University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2014 *Paper* 112

Clustering Survival Outcomes using Dirichlet Process Mixture

Lili Zhao* Jingchunzi Shi†

Tempie H. Shearon‡ Yi Li**

*University of Michigan School of Public Health

†University of Michigan School of Public Health

‡University of Michigan School of Public Health

**University of Michigan School of Public Health

Clustering Survival Outcomes using Dirichlet Process Mixture

Lili Zhao, Jingchunzi Shi, Tempie H. Shearon, and Yi Li

Abstract

Motivated by the national evaluation of mortality rates at kidney transplant centers in the United States, we sought to assess transplant center long- term survival outcomes by applying a methodology developed in Bayesian non-parametrics literature. We described a Dirichlet process model and a Dirichlet process mixture model with a Half-Cauchy for the estimation of the risk- adjusted effects of the transplant centers. To improve the model performance and interpretability, we centered the Dirichlet process. We also proposed strategies to increase model's classification ability. Finally we derived statistical measures and created graphical tools to rate transplant centers and identify outlying centers with exceptionally good or poor performance. The proposed method was evaluated through simulation, and then applied to assess kidney transplant centers from a national organ failure registry.

Clustering Survival Outcomes using Dirichlet Process Mixture

Lili Zhao [*], Jingchunzi Shi [*], Tempie H. Shearon [*], and Yi Li [*]

[*]Department of Biostatistics, University of Michigan, Ann Arbor

March 5, 2014

Abstract

Motivated by the national evaluation of mortality rates at kidney transplant centers in the United States, we sought to assess transplant center long-term survival outcomes by applying a methodology developed in Bayesian nonparametrics literature. We described a Dirichlet process model and a Dirichlet process mixture model with a Half-Cauchy for the estimation of the risk-adjusted effects of the transplant centers. To improve the model performance and interpretability, we centered the Dirichlet process. We also proposed strategies to increase model's classification ability. Finally we derived statistical measures and created graphical tools to rate transplant centers and identify outlying centers with exceptionally good or poor performance. The proposed

method was evaluated through simulation, and then applied to assess kidney transplant centers from a national organ failure registry.

Key Words: Clustering survival data; DP; DPM; Mixture model;

1 INTRODUCTION

An issue of substantial importance is the monitoring and improvement of health care centers such as hospitals, nursing homes, dialysis facilities or surgical wards. This study is motivated by the need for the evaluation of kidney transplant centers in the United States with respect to their mortality rates after transplantation. The data include patients in the Scientific Registry of Transplant Recipients (SRTR) who received their kidney from 2008 to 2011. A total of 56455 kidney transplants were performed at 242 transplant centers.

There is a large amount of literature describing methods for the evaluation of center performances and identification of outlying centers with extremely good or poor performance. The application includes a variety of data: (standardized) mortality, proportions, counts of adverse events, categorical data or continuous data measuring quality of life. Some examples of using parametric approaches can be found in Liu et al. (2003); Jones and Spiegelhalter (2011); Kalbfleisch and Wolfe (2013); He et al. (2013). Among these articles, Liu et al. (2003); Jones and Spiegelhalter (2011) used a normal hierarchical (random effects) model for the center effects. As we know, random effects models improve estimation by borrowing information across transplant centers, and thus shrinking estimates of the center effects toward the overall mean and leading to a reduced variation of the estimates. However, the smaller variance is achieved at the cost of bias and inappropriate shrinkage could prevent the centers with exceptionally good or poor performance from being identified. For this reason, Kalbfleisch and Wolfe (2013); He et al. (2013) prefer a model with center effects being considered as fixed, leading to independent (no shrinkage) center estimates. It seems

that a desirable model would combine the advantages of both the fixed and random effects models, in the sense that it would allow borrowing strength across similar centers, but avoid shrinking outlying centers towards the population mean.

Moreover, in both random or fixed effects models, it is not immediately clear how unusual centers, i.e., any with exceptionally good or poor performance, can be identified. An common strategy is to measure the deviation of each transplant center relative to the population average using a p-value or an (adjusted) Z-score derived from an assumed parametric or empirical null distribution (see D et al. (2012); Kalbfleisch and Wolfe (2013)). However, ideally a model would provide an in-built diagnostic measure for centers with unusual outcomes.

Ohlssen et al. (2007) applied a Dirichlet process (DP) model and a Dirichlet process mixture (DPM) model to the problem of hospital comparisons using mortality rates. These Bayesian non-parametric approaches satisfies the above requirements. The efficiently accommodates outlying centers and allows for a more flexible distribution of center effects with the possibility of skewness and multimodality. Furthermore, the embedded clustering feature in Dirichlet process models provides inherent diagnostic measures to identify outlying centers. However, Ohlssen et al. (2007) considered mortality rates as binary outcomes. In our application, majority of data are censored; as of Jan 31, 2013, 93% patients were still alive. Therefore, we extended their work to survival-time data, defined from the time of kidney transplantation to death; patients who are alive at the last follow-up time point were considered to be right centered. Center effects were represented as random effects (frailties) in a Cox proportional hazard model. In the Cox model, we included important patient-level characteristics and dealt with missing data. Due to the large number of transplants (> 50000) and the large dimension of patient-level covariates, Kalbfleisch and Wolfe (2013) used a two-stage approach to obtain the risk-adjusted center effects. In the firs stage, they estimated patient-level covariates from a Cox model stratified by transplant centers; in the second stage, they derived center effects by fixing the covariate effects obtained from the first stage. However, we used a fully Bayesian approach, in which we used a Gibbs algorithm that alternates between (1) updating effects of

covariates with a Metroplis-Hasting algorithm conditional on estimated center effects, and then (2) updating center effects conditional on estimated covariate effects using a DP or DPM model.

