The Sensitivity of Priors in Bayesian Variable Selection for Parametric AFT Models in High Dimensions

Md. Hasinur Rahaman Khan and Ewart Shaw

Department of Statistics
University of Warwick, UK

$3o^{th}$ ISCB Conference, August 26, 2009
PRAGUE

Outline

Outline

What is Parametric AFT Model?

General form of **Accelerated Failure Time** (AFT) model is

$$\log(T_i) = \alpha + X_i'\beta + \varepsilon_i, \quad i = 1, ..., n \qquad (1)$$

Where,

- T is the survival time
- α is the intercept
- X_i' is a p-vector of covariates
- β is the vector of regression parameters
- ϵ_i's are iid r.vs. which has **parametric** form (or unspecified).

What is Parametric AFT Model?

General form of **Accelerated Failure Time** (AFT) model is

$$\log(T_i) = \alpha + X_i'\beta + \varepsilon_i, \quad i = 1, ..., n \qquad (1)$$

Where,

- T is the survival time
- α is the intercept
- X_i' is a p-vector of covariates
- β is the vector of regression parameters
- ϵ_i's are iid r.vs. which has **parametric** form (or unspecified).

Why AFT Model?

- Model procedures have **sound theoretical justification.**
- Can be implemented with an **efficient numerical method.**
- Has intuitive **physical interpretation.**
- It is parametric.
  - has **simplicity**
  - the **availability** of likelihood-based inference procedures
  - ease of use for **description/comparision/prediction** or **decision**

FOR DETAILS: See the paper by **Wei** (1992).

Why AFT Model?

- Model procedures have **sound theoretical justification.**
- Can be implemented with an **efficient numerical method.**
- Has intuitive **physical interpretation.**
- It is parametric.
  - has **simplicity**
  - the **availability** of likelihood-based inference procedures
  - ease of use for **description/comparision/prediction** or **decision**

FOR DETAILS: See the paper by **Wei** (1992).

What are the AFT Models?

AFT models are defined based on the distributional assumption of $Log(T_i)$ or alternatively

- **Log-normal** if ϵ has Normal distribution.
- **Log-t** if ϵ has t-distribution.
- **Log-logistic** if ϵ has Logistic distribution.
- **Weibull** if ϵ has Extreme Value distribution with 2 parameters.
- **Exponential** if ϵ has Extreme Value distribution with 1 parameter.
- **Gamma** if ϵ has Log-gamma distribution.

What are the AFT Models?

AFT models are defined based on the distributional assumption of $Log(T_i)$ or alternatively

- **Log-normal** if ϵ has Normal distribution.
- **Log-t** if ϵ has t-distribution.
- **Log-logistic** if ϵ has Logistic distribution.
- **Weibull** if ϵ has Extreme Value distribution with 2 parameters.
- **Exponential** if ϵ has Extreme Value distribution with 1 parameter.
- **Gamma** if ϵ has Log-gamma distribution.

Outline

Variable Selection Problem in Survival Analysis

- The main problem of any variable selection (subset selection) arises when one wants to model the relationship between response and a subset of $X_1, ..., X_p$ but **there is uncertainty** about which subset to use.
- This is absolutely true when $p \gg n$ (e.g. p=10,000, n=100).
- This is more absolutely true when number of potential significant variables is small (e.g. $p_\gamma = 10$).
- The above problems arise nowadays in many fields with high-dimensional survival data:
  - Genomic (Microarray Gene expression data)
  - Proteomic
  - Meta-data (e.g. networks and pathways)

Variable Selection Problem in Survival Analysis

- The main problem of any variable selection (subset selection) arises when one wants to model the relationship between response and a subset of $X_1, ..., X_p$ but **there is uncertainty** about which subset to use.
- This is absolutely true when $p \gg n$ (e.g. p=10,000, n=100).
- This is more absolutely true when number of potential significant variables is small (e.g. $p_\gamma = 10$).
- The above problems arise nowadays in many fields with high-dimensional survival data:
  - Genomic (Microarray Gene expression data)
  - Proteomic
  - Meta-data (e.g. networks and pathways)

