Variational Inference for Spatial Correlated Failure Time Data under Bayesian Framework

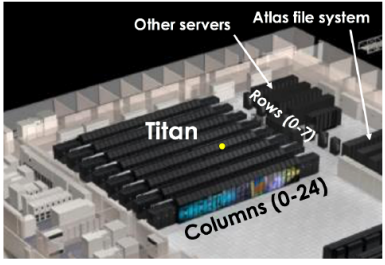
Yueyao Wang[1]

The 2024 JRC
June 19th, 2024

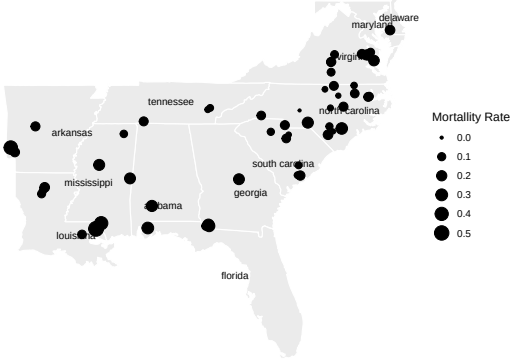
Joint work with Yili Hong[2], Xinwei Deng[2], and Laura Freeman[2]

[1]School of Statistics and Mathematics, ZhejiangGongshang University, Hangzhou, PRC
[2]Department of Statistics, Virginia Tech, Blacksburg, VA

The Motivating Examples



(a) The Titan GPU Failure Time Data



(b) Pine Tree Survival Data

Background

- The **spatial dependence** among spatial survival times needs to be properly accounted with **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) in terms of **computational efficiency** and **statistical inference performance**.
- We focus on two models, the **proportional hazards model** and the **cumulative exposure model**.

Data Notations

Suppose there are m distinct locations $s_1, \ldots, s_m$.

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

Spatial Cumulative Exposure Model (CEM)

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

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

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

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

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

Spatial Proportional Hazards Model (PH)

- The hazard function of the jth unit in ith location is modeled as

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

- $h_0(t)$: 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: $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)^\top$

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

- σ_γ^2: the overall spatial variability.
- $\Omega = \left(\rho_{i,i'}\right)_{m \times m}$: 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

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

For spatial CEM:

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

For spatial PH:

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

Variational Inference

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

Divergences

KL Divergence

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

The Evidence Lower Bound (ELBO):

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

$\alpha-$Divergence: $\alpha \to 1$, $\alpha-$divergence $\to$ KL divergence.

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

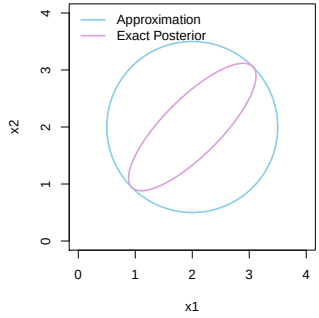
The variational Rènyi (VR) bound:

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

$\alpha-$Divergence Characteristics

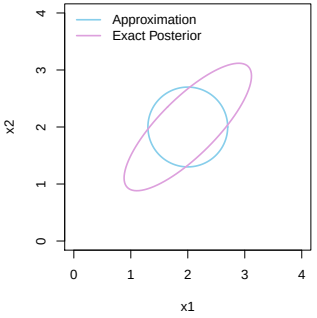
Mass Covering

$\alpha \to -\infty$



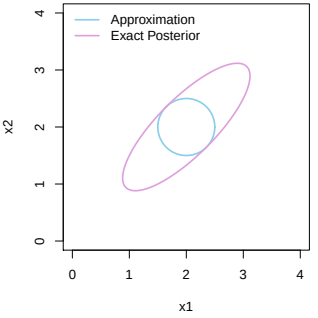
$\Rightarrow$

$\alpha \to 1$



Zero Forcing

$\alpha \to \infty$



Variational Parameter Estimation in VI

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 $\theta_k \sim q(\theta|\eta_{r-1}), k = 1, \ldots, h$ and compute a stochastic estimate of $\widehat{\mathcal{L}}_\alpha$.
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

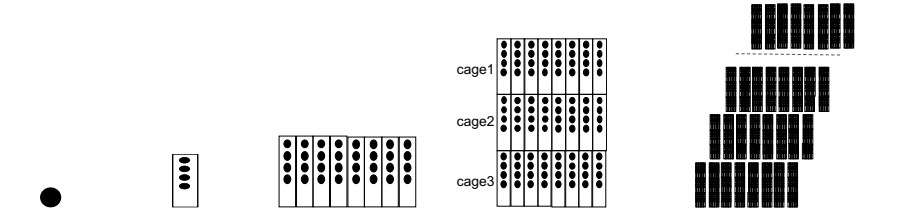


Figure: The physical organization of Titan supercomputer.

GPU Lifetime Dataset

- 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 $\alpha-$divergence with $\alpha = 0.8$ performance are compared.

VI for AFT model with GPU Data

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

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

$$\begin{aligned}
\boldsymbol{\beta} &\sim \text{MVN}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}), \text{where } \Sigma_{\boldsymbol{\beta}} = \text{Diag}(\boldsymbol{\sigma}_{\boldsymbol{\beta}}^2),\\
\boldsymbol{\gamma} &\sim \text{MVN}(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \Sigma_{\boldsymbol{\gamma}}), \text{where } \Sigma_{\boldsymbol{\gamma}} = \text{Diag}(\boldsymbol{\sigma}_{\boldsymbol{\gamma}}^2),\\
\sigma_l &\sim \text{N}(\mu_{\sigma}, \sigma_{\sigma}^2), \text{where } \sigma_l = \log(\sigma)\\
\sigma_{\gamma}^2 &\propto \text{IGAM}(c_{\sigma}, d_{\sigma}),\\
\nu &\propto \text{IGAM}(c_{\nu}, d_{\nu}).
\end{aligned}$$

- The variational distribution is:

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

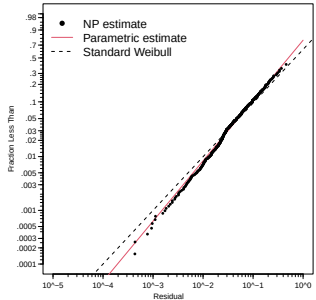
The Cox-Snell Residual

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

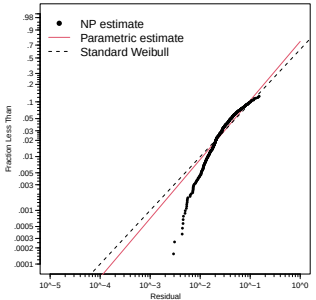
$$\widehat{\epsilon}_{ij} = \frac{\log(t_{ij}) - \boldsymbol{x}_{ij}^{\top}\widehat{\boldsymbol{\beta}} - \gamma_i}{\widehat{\sigma}}.$$

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

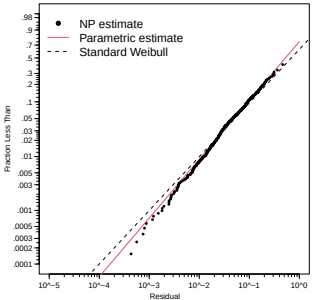
Residual Plot for GPU Lifetime Data



(a) $\alpha = 0.8$



(b) KL



(c) HMC

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

Methods Comparisons with GPU Data

Table: The negative log likelihood and computing time of $\alpha-$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

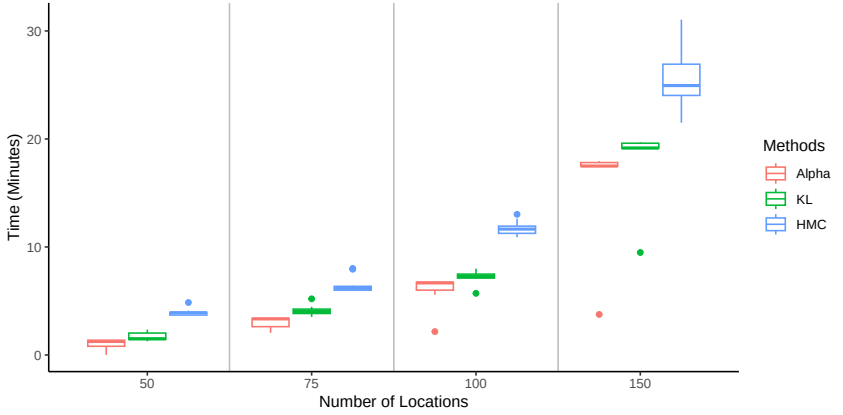


Figure: The computing time of three inference methods versus the number of locations with the GPU data.