In this article, we proposed strategies to improve model performance and increase model's classification ability. To our knowledge, this approach has not yet been considered for the national evaluation of survival outcomes.

The remaining of the article is organized as follows. Section 2 describes a DP prior and a DPM prior for modelling center effects in a Cox proportional hazard model, proposes strategies to improve model performance and creates graphical tools to evaluate centers. Section 3 presents simulation studies and investigates shrinkage effects for data with different clustering structures. Section 4 illustrates the analysis on the Kidney transplant data. Section 5 is the concluding discussion.

2 MODEL

2.1 Cox Proportional Hazard Model

The data are denoted by $\{(t_{ij}, \delta_{ij}, x_{ij}), i = 1, \cdots, N; j = 1, \cdots, n_i\}$, where t_{ij} is the observed event time for patient j in transplant center i; $\delta_{ij} = 1$ if t_{ij} is an observed failure time and 0 if it the failure time is right censored at t_{ij} , and x_{ij} is a p-dimensional vector of covariates.

Under the proportional hazards model, we have

$$\lambda(t_{ij}) = \lambda_0(t_{ij}) \exp\{\alpha x_{ij} + \beta_i\},$$

where β_i is effect for center i after adjusting for the covariate vector x_{ij}. Often the baseline hazard is assumed to be piecewise constant on a partition comprised of K

disjoint intervals, yielding the piecewise exponential model (See, for instance, Jara (2011), Walker and Mallick (1997), Aslanidou, Dey, and Sinha (1998), and Qiou et al. (1999)). Assume that $\lambda_0(t) = \sum_{k=1}^{K} \lambda_k \mathrm{I}(a_{k-1} < t \leq a_k)$, where $a_0 = 0$ and $a_K = \max\{t_{ij}\}$. Let $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_K)$, the likelihood for $(\alpha, \boldsymbol{\lambda}, \beta)$ is given by

$$L(\alpha, \boldsymbol{\lambda}, \beta) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} f(\boldsymbol{N}_{ij}, x_{ij}, \boldsymbol{\Delta}_{ij}; \alpha, \boldsymbol{\lambda}, \beta)$$

and

$$\begin{aligned} f(\boldsymbol{N}_{ij}, x_{ij}, \boldsymbol{\Delta}_{ij}; \alpha, \boldsymbol{\lambda}, \beta) &= \prod_{k=1}^{K} \exp\{-\exp\{\log \lambda_k + \alpha x_{ij} + \beta_i\}\Delta_{ijk}\} \\ &\qquad \{\exp\{\log\{\lambda_k\} + \alpha x_{ij} + \beta_i\}\Delta_{ijk}\}^{N_{ijk}}, \end{aligned} \tag{1}$$

where $\boldsymbol{N}_{ij} = (N_{ij1}, \cdots, N_{ijK})$, and N_{ijk} takes a value of one if $t_{ij} \in (a_{k-1}, a_k]$ and $\delta_{ij} = 1$ and N_{ijk} is zero otherwise. Define $\boldsymbol{\Delta}_{ij} = (\Delta_{ij1}, \cdots, \Delta_{ijK})$, and $\Delta_{ijk} = (\min\{a_k, t_{ij}\} - a_{k-1})_+$ with x_+ as $\max(x, 0)$.

The gamma process is used as a prior for the cumulative baseline hazard function Λ_0 (Kalbfleisch, 1978), i.e., $\Lambda_0 \sim \mathcal{GP}(c_0\Lambda_0^*, c_0)$, where Λ_0^* is often assumed to be a known parametric function. For example, $\Lambda_0^* = \eta y^{k_0}$ corresponds to the Weibull distribution, and c_0 represents the degree of confidence in this prior guess.

2.2 Truncated Stick-breaking Process

In this article the attention is focused on modelling random effects (frailties) of $\beta_1, \cdots, \beta_N$. Often the random effects $\beta_i' s$ (or e^{β_i}) are assumed to be generated from some parametric distribution such as normal, gamma, positive stable, etc. Here we

consider a Dirichlet process prior as

$$\begin{aligned} \beta_1, \cdots, \beta_N &\sim G \\ G &\sim \mathrm{DP}(a, G_0), \end{aligned}$$

where G_0 corresponds to a best guess for G as a priori and a expresses confidence in this guess.

The stick-breaking representation (Sethuraman, 1994) implies that $G \sim \mathrm{DP}(a, G_0)$ is equivalent to

$$G = \sum_{h=1}^{\infty} \pi_h \delta_{\beta_h}, \quad \beta_h \sim G_0, \text{ and } \sum_{h=1}^{\infty} \pi_h = 1, \tag{2}$$

where $\pi_h = V_h \prod_{l<h}(1 - V_l)$ is a probability weight formulated from a stick breaking process, with $V_h \sim \mathrm{Beta}(1, a)$, for $h = 1, \cdots, \infty$, and δ_β is a point mass at β. Small values of a (such as $a = 1$) favor a small number of clusters and large values of a (such as $a = 10$) would result in as many as 50 clusters. Since our interest lies in the identification of a few subgroups, we fixed $a = 1$, a widely used choice in applications that favors a few clusters Dunson (2010).

