$\mathcal{I}^2\mathcal{SDS}$
The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2008-20
November 14, 2008

Assessment of Mortgage Default Risk via Bayesian Reliability Models

Refik Soyer
Department of Decision Sciences
The George Washington University, USA

Feng Xu
Department of Decision Sciences
The George Washington University, USA

Assessment of Mortgage Default Risk via Bayesian Reliability Models

Refik Soyer
Department of Decision Sciences
The George Washington University, USA

Feng Xu
Department of Decision Sciences
The George Washington University, USA

Abstract

In this paper we consider duration type models and their generalizations for modeling default risk. The models are motivated by noting similarities between reliability/survival analysis and mortgage default risk. We present Bayesian modeling strategies used in reliability analysis for describing time to default data. Our models include proportional hazards type generalized gamma and mixture models which are capable of capturing nonmonotonic default rates. We develop Bayesian inference for our models and illustrate their implementation using actual time to default data from the US mortgage market.

1. Introduction and Overview

As recent events suggested, performance of the residential mortgage market is a key to the stability of the U.S. economy and financial markets. Increase of risk in the mortgage market, as represented by increase of residential mortgage delinquency and foreclosure rates, has significant impact on the overall economy.

With the exception of early 1980s the U.S. national homeownership rate has increased from around 63% to above 68% in the last four decades. The U.S. residential mortgage market has developed tremendously in size during this period. The outstanding debt of single-family mortgage loans in the U.S. has grown from around $2.6 trillion in

1990 to above $9.8 trillion in the second quarter of 2006, representing an increase from about 45% to 74.5% of its share in the U.S. GDP during the period.

Due to the significant costs to mortgage loan lenders, investors of mortgage backed securities and borrowers, resulting from default, how to estimate and manage the default risk is one of the primary concerns for financial institutions and policy makers. There are alternate definitions of mortgage default used in the literature. The legal definition is given by Giliberto and Houston (1989) as the "transfer of the legal ownership of the property from the borrower to the lender either through the execution of foreclosure proceedings or the acceptance of a deed in lieu of foreclosure." Others who focus on modeling of default risk simply define default as being delinquent in mortgage payment for 90 days; see for example, Ambrose and Capone (1998). In this paper, we adopt the latter definition to distinguish default from foreclosure.

There exits a rich literature on morgage default risk. A detailed review is given by Quercia and Stegman (1992). An overview of more recent developments can be found in Leece (2004). One of the primary objectives of mortgage default modeling is identification of key individual borrower, property and loan characteristics affecting the likelihood of default.

One class of models is based on the *ruthless default assumption* which states that a rational borrower would maximize his/her wealth by defaulting on the mortgage if the market value of the mortgage exceeds the house value, and by prepaying via refinancing if the market value of the house exceeds book value of the house. Such models use an option theoretic approach and assume that the mortgage value and the prepayment and default options are determined by the stochastic behavior of variables such as property prices and the interest rates; see for example, Kau *et al.* (1990). Thus, under the option theoretic approach, other factors, such as the transaction costs, borrower's characteristics, etc., are assumed to have no impact on values of the mortgage and the property underlined. The ruthless default assumption is not universally accepted in the literature

2

and evidence against the validity of the assumption has been presented by many authors. Furthermore, implementation of this class of models requires availability of performance level data on individual loans over time which is typically difficult to obtain.

The alternate point of view, that does not subscribe to the ruthless default assumption, favors direct modeling of time to default of the mortgage. This approach involves hazard rate based models and also considers more direct determinants of mortgage default. This class of models includes competing and proportional hazards models of Lambrecht, Perraudin and Satchell (2003) and duration models of Lambrecht, Perraudin and Satchell (1997) that take into account individual borrower and loan characteristics.

In this paper we consider duration type models and their generalizations for modeling default risk. In so doing, we discuss connections between reliability analysis and mortgage default modeling and present Bayesian modeling strategies used in reliability literature for describing mortgage default risk. Use of Bayesian methods in residential mortgage default modeling has been limited to few papers such as Herzog (1988) who introduced some basic Bayesian concepts, Young and Kazarian (1997) who considered binary time-series regression models and more recently, Popova, Popova and George (2008) who proposed Bayesian methods for forecasting mortgage prepayment rates. Thus, the Bayesian models that are presented in this paper represent contributions to the literature in mortgage default risk. Our models include proportional hazards type generalized gamma and mixture models which have not been considered in the mortgage default literature. These models are capable of dealing with nonmonotonic default rates that are expected in mortgage time to default data.

A synopsis of our paper is as follows: In Section 2 we note the similarities between concepts in reliability analysis and mortgage default analysis. In so doing, we discuss characteristics of mortgage default data and propose models from reliability

3

literature that can capture these characteristics. In Section 3 we introduce generalized gamma proportional hazards models and mixtures of proportional hazards models for describing behavior early payment defaults. For both classes of models we develop Bayesian inference using Markov chain Monte Carlo (MCMC) methods in Section 4. Implementation of the models and the Bayesian methods to actual default data[1] are presented in Section 5.

