}}(\sigma_\gamma^2|c_\sigma, d_\sigma) f_{\text{IGAM}}(\nu|c_\nu, d_\nu)$$

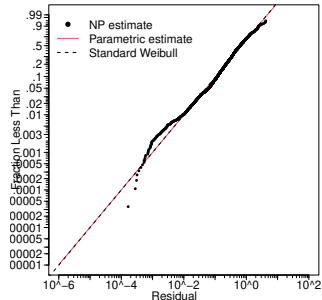
- For a PH model with time-dependent covariates, the Cox-Snell residual is defined as

$$\hat{\epsilon}_{ij} = \hat{H}_0(T_{ij}) \int_0^{T_{ij}} \exp[\mathbf{x}_{ij}(t)^\top \hat{\boldsymbol{\beta}} + \hat{\gamma}_i] dt,$$

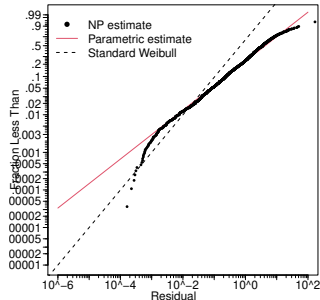
where $\hat{H}_0(T_{ij})$ is the estimated cumulative baseline hazard rate by plugging in \hat{a} and \hat{b} .

- If the model is correct, then $\hat{\epsilon}_{ij}$ approximately follows exponential distribution with $\lambda = 1$ and censoring, which is a special case of Weibull distribution.

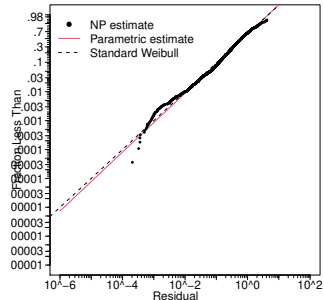
Pine Tree Survival Data Analysis Results



(a) $\alpha = 0.8$



(b) KL



(c) HMC

Figure: Weibull probability plot of residuals for α -divergence, KL divergence and HMC with the pine tree data.

Pine Tree Survival Data Analysis Results

Table: The negative log likelihood and computing time of α -divergence, KL divergence and HMC inferences with pine tree data.

	$\alpha = 0.8$	KL	HMC
NLL	12341.71	18985.41	12332.84
Time (hours)	7.97	8.52	16.29

Conclusions and Future Directions

- Compared to KL divergence, α -divergence encourages a more flexible variational distribution, thus it has better performance regarding statistical inference.
- Based on these two applications, we find α -divergence with $\alpha < 1$ has comparable performance as HMC but with better computational efficiency.
- In the future, it will be interesting to study how the statistical inference performance changes with different α values and how to choose α .

Thank you!