Bayesian Variable Selection in Survival Analysis

- With 2^p possible subsets (when $p \gg n$) standard frequentist methods for the variable selection process in survival analysis can be **unstable** and result in **over-fitting**.
- But Bayesian methods with the help of MCMC can make the analysis **stable** and **more convenient.**
- With $p \gg n$ settings Bayesian approach can deal with
  - Conjugate priors
  - Non-conjugate priors (Involve complex MCMC and intensive computations)
  - Others: Co-linearity among covariates, Censoring etc.

Outline

Prior Specification in Variable Selection

Specifying meaningful prior distributions for parameters of **each model** and **models themselves** is a difficult task because

- It requires contextual interpretations of a large number of parameters.
- There is no unique way of choosing a prior dist. and that the resulting inference may be influenced.
- Generally there be some uncertainty in the choice of prior, especially when there is little information from which to construct such dist.
- Even in the absence of strong prior information prior specification has to be done at the appropriate scale of biological interest.

Outline

Main Purpose and Assumptions

We work with Log-normal AFT model (ε_i's $\sim N(0, \sigma^2)$ in Equation (1)).

- This work builds on ideas in **Sha et al.** (2006).
- This research is a further extension of their work where mainly the sensitivity of prior specification for the parameters has been investigated with simulated data (when $p \gg n$) for the underlying method.
- There is *Type-I* right censoring in the data.
- Out of p variables, only few variables (p_γ) are assumed to be related to $Log(T_i)$.
- It is also assumed there is co-linearity (correlation) between p_γ and $(p - p_\gamma)$

Main Purpose and Assumptions

We work with Log-normal AFT model (ε_i's $\sim N(0, \sigma^2)$ in Equation (1)).

- This work builds on ideas in **Sha et al.** (2006).
- This research is a further extension of their work where mainly the sensitivity of prior specification for the parameters has been investigated with simulated data (when $p \gg n$) for the underlying method.
- There is *Type-I* right censoring in the data.
- Out of p variables, only few variables (p_γ) are assumed to be related to $Log(T_i)$.
- It is also assumed there is co-linearity (correlation) between p_γ and $(p - p_\gamma)$

Basic Methods

- **Data augmented approach** (**Tanner and Wong**, 1987) is used to impute the censored data.
  - The augmented data, $w_i = \log(t_i)$ is generated so that

$$\begin{cases} w_i = \log(t_i^*) & \text{if } \delta_i = 1 \\ w_i > \log(t_i^*) & \text{if } \delta_i = 0 \end{cases}$$

  where, $t_i^* = \min(t_i, c_i)$, $\delta_i = I(t_i \leq c_i)$, and c_i is the censoring time.

- Then full data are normally distributed i.e. $W|X, \alpha, \beta, \sigma^2 \sim N(\alpha J + X\beta, \sigma^2 I)$.

Outline

Prior Distributions for Model Parameters $(\alpha, \beta, \sigma^2)$

- Conjugate priors are used for model parameters-
  - $\alpha|\sigma^2 \sim N(\alpha_0, h_0\sigma^2)$
  - $\beta|\sigma^2 \sim N(\beta_0, \sigma^2 \sum_0)$
  - $\sigma^2 \sim IG(v_0/2, v_0\sigma_0^2/2)$
- Vague or **weakly informative priors** are obtained for α and β specifying hyperparameters as:
  - $\alpha_0 = 0$, h_0 large,
  - $\beta_0 = 0$ and $\Sigma_0 = hI$,
  - here h **regulates the amount of shrinkage** in the model and I is the identity matrix.
- **Weakly informative prior** is obtained for σ^2 specifying a small value for v_0.

Model Space Priors for Variable Selection

This study follows **mixture prior method** (**George and McCulloch** (1993)). Accordingly, the β's arise from mixture of a point mass at 0 and a normal density i.e.

$$\beta_j|\gamma_j, \sigma^2 \sim (1-\gamma_j)I(0) + \gamma_j N(0, \sigma^2\tau_j), \qquad j = 1, ..., p,$$

where

- γ is a latent *p*-vector with binary entries: 1 and 0 depending on whether the variable is in the model or not respectively.
- τ_j is the *j*-th diagonal element of Σ_0.
- γ_j's are iid Bernoulli random variables with ω elicited as the proportion of variables expected a *priori* in the model.

Outline

Simulation Strategy

For simulation we follow the strategy described in **Gui and Li** (2005). Total variables= p, potential predictors= p_γ, co-linearity exists between p_γ and $p - p_\gamma$.

- $A_{n \times p} \sim U(-1.5, 1.5)$ of which $\{\vartheta_1, ..., \vartheta_{p_\gamma}, \varrho_1, ..., \varrho_{n-p_\gamma}\}$ is the normalized orthogonal basis.
- By Cauchy's inequality, for any T of $p_\gamma \times (n - p_\gamma)$ matrix $\text{corr}(\vartheta y, (\varrho + \vartheta T)x) \leq \rho/\sqrt{1 + \rho^2}$, for $\forall y \in R^{p_\gamma}$, $\forall x \in R^{n-p_\gamma}$ and where ρ^2 is the largest eigenvalue of $T'T$.
- Then $p - p_\gamma$ variables not related to the risk of the failure are generated from the linear space $C = \{\varrho + \vartheta T\}$ with the appropriate choice of the maximum eigenvalue of $T'T$.

Simulation Strategy

For simulation we follow the strategy described in **Gui and Li** (2005). Total variables= p, potential predictors= p_γ, co-linearity exists between p_γ and $p - p_\gamma$.

- $A_{n\times p} \sim U(-1.5, 1.5)$ of which $\{\vartheta_1, ..., \vartheta_{p_\gamma}, \varrho_1, ..., \varrho_{n-p_\gamma}\}$ is the normalized orthogonal basis.
- By Cauchy's inequality, for any T of $p_\gamma \times (n - p_\gamma)$ matrix $\text{corr}(\vartheta y, (\varrho + \vartheta T)x) \leq \rho/\sqrt{1+\rho^2}$, for $\forall y \in R^{p_\gamma}$, $\forall x \in R^{n-p_\gamma}$ and where ρ^2 is the largest eigenvalue of $T'T$.
- Then $p - p_\gamma$ variables not related to the risk of the failure are generated from the linear space $C = \{\varrho + \vartheta T\}$ with the appropriate choice of the maximum eigenvalue of $T'T$.

Outline

Simulation Example

We consider-

- $n = 100$, $p = 1000$ and $p_\gamma = 10$, ρ=higher (>0.90)
- 20% data are censored at time $c_i \sim U(0, t_i)$
- $\alpha_0 = 0,\ \beta_0 = 0, \Sigma_0 = hI$, $v_0 = 3$, $\sigma_0^2 = 1$, and ϵ_i's $\sim N(0, 1)$.
- For prior on γ, Bernoulli prior with expected number of variable expected to 15 is considered.
- MCMC for 200000 iterations with a starting model with 40 randomly selected variables.

Simulation Example Continued

How sensitive would the results be to the specification of h_0 and h?

We have found in the results

- No sensitivity to the choice of h_0.
  - We used $h_0 = 10^4$
- Some sensitivity to the choice of h.
  - We found h within [0.05, 11.5] provides good mixing of MCMC sampler.
  - Our result is a bit different compared to [0.1, 10] proposed by **Sha et al.**(2006).
  - For best mixing of MCMC, we propose h's value 5.7, the median value but not 1 which was proposed by **Sha et al.**

Simulation Example Continued

How sensitive would the results be to the specification of h_0 and h?

We have found in the results

- No sensitivity to the choice of h_0.
    - We used $h_0 = 10^4$
- Some sensitivity to the choice of h.
    - We found h within [0.05, 11.5] provides good mixing of MCMC sampler.
    - Our result is a bit different compared to [0.1, 10] proposed by **Sha et al.**(2006).
    - For best mixing of MCMC, we propose h's value 5.7, the median value but not 1 which was proposed by **Sha et al.**

Simulation Example Continued

How sensitive would the results be to the specification of h_0 and h?

We have found in the results

- No sensitivity to the choice of h_0.
  - We used $h_0 = 10^4$
- Some sensitivity to the choice of h.
  - We found h within [0.05, 11.5] provides good mixing of MCMC sampler.
  - Our result is a bit different compared to [0.1, 10] proposed by **Sha et al.**(2006).
  - For best mixing of MCMC, we propose h's value 5.7, the median value but not 1 which was proposed by **Sha et al.**

Simulation Example Continued

How sensitive would the results be to the specification of h_0 and h?

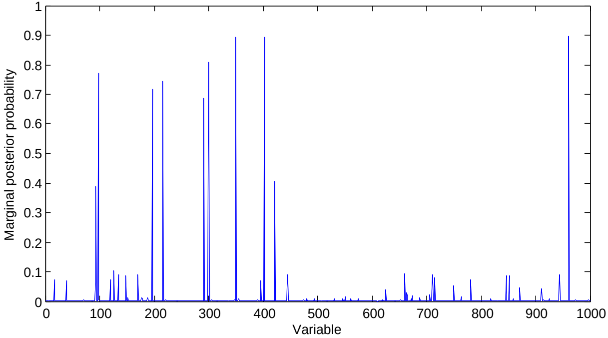
We have found in the results

- No sensitivity to the choice of h_0.
  - We used $h_0 = 10^4$
- Some sensitivity to the choice of h.
  - We found h within [0.05, 11.5] provides good mixing of MCMC sampler.
  - Our result is a bit different compared to [0.1, 10] proposed by **Sha et al.**(2006).
  - For best mixing of MCMC, we propose h's value 5.7, the median value but not 1 which was proposed by **Sha et al.**

Simulation Example Continued

Marginal Probabilities are greater than 0.5 for 9 variables, all in the set of 10 variables related to failure time.

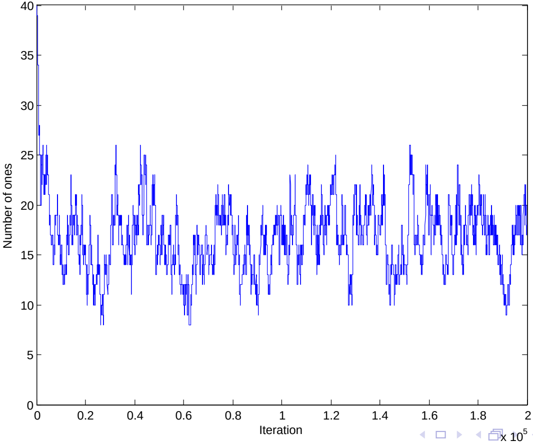
Figure 1: Marginal posterior probabilities of inclusion



Simulation Example Continued

MCMC chain mixes well and concentrating mostly on 10 to 24 variables.

Figure 2: Number of included variables



Future Plans

To perform sensitivity analysis for prior specifications with Log-normal AFT model with respect to

- Various sample sizes and covariates sizes
- Different correlation structures
- Several prior structures of the model space parameter for the underlying methods
- Several potential Bayesian methods (i.e. Bayesian LASSO; **Park and Casella**, 2008)

References

- E. George and R. McCulloch. Variable sampling via gibbs sampling. *J. Amer. Stat. Assoc.*, 88:881-889, 1993.
- J. Gui and H. Li. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21:3001-3008, 2005.
- T. Park and G. Casella. The Bayesian lasso. *J. Amer. Stat. Assoc.*, 103:681-686, 2008.
- Sha et al. Bayesian variable selection for the analysis of microarray data with censored outcome. *Bioinformatics,* 22(18):2262-2268, 2006.
- L. J. Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 11:1871-1879, 1992.