One potential issue with this representation of a mixture of point mass is that it assumes a discrete distribution for the random effects so that different centers in a cluster have exactly the same random effect values. It may be more realistic to assume that centers in a cluster have similar, but not identical, random effect values. To accomplish this, let $\beta_i \sim N(\mu_h, \sigma_h^2)$ and $(\mu_h, \sigma_h^2) \sim G$. That is, (2) becomes

$$G = \sum_{h=1}^{\infty} \pi_h \mathrm{N}(\mu_h, \sigma_h^2), \quad (\mu_h, \sigma_h^2) \sim G_0, \text{ and } \sum_{h=1}^{\infty} \pi_h = 1 \tag{3}$$

In this case, the random distribution, G, is characterized as a DP mixture (DPM) of normals (Escobar and West, 1995). A mixture of normals allow a flexible continuous

random-effects distribution of the center effects. (Readers can refer to a nice book by Dunson (2010) for a detailed review of the DP and DPM model).

Recent research has focussed on using the constructive definition of the DP to produce practical MCMC algorithms (Ishwaran and James, 2001). The principle is to approximate the full process by truncating the DP(M) at a maximum number of components H, so that

$$G = \sum_{h=1}^{H} \pi_h \delta_{\beta_h} \quad \text{in DP} \quad \text{and} \quad G = \sum_{h=1}^{H} \pi_h \mathrm{N}(\mu_h, \sigma_h^2) \quad \text{in DPM}$$

Closely related to a, a large H provides an accurate approximation to the full DP(M) but requires a large computation effort. Ohlssen et al. (2007) provided strategies to specify H. Since we are interested in detecting a few subgroups, i.e., two groups below and two above the population average, we fixed $H = 5$ and found that $H = 5$ worked very well for both the cases studied in the simulation and the kidney transplant data.

2.3 Hyperpriors

In PD model G_0 is often chosen to have a normal distribution, i.e., $G_0 \sim N(\mu_0, \sigma_0^2)$. The hyperparemters (μ_0, σ_0^2) can be fixed, or assigned a normal-inverse gamma hyperprior. A hyperprior would allow the base distribution having unknown mean and variance and provide a shrinkage of center effects towards the overall mean. In DPM model we assumed that $\mu_h \sim N(\mu_0, \sigma_0^2)$, with a normal hyperprior for μ_0 and a Half-Cauchy prior for σ_0^2 (Gelman, 2006), i.e., $f(\sigma_0) \propto \left(1 + (\frac{\sigma_0}{A})^2\right)^{-1}$, with a smaller A indicating a stronger prior information and a greater shrinkage. This Cauchy prior behaviors well for a small number of components (clusters), such as $H = 5$, and it restricts σ_0^2 away from very large values and have better behavior near zero, compared to the inverse-gamma family (Gelman, 2006). We also assumed that $1/\sigma_h^2 \sim \text{Gamma}(e_0, f_0)$ and fixed hyperparemters (e_0, f_0) to be weakly informative, since fully non-informative priors are not possible in a mixture context (Richardson

and Green, 1997).

2.4 Centered Stick-breaking Process

In parametric hierarchical models, it is a standard practice to place a mean constraint on the latent variable distribution for sake of identifiability and interpretability (Yang et al., 2010; Yisheng Li and Lin, 2011). In this article we centered the Dirichlet process to have zero mean. Following Yang et al. (2010), we estimated the mean of the process, μ_G^m, at the m^{th} MCMC iteration as

$$\mu_G^m = \sum_{h=1}^{H} V_h^m \prod_{l<h} (1 - V_l^m)\beta_h^m,$$

where V_h^m and β_h^m are the posterior samples from the un-centered process defined in (2), and $\beta_i^m - \mu_G^m$ $(h = 1, \cdots, H)$ is the "centered" estimate for center i at the m^{th} iteration. The same idea applies to the DPM model.

Centering the process improves model performance in two aspects. First, it improves MCMC convergence and mixing rates. Second, an "centered" estimate can be interpreted as a deviation from the population average.

2.5 Posterior Computation

The blocked Sampler of Ishwaran and James (2001) was used to allocate each center to one of the components by sampling the label Z_i from a multinomial conditional posterior. In the DP model, probabilities in the multinomial distribution are :

$$Pr(Z_i = h|-) = \frac{\{V_h \prod_{l<h}(1 - V_l)\} \prod_{j=1}^{n_i} f(\boldsymbol{N}_{ij}, \boldsymbol{\Delta}_{ij}, x_{ij}; \alpha, \boldsymbol{\lambda}, \beta_h)}{\sum_{r=1}^{H} \{V_r \prod_{l<r}(1 - V_l)\} \prod_{j=1}^{n_i} f(\boldsymbol{N}_{ij}, \boldsymbol{\Delta}_{ij}, x_{ij}; \alpha, \lambda, \beta_r)},$$

where $f(\boldsymbol{N}_{ij}, \boldsymbol{\Delta}_{ij}, x_{ij}; \alpha, \boldsymbol{\lambda}, \beta)$ is defined in (1).

In the DPM, the probabilities are

$$Pr(Z_i = h|-) = \frac{\{V_h \prod_{l<h}(1-V_l)\} \prod_{j=1}^{n_i} N^{\eta}(\beta_i; \mu_h, \sigma_h^2)}{\sum_{r=1}^{H} \{V_r \prod_{l<r}(1-V_l)\} \prod_{j=1}^{n_i} N^{\eta}(\beta_i; \mu_r, \sigma_r^2)}$$

We used $\eta = 2$ to facilitate efficient classification of the label indicators, and thus avoiding all centers to be classified into a single component (cluster). We have found that a small η ($\eta > 1$) improved the clustering performance especially when the prior for component parameters are weak. Intuitively, $\eta = 2$ implies that each patient information is doubled, and thus increases the effect size. It is worth mentioning that different choices of η generally will not affect the posterior estimation of the center effects, and estimates with $\eta = 2$ is very similar to that with $\eta = 1$. The posterior calculations for other parameters are shown in the Appendix.

2.6 Statistical Measures to Rate Centers

A useful metric to rate transplant centers is their ranks. In each MCMC iteration, β_i ($i = 1, \cdots, N$) was ranked; without ties, the smallest β_i had rank 1 and the largest β_i has rank N. Over all MCMC interactions, we obtained a distribution of ranks for each center. Ohlssen et al. (2007) proposed a measure to assess pairwise clustering between centers by creating a $N \times N$ matrix of posterior probabilities of the two centers being classified into the same cluster. We combined the above two measures and graphically represented the $N \times N$ matrix using a heat map where transplant centers are ordered by their posterior means of the ranks. This heat map reveals a clustering structure of the studied centers that facilitates rating centers, as well as identifying outlying centers.

Additionally, in order to visually detect outlying centers, we calculated the proportion of centers in the same cluster as center i, denoted by PS. Together with the

rank (percentile) statistics, we created a graph that helps identify centers that are in isolated small clusters with exceptionally low or high ranks.

3 SIMULATION STUDIES

As discussed in Dunson (2010), the clustering is sensitive to hyperparameters, and different hyperarameters induce different model shrinkage. However, it is not immediately clear how hyperarameters affect shrinakge in survival data and what is a desirable shrinkage for a model. In this section we conducted simulation studies to investigate the performance of DP and DPM models that were designed to have different degree of shrinkage.

- A DP model with fixed hyperparameters, i.e., (μ_0, σ_0^2) are fixed (denoted by DP), or a random normal-inverse gamma hyperprior for (μ_0, σ_0^2) (denoted by DP-HP).
- A DPM model with $\mu_h \sim (\mu_0, \sigma_0^2)$, where μ_0 has a normal hyperprior and σ_0^2 has a Half-Cauchy prior with $A = 1$ or 5. We also assigned a very weak prior for σ_h^2, i.e., $1/\sigma_h^2 \sim \text{Gamma}(0.01, 0.01)$.

The DP model, with fixed hyperparameters, do not induce shrinkage between clusters, but shrinks centers within the same cluster to a single estimate. In contrast, DPM allows shrinkage between- and within-clusters with a smaller A indicating a stronger shrinkage. Intuitively, DP-HP could have a stronger shrinkage than DPM since DP-HP has the strongest shrinkage within cluster as well as a between-cluster shrinkage that is induced by a hyperprior.

In all simulations, survival times were generated from a Cox model (Bender et al., 2005), $S(t|center_i) = \exp[-\Lambda_0(t)\exp(\beta_i),]$ where Λ_0 is the cumulative hazard function of a Weibull distribution, with a scale parameter of one and a shape parameter of 0.8, suggesting the mortality rate decreases over time, which is observed in the Kidney

Table 1: Simulation set-up for Scenario I-III

	C1	C2	C3
Scenario I	$n_1 = 16$ $N(0.69, 0.2^2)$	$n_2 = 16$ $N(0, 0.2^2)$	$n_3 = 16$ $N(0.69, 0.2^2)$
Scenario II	$n_1 = 8$ $N(0.69, 0.2^2)$	$n_2 = 24$ $N(0, 0.2^2)$	$n_3 = 16$ $N(0.69, 0.2^2)$
Scenario III	$n_1 = 4$ $N(0.69, 0.1^2)$	$n_2 = 40$ $N(0, 0.2^2)$	$n_3 = 4$ $N(0.69, 0.1^2)$

transplant data. For an illustrative purpose, we did not include covariates in the simulation.

We generated data under four data clustering structures, each consists of 48 centers and each center has 20 or 40 patients. Scenario I-III have three clusters but with different sizes (the size refers to number of centers in a cluster). Table describes the size of the cluster and the distribution of β_i's within each cluster. For instance, the first n_1 centers form a cluster named as C1, with $\beta_i's$ simulated from a normal distribution with a mean -0.69 and a standard deviation of 0.2; likewise, the next n_2 centers form a cluster named as C2 and the last n_3 centers form a cluster named C3. We also include Scenario IV in the simulation, in which all β_i's were generated a single cluster with $\beta_i \sim N(0, 0.3)$. Within each cluster, the first half centers have $n = 20$ and other half centers have $n = 40$. A large β_i means poor center performance. β_i of -0.69 and 0.69 correspond to a hazard ratio of 0.5 and 2 relative to the population average, respectively. These clinically meaningful ratios are expected to be detected in the real data analysis.

All normal priors were assumed to have a mean of zero and variance of 100. The baseline hazard was assumed to have an exponential distribution with a rate of 1 and $c_0 = 0.1$, and the time axis was partitioned into 5 intervals based on the observed quantiles. All the priors were set to be quite weak. With a burn-in of 1000 iterations,

Table 2: Parameter estimation and performance statistics with respect to the absolute bias (Bias), standard deviation (SD) and mean square error (MSE), based on 200 simulated datasets.

			DP			DPM ($A = 1$)		
		n	C1	C2	C3	C1	C2	C3
I	Bias	20	0.15	0.03	0.12	0.11	0.02	0.08
		40	0.08	0.02	0.06	0.06	0.02	0.02
	SD	20	0.24	0.28	0.25	0.30	0.27	0.27
		40	0.23	0.21	0.22	0.26	0.18	0.24
	MSE	20	0.08	0.08	0.08	0.10	0.07	0.09
		40	0.06	0.04	0.05	0.07	0.03	0.06
II	Bias	20	0.11	0.09	0.25	0.05	0.09	0.21
		40	0.02	0.11	0.17	0.01	0.10	0.15
	SD	20	0.31	0.21	0.26	0.33	0.22	0.28
		40	0.29	0.15	0.25	0.28	0.17	0.25
	MSE	20	0.11	0.05	0.13	0.11	0.06	0.12
		40	0.09	0.04	0.09	0.08	0.04	0.08
III	Bias	20	0.35	0.00	0.30	0.28	0.01	0.25
		40	0.21	0.01	0.15	0.16	0.01	0.15
	SD	20	0.30	0.12	0.27	0.34	0.14	0.26
		40	0.31	0.09	0.26	0.30	0.12	0.25
	MSE	20	0.21	0.01	0.16	0.20	0.02	0.13
		40	0.14	0.01	0.09	0.12	0.01	0.09

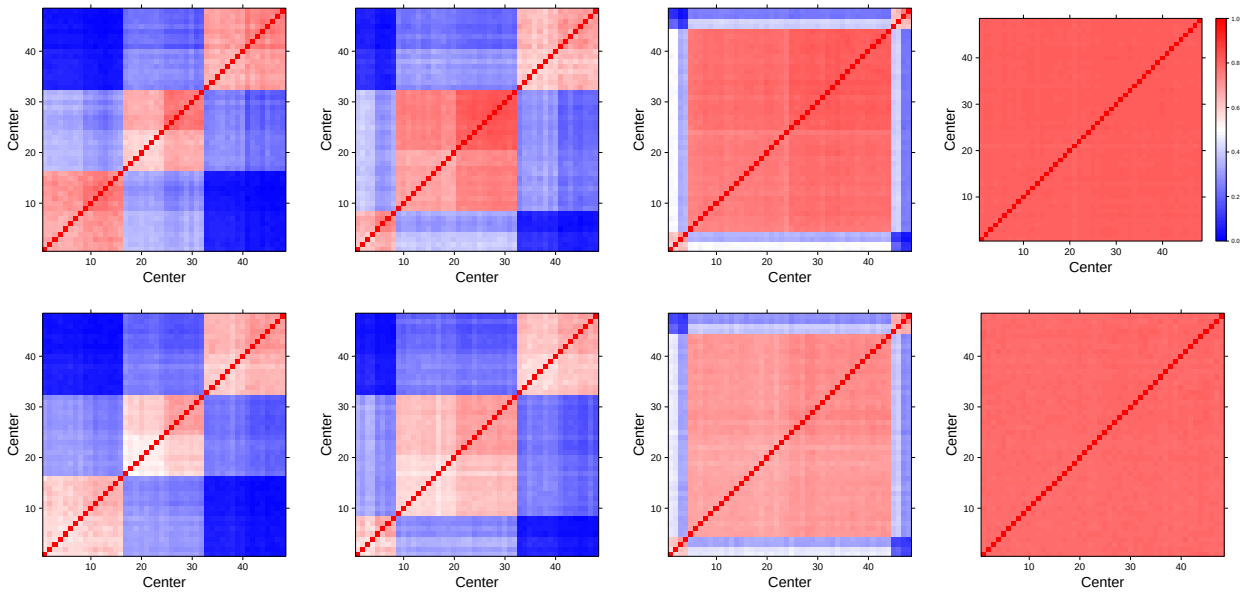


Figure 1: Pairwise posterior probabilities of two centers assigned to the same cluster under four scenarios using the DP model (the first row) and the DPM($A = 1$) model (the second row). White, red and blue color corresponds to a probability of equal to, larger and less than 0.5, respectively; The darker the red, the closer the probability is to 1; the darker the blue, the closer the probability is to 0. The heat map of DP-HP DPM ($A = 5$) are very similar to DP and DPM ($A = 1$), respectively, so their heat maps are not presented.

Table 3: DIC_3 under four scenarios.

	DP	DP-HP	DPM ($A = 1$)	DPM ($A = 5$)
I	6174	6177	6164	**6149**
II	6435	6437	**6413**	6422
III	**6320**	6327	6330	6329
IV	6390	**6371**	6387	6386

an additional 2000 iterations were used for inference.

Table 2 shows the parameter estimations and performance statistics with respect to the absolute bias (Bias), standard deviation (SD) and mean square error (MSE), based on 200 repeated datasets. As expected, centers with $n = 40$ have more accurate estimates compared to centers with $n = 20$, as evidenced by a smaller bias, SD and MSE. Surprisingly, parameter estimations are very similar between DP and DPM models, so we only presented the DP and DPM ($A = 1$) for illustration. It is also interesting to note that, in both DP and DPM models, estimates in a small cluster can be significantly biased toward a large cluster.

Figure 1 displays the estimated clustering structure for the DP and DPM model under four scenarios over 200 repeated datasets. Each heat map was created based on 48×48 matrix, containing pair-wise posterior probabilities for two centers being classified into the same cluster. Since we know the true cluster status for each center, centers are ordered by their true IDs. A red square represents a cluster (subgroup). It is apparent that the true clustering structure is well represented in all scenarios in both DP and DPM models. As expected, centers with a large sample size ($n = 40$) are more likely to be classified correctly than centers with a small sample size ($n = 20$), as demonstrated by large centers having in a cluster and a darker blue between clusters. Upon closer inspection, DP models generally have higher pair-wise probabilities than DPM models, as evidenced by a darker red and a lighter blue in the DP model. We observed that the pair-wise probabilities are less influenced by undue shrinkage than the estimates of the center effects seen in Table 2. Here, the average pair-wise probabilities are between 65%-70% for all three clusters. For example, in Scenario III, probabilities for C1, C2 and C3 are 0.65, 0.70 and 0.70, respectively in the DPM model and 0.65, 0.78 and 0.69, respectively in the DP model.