2. Mortgage Default and Reliability Risk

Relationship between reliability (survival) analysis and financial risk has been noted by others in the reliability literature. For example, Lynn (2004) points out that concept of default of a bond in finance is analogous to failure in reliability analysis and proposes a counting process for describing the number of defaults of bond issuers. More recently, Singpurwalla (2007) notes the relationship between the survival function in reliability and the asset pricing formula of fixed income instrument such as a risk-free zero coupon bond in finance.

Similar to the above, we can see connections between mortgage default risk and reliability risk. The event of default of a mortgage, that is, being delinquent in mortgage payment for 90 days, is similar to failure of a system or a component in reliability analysis. Modeling time to default of a mortgage is analogous to modeling time to failure of a component. Thus, it is not surprising to find uses for reliability/survival analysis models in the mortgage default literature.

If T_i denotes time to default of the mortgage loan i with density function $f(t)$ and distribution function $F(t)$ then the failure rate is defined by

$$\lambda(t) = \frac{f(t)}{1 - F(t)}. \tag{2.1}$$

---

[1]The authors would like to thank Dr. Thomas N. Herzog and Teri Hines at FHA for providing the data for this work.

In (2.1) $\lambda(t)$ is sometimes referred to as the default rate. In the mortgage default literature it is not uncommon to model the default (hazard) rate using the proportional hazards model (PHM) of Cox (1972). Thus, for mortgage loan i with covariate vector $\boldsymbol{X}_i(t)$, the failure (default) rate is given by

$$\lambda_i(t) = \lambda_0(t)e^{\boldsymbol{\beta}'\boldsymbol{X}_i(t)} \tag{2.2}$$

where $\lambda_0(\bullet)$ is the baseline failure rate, and $\boldsymbol{\beta}$ is the parameter vector. In the above $\lambda_0(t)$ represents the effect of age of the mortgage on the probability of default whereas $e^{\boldsymbol{\beta}'\boldsymbol{X}_i(t)}$ represents the effects of different covariates. The attractive feature of the PHM is that it allows for incorporating covariate effects in modeling the hazard rate. Under the PHM, the ratio of the default rates for two mortgages, say i and $j,$ at the same age t is given by

$$\frac{\lambda_i(t)}{\lambda_j(t)} = e^{\boldsymbol{\beta}'[\boldsymbol{X}_i(t) - \boldsymbol{X}_j(t)]}. \tag{2.3}$$

Thus, the ratio of default rates for two mortgages with different risks, as implied by (2.3), is proportional to a function of the respective covariates. In our development and data analysis we will be using time independent covariates. Thus, in what follows the covariate vector will be written as $\boldsymbol{X}_i$.

Duration models have recently got much attention in the literature for modeling mortgage default. In recognition of the fact that the hazard rate is not monotonic for mortgage default, Lambrecht et al. (1997) suggest a generalization of the Weibull failure rate to describe time to default in the U.K. mortgage market. The authors define the baseline failure rate as

$$\lambda_0(t) = \alpha t^{\alpha-1} exp(-\gamma t), \tag{2.4}$$

which reduces to the Weibull hazard for $\gamma = 0$. The proposed failure model implies that default rate increases in early years of the mortgage and then decreases afterwards. The

model also included other determinants of default by introducing a covariate component $e^{\boldsymbol{\beta}' \boldsymbol{X}_i(t)}$ in the scale parameter in (2.4).

The nonmonotonic failure rate behavior described by (2.4) can be obtained by using more flexible class of models that are used in reliability literature. In what follows, we will present the generalized gamma and mixture models and discuss their characteristics. To the best of our knowledge both classes of models have not been previously considered in the mortgage default literature.

3. Reliability Models for Time to Default

In view of our discussion of nonmonotonic failure rates of time to default data we will first present the class of *generalized gamma models* that includes many known duration models such as exponential, gamma, lognormal and Weibull as special cases. Our discussion will follow with the *mixture models* that are common in the reliability literature when pooling heterogeneous failure data; see for example Gurland and Sethuraman (1994). For both classes of models we will consider PHM extensions.

3.1 Generalized Gamma Models

As noted by Dadpay et al. (2007) the generalized gamma distribution offers lot of flexibility in duration modeling as it allows for various hazard patterns. The distribution is originally introduced by Stacy (1962) in reliability modeling. It has been considered in the economics [see Jaggia (1991)] and marketing literatures [see Allenby et. al (1999)]. Zhang et al. (2001) considered the generalized gamma distribution for modeling the duration of stock transactions. In what follows, we will consider it as a model for time to mortgage default.

6

As before we denote time to default of loan i by T_i and assume a generalized gamma model with density function

$$f(t \mid \alpha, \gamma, \lambda) = \frac{\gamma \lambda^\alpha t^{\alpha\gamma - 1}}{\Gamma(\alpha)} exp\{-\lambda t^\gamma\}, \tag{3.1}$$

where $\alpha, \gamma, \lambda > 0$. An attractive feature of the generalized gamma distribution is that many of the well-known duration models are obtained as special cases of the density given by (3.1). For example, exponential model ($\gamma = \alpha = 1$), Weibull model ($\alpha = 1$), gamma model ($\gamma = 1$) and the half normal model ($\gamma = 2$, $\alpha = 1/2$). The lognormal model is also obtained as $\alpha \to \infