

Bayesian semiparametric inference for the AFT model, using *N-IG mixture* priors

*Inferenza Bayesiana semiparametrica per il modello AFT, con
distribuzione a priori N-IG mistura*

Raffaele Argiento, Antonio Pievatolo, Fabrizio Ruggeri ¹

CNR-IMATI, Milano, e-mail: raffaele@mi.imati.cnr.it,

marco@mi.imati.cnr.it, fabrizio@mi.imati.cnr.it

Alessandra Guglielmi ²

Dip. di Matematica, Politecnico di Milano,

e-mail: alessandra.guglielmi@mate.polimi.it

Riassunto: In questo lavoro seguiamo un’impostazione semiparametrica nell’ambito della statistica Bayesiana per un modello AFT (Accelerated Failure Time), che individua una relazione log-lineare fra i tempi di sopravvivenza e un vettore di covariate. La componente d’errore nel modello viene considerata come una mistura di densità parametriche con supporto sui reali positivi, con distribuzione misturante non parametrica, per la quale proponiamo la legge del processo N-IG (Normalized inverse-Gaussian), costruita a partire dalle leggi finito dimensionali gaussiane inverse (modello *N-IG mistura*). In particolare confrontiamo l’efficacia del modello proposto con quella del corrispondente modello costruito a partire dal processo di Dirichlet (modello DPM).

Keywords: AFT regression models, Bayesian semiparametrics, Mixture models, MCMC algorithms.

1. Extended Summary

Let $\mathbf{T} = (T_1, \dots, T_n)$ be the vector of the survival times of n subjects, and let $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})$ be a p -vector of covariates. In the survival literature, the AFT model is usually meant as the multiplicative effect of the vector $\mathbf{x}_i$ on the failure time T_i , i.e. $T_i = e^{-\mathbf{x}_i\beta} V_i$ for $i = 1, \dots, n$, where $\beta = (\beta_1, \dots, \beta_p)$ is the vector of regression coefficients, and V_i denote the errors. We consider a nonparametric hierarchical mixture (NPHM) model for the errors V_i , i.e. for $i = 1, \dots, n$:

$$V_i|\theta_i \stackrel{ind}{\sim} k(\cdot; \theta_i), \quad \theta_i|G \stackrel{iid}{\sim} G, \quad G \sim q, \quad (1)$$

where $k(\cdot|\theta_i)$ is a parametric family of densities on $\mathbb{R}^+$, and q is the prior distribution of the random distribution function G . In particular we assume that G is a Dirichlet process, or a N-IG process.

The Dirichlet process is maybe the most studied non parametric prior in Bayesian statistics. One of the reasons of this popularity is its “relative” simplicity. It has almost sure

¹Address of correspondence: Via Bassini, 15, Milano

²Address of correspondence: Piazza Leonardo da Vinci, 32, Milano

discrete realizations; this property, generally considered as a drawback, is instead a desirable feature when it is used as prior in NPHM modeling. The N-IG process, recently introduced in literature by Lijoi *et al.* (2005), represents a valid alternative to the Dirichlet process. Indeed, it preserves almost the same tractability and it has almost sure discrete realizations. Moreover it has an interesting clustering property that makes use, in a predictive sense, of all the information contained in the data.

The work mainly focuses on the performances of the two NPHM models by comparing the Bayesian inference on the regression parameters and the predictive distributions.

Our estimates are based on the predictive distribution, i.e the distribution of a new observation T_{n+1} , with covariates $\mathbf{x}_{n+1}$, given the sample $\mathbf{T}$, with density that can be written as $f(t_{n+1}|\mathbf{T}) = \int f(t_{n+1}|\beta, \underline{\theta}, \mathbf{x}_{n+1})d\pi(\beta, \underline{\theta}|\mathbf{T})$. A sample from the posterior $\pi(\cdot, \cdot|\mathbf{T})$ is obtained through MCMC simulations. In particular for the full conditional $\pi(\cdot|\beta, \mathbf{T})$, concerning the posterior sample of $\underline{\theta} = (\theta_1, \dots, \theta_n)$, we use a Pólya urn Gibbs sampler procedure, for both the N-IG and the Dirichlet prior model (see, Ishwaran and James, 2001, Neal, 2000). These sample schemes are based on the Pólya urn representation of the predictive distribution of θ_{n+1} given $\underline{\theta}$, written as $\mathbb{P}(\theta_{n+1} \in \cdot | \underline{\theta}) = w_{0,q}^{(n)} G_0(\cdot) + w_{1,q}^{(n)} \delta_{\theta^*}(\cdot)$, where $w_{0,q}^{(n)}$ and $w_{1,q}^{(n)}$ are suitable weights, $G_0(\cdot) := \mathbb{E}_q(G(\cdot))$, and $\theta_1^*, \dots, \theta_k^*$ denote the k distinct observations within $\underline{\theta}$. In our models the random variable k has a key role, indeed it represents the number of components of the mixture model under study. A priori, the N-IG leads to a less informative prior for k than the Dirichlet prior.

We test the two competing models on both simulated and real data. The Pólya urn Gibbs sampler scheme is particularly simplified when the “mean” distribution $G_0(\cdot)$ and the kernel $k(\cdot; \cdot)$ are conjugate in θ . We implement both the conjugate model (Argiento *et al.*, 2006), and the non conjugate model (Argiento, 2007).

Once the predictive estimates have been obtained, we measure the predictive power of the two proposed models. In the simulated data case, we consider the distance in the uniform metric between the predictive and the “target” distribution. In the real data set case we use a *cross validation* method, quantifying in practice, how far the predictions are from the data. From a predictive point of view, our experiments show that there is not a substantial difference between the DPM and the N-IG mixture model. However, the N-IG process prior seems to be more effective in the estimation of the number of components in the mixture.

References

- Argiento R. (2007) *Bayesian semiparametric inference for accelerated failure time models*, Ph.D. thesis, Bocconi University.
- Argiento R., Guglielmi A., Pievatolo A. and Ruggeri F. (2006) Bayesian semiparametric inference for the accelerated failure time (aft) model using hierarchical mixture modeling with n-ig priors, in: *Joint Statistical Meetings, Seattle.*
- Ishwaran H. and James L.F. (2001) Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association*, 96, 161–173.
- Lijoi A., Mena R.H. and Prünster I. (2005) Hierarchical mixture modeling with normalized inverse-Gaussian prior, *Journal of the American Statistical Association*, 100, 1278–1291.
- Neal R.M. (2000) Markov Chain sampling methods for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics*, 9, 249–265.