We also computed two Bayesian model comparison criteria for selecting the best model. one is DIC_3 Celeux et al. (2006) and the other is the log-pseudo marginal likelihood (LPML) Ibrahim et al. (2001). The best model should have the smallest DIC_3 and largest LPML. We only presented DIC_3 since LPML showed exactly the

same ordering as the DIC_3. As indicated by the performance statistics, when data consist of a few clusters, a DP model is preferred if a big cluster is accompanied by a few small outlying clusters (such as Scenario III) and a DPM is a better choice otherwise. In the last scenario when there is only a single cluster, DP-HP is the best.

4 APPLICATION

We applied our model to data on 213 transplant centers in the US, excluding centers with fewer than 10 patients. The number of patients per center has a median of 198 and an interquartile range of (111, 356). We selected nine patient-level covariates using a forward selection algorithm, including Cold ischemia time, Peak renal reactive antibody level (PRRA), Body mass index, Time on renal replacement therapy (TRRT), Donor race, Recipient race, Donor history of Diabetes, Previous Solid Organ Transplant, Recipient Diagnosis. Due to the retrospective nature of the analysis, values were missing for some of those characteristics. For instance, there are 16.47% missing data in TRRT and 2.14% missing in PRRA. In order to include patients with partially missing covariates while reserving the original covariate distributions, we created a binary variable for each covariate indicating if the data is missing for each subject. For example, a continuous covariate was created into two variables with one variable containing the original value and the other variable containing one if the data is missing and zero otherwise. By doing so, we created 12 covariates.

The priors used in the application are the same as in the simulation studies except that $1/\sigma_h \sim \text{Gamma}(3, 0.5)$, which is also weak relative to the likelihood. With a burn-in of 10000 iterations, an additional 20000 iterations were used for posterior inference. We observed that the chain mixes well and the results are robust to different choices of the initial values.

As illustrated in Table 4, A DPM model with a stronger shrinkage (e.g., $A = 1$) is preferred for the kidney transplant data. In this application, we also tried a larger

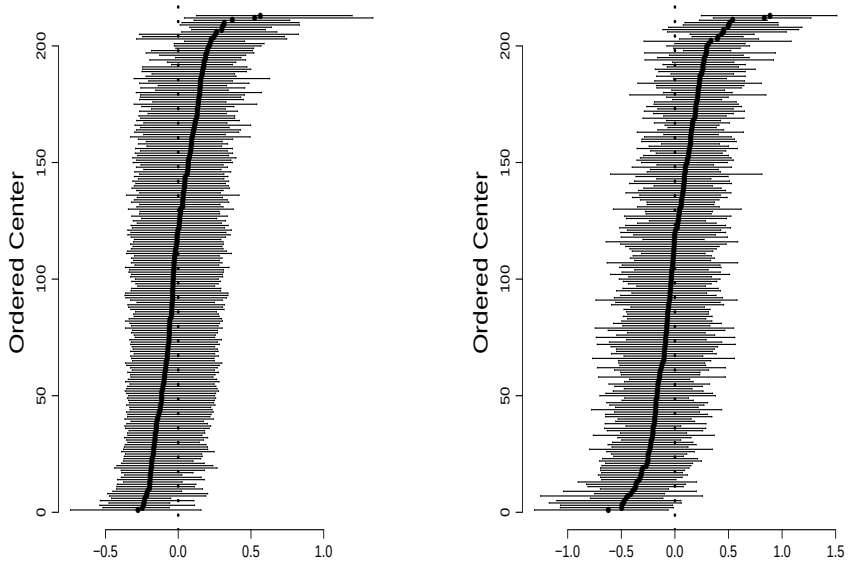
Table 4: Diagnosis statistics for different models

Model	DIC_3 (pD_3)	LPML
DP	41793 (113)	-20897
DPM($A = 1$)	41763 (137)	-20883
DPM($A = 5$)	41765 (138)	-20884

H ($H = 20$ and $H = 50$) and considered a random a, respectively. We found that a large H did not improve the model performance (DIC_3 and LPML were the same as fixing $H = 5$), and data seemed to contain little information about estimating the parameter a, leading to the same, or slightly worse DIC_3 and LPML, compared to a fixed a (data not shown).

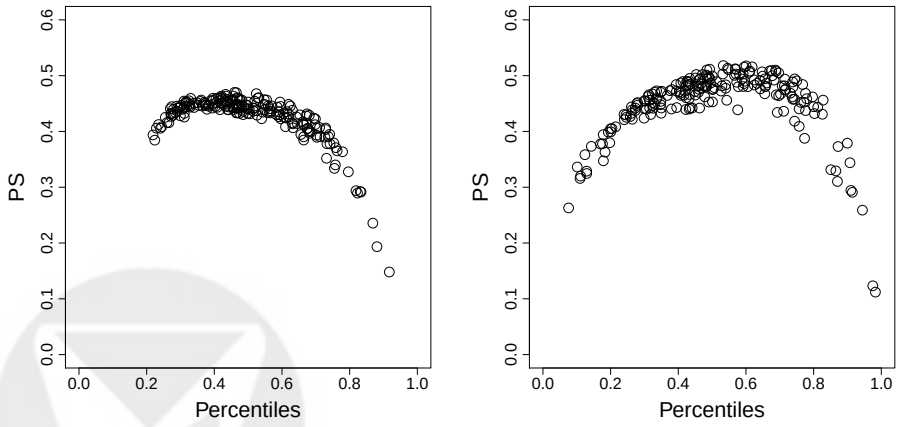
Figure 2 presents caterpillar plots of the center estimates and Figure 3 depicts outlying centers at two tails, i.e., centers with very low and high percentiles and small probabilities of being in the same cluster as other centers. In both DP and DPM models, two transplant centers (with id 652 and 437) have the worst outcomes ($PS < 0.2$ and Percentile> 0.8). It is also interesting to note that a few centers with exceptionally good performance are observed in DPM model but not in DP model.

The heat map in Figure 4 is based on pair-wise probabilities and ordered by rank statistics as described in Section 2.6. It appears that a few centers at the upper right corner form an isolated cluster (this is clearer when the plot was magnified to 200%), which performed significantly worse than the population average. Moving along the diagonal toward the lower left corner, the next cluster consists of approximately 60 centers that performed better than the previously mentioned small outlying cluster but still worse than the population average. No isolated subgroups of centers seem to perform significantly better than the population average; however, a big group of around 60 centers performed above average. We found that the DP and DPM model identified very similar outlying centers with slightly different ordering of across the centers.



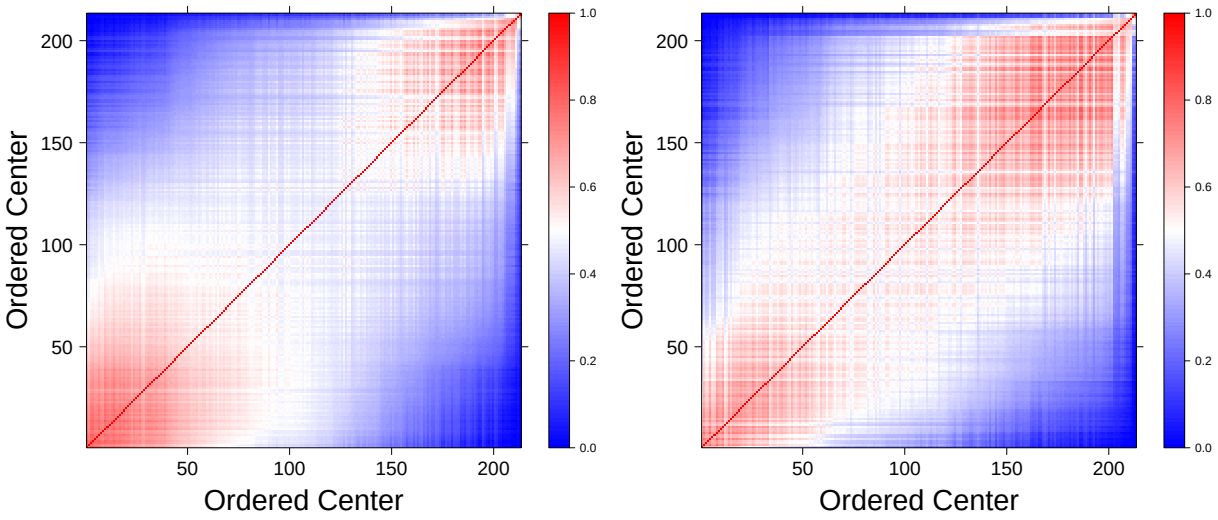
(a) (b)

Figure 2: Caterpillar plots of 95% credible intervals for estimates of the center effects in kidney transplant data using the DP model in (a) and the DPM ($A = 1$) model in (b). The transplant centers were placed in the order of their posterior means. The dotted vertical line corresponds to the population average.



(a) (b)

Figure 3: The x axis is the mean percentile and the y axis is the mean percentage of centers being in the same cluster as center i for the kidney transplant data using DP model (a) and DPM ($A = 1$) model (b).



(a) (b)

Figure 4: Heat maps representing pair-wise posterior probabilities of the two centers are classified into the same cluster; centers are ordered based on their mean ranking scores. (a) and (b) are from DP and DPMG model, respectively.

5 DISCUSSION

In this article we evaluated long-term mortality rates of kidney transplant centers from a national organ failure registry. We proposed a fully Bayesian approach to model patient-level covariates and center effects in a Cox Proportional Hazard model. In modelling the risk-adjusted center effects, we described a Dirichlet process model and a Dirichlet process mixture model with a Half-Cauchy. We also derived statistical measures and presented graphical tools to rate transplant centers and identify outlying center with exceptionally good or poor performance.

To improve the model performance and interpretability for the survival data, we centered the Dirichlet process. Without centering, MCMC chains exhibit very high autocorrelation which has no hope of yielding any meaningful estimates. Centering the DP process by constraining the mean to zero dramatically improved model convergence and interpretability of the estimates of the center effects. To increase model's classification ability, we raised the normal density to a power of two for updating component labels in DPM. In practice this η can be considered as a tuning parameter, say try $\eta = 1.5, 2 or 2.5$, and choose the η with the best model performance assessed by the heat map and LPML statistics. We found that this strategy dramatically improved model's classification ability and it outperformed alternative strategies, such as forcing a common variance across components ($\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_H^2$) and assigning strongly informative priors for $\sigma_h^2 (h = 1, \cdots, H)$.

The graphical tools and statistical measures proposed in this article are particularly useful to reveal a clustering structure of all transplant centers and help identify outlying centers. We can set up hypothesis testing to further quantify the outlying centers. This is not our focus in this article, but it would be an interesting future work.