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.

VI for the Bayesian PH Model

- We assume the following variational distribution assumptions:

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \text{MVN}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}), \text{where } \Sigma_{\boldsymbol{\beta}} = \text{Diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\beta}}), \\
\boldsymbol{\gamma} &\sim \text{MVN}(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \Sigma_{\boldsymbol{\gamma}}), \text{where } \Sigma_{\boldsymbol{\gamma}} = \text{Diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\gamma}}), \\
a_l &\sim \text{N}(\mu_a, \sigma^2_a), \\
b_l &\sim \text{N}(\mu_b, \sigma^2_b), \\
\sigma^2_\gamma &\propto \text{IGAM}(c_\sigma, d_\sigma), \\
\nu &\propto \text{IGAM}(c_\nu, d_\nu),
\end{aligned}
$$

- The variational distribution is:

$$q(\boldsymbol{\theta}|\boldsymbol{\eta}) = f_{\text{MVN}}(\beta|\boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}) f_{\text{MVN}}(\boldsymbol{\gamma}|\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \Sigma_{\boldsymbol{\gamma}}) f_{\text{N}}(a_l|\mu_a, \sigma^2_a) f_{\text{N}}(b_l|\mu_b, \sigma^2_b) f_{\text{IGAM}}(\sigma^2_\gamma|c_\sigma, d_\sigma) f_{\text{IGAM}}(\nu|c_\nu, d$$

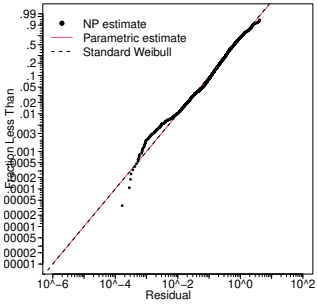
The Cox-Snell Residual

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

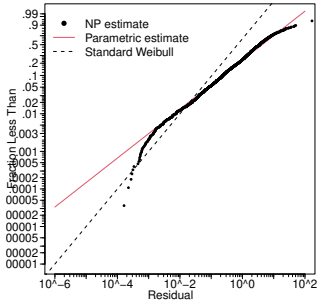
$$\widehat{\epsilon}_{ij} = \widehat{H}_0(T_{ij}) \int_0^{T_{ij}} \exp[\boldsymbol{x}_{ij}(t)^\top \widehat{\boldsymbol{\beta}} + \widehat{\gamma}_i] dt,$$

  where $\widehat{H}_0(T_{ij})$ is the estimated cumulative baseline hazard rate by plugging in $\widehat{a}$ and $\widehat{b}$.
- If the model is correct, then $\widehat{\epsilon}_{ij}$ approximately follows exponential distribution with $\lambda = 1$ and censoring, which is a special case of Weibull distribution.

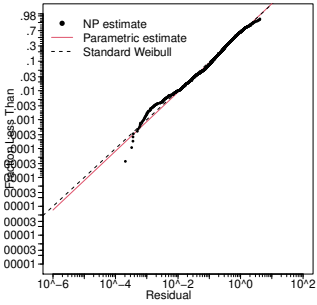
Pine Tree Survival Data Analysis Results



(a) $\alpha = 0.8$



(b) KL



(c) HMC

Figure: Weibull probability plot of residuals for $\alpha-$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 $\alpha-$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, $\alpha-$divergence encourages a more flexible variational distribution, thus it has better performance regarding statistical inference.
- Based on these two applications, we find $\alpha-$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.

Questions

Thank you!