As discussed in Ohlssen et al. (2007), the modelling should be seen as a starting point for further investigation into the reasons for unusual results or apparent clustering. This might include checking the data for coding errors, investigation of

covariates of centers which appear to form clusters and more detailed inspection of transplant centers with unusual outcomes.

APPENDIX

A.1 Gibbs Sampler

In DP model,

1. Update the stick-breaking weights from conditionally conjugate beta posterior distributions:

$$V_h|- \sim \text{Beta}\left(1 + \sum_{i=1}^{N} \text{I}(Z_i = h), a + \sum_{i=1}^{N} \text{I}(Z_i > h)\right), h = 1, \cdots, H$$

2. Given the centers with labels specific to cluster h,, update β_h by the adaptive rejection algorithm and $\beta_h \sim N(0, 100)$ as a priori (see Gilks et al. (1995); Ibrahim et al. (2001)).

3. Update baseline hazard in interval $k(k = 1, \cdots, 5)$ from Gamma$(1 \times 0.1 + D_k, 0.1 + \sum_{i \in R_k} exp\{\alpha x_{ij} + \beta_i\}\Delta_{ijk})$, and D_k and R_k represents the number of death and the number subjects at risk in interval k.

4. The vector of covariates was divided into 3 groups with 6 covariates per group, and α was updated by groups. Within each group, the corresponding α was updated using the adaptive Metropolis-Hastings algorithm (Haario et al., 2001). The initial estimates of α was calculated from a Cox model stratified by centers. The multivariate normal proposal density centered at the previous value, and the covariance in the proposal was "refined" by using the empirical covariance from an extended burn-in period.

5. Update a from a Gamma distribution

$$a \sim \text{Gamma}\left(1, a_0 + H - 1, b_0 - \sum_{r=1}^{H-1} \log(1 - V_r)\right) \text{I}(0.3, 10)$$

The prior for a is gamma with hyperparameters a_0 and b_0, which are constrained in the range from 0.3 to 10.

6. In DP-HP, given $\beta_1, \cdots, \beta_h, \cdots, \beta_H$, update (μ_0, σ_0^2) using the normal-inverse-gamma conjugacy form (see Carlin and Louis (2000)).

Compared to the DP model, there are some changes in the DPM model,

2. Update β_i by the adaptive rejection algorithm and with $\beta_i \sim (\mu_{Z_i}, \sigma_{Z_i}^2)$ as a priori.

7. Gibbs sampling of component-specific parameters and hyperparameters in G_0 using the Cauchy prior can be found in Gelman (2006).

References

Bender, R., Augustin, T., and Blettner, M. (2005), "Generating survival times to simulate Cox proportional hazards models." *Statistics in Medicine*, 24, 1713–1723.

Carlin, B. and Louis, T. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall/CRC, 2nd edition.

Celeux, G., Forbes, F., Robert, C., and Titterington, D. (2006), "Deviance information criteria for missing data models," *Bayesian Analysis*, 1, 651–674.

D, D. S., Sherlaw-Johnson, C., Bardsley, M., Blunt, I., C, C. W., and Grigg, O. (2012), "Statistical methods for healthcare regulation: rating, screening and surveilliance," *J Royal Stat Soc A*, 175, 1–47.

Dunson, D. (2010), *Nonparametric Bayes applications to biostatistics*, Cambridge: Cambridge University Press.

Escobar, M. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Gelman, A. (2006), "Prior distributions for variance parameters in hierachical models," *Bayesian Analysis*, 1, 515–533.

Gilks, W., Best, N., and Tan, K. (1995), "Adaptive rejection Metropolis sampling within Gibbs sampling," *Applied Statistics*, 44, 455–472.

Haario, H., Saksman, S., and Tamminen, J. (2001), "An adaptive Metropolis algorithm," *Bernoulli*, 7, 223–242.

He, K., Kalbfleisch, J. D., Li, Y., and Li, Y. (2013), "Evaluating hospital readmission rates in dialysis facilities with and without adjusting for hospital effects," *Lifetime Data Anal*, 19, 490–512.

Ibrahim, J. G., Chen, M.-H., and D.Sinha (2001), *Bayesian Survival Analysis*, New York: Springer.

Ishwaran, H. and James, L. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 101, 179–194.

Jones, H. E. and Spiegelhalter, D. J. (2011), "The identification of Unusual Heath-Care Providers from a Hierarchical Model," *Lifetime Data Anal*, 65, 154–163.

Kalbfleisch, J. (1978), "Non-Parametric Bayesian Analysis of Survival Time Data," *Journal of the Royal Statistical Society. Series B*, 40, 214–221.

Kalbfleisch, J. D. and Wolfe, R. A. (2013), "On monitoring outcomes ofmedical provider," *Stat Biosciences*, 40, 1–30.

Liu, J., Louis, T., Pan, W., Ma, J., and Collins, A. (2003), "Methods for estimating and interpreting provider-specific standardized mortality ratios," *Health Serv Outcomes Res Methodol*, 1754, 135–149.

Ohlssen, D., Sharples, L., and Spiegelhalter, D. (2007), "Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons," *Statistics in Medicine*, 26, 2088–112.

Richardson, S. and Green, P. (1997), "On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society, Series B*, 59, 731–792.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Yang, M., Dunson, D. B., and Baird, D. (2010), "Semiparametric Bayes hierarchical models with mean and variance constraints," *Computational Statistics and Data Analysis*, 54, 2172–2186.

Yisheng Li, P. M. and Lin, X. (2011), "Center-adjusted inference for a Non-parametric Bayesian random effect distribution," *Statistica Sinica*, 21, 1201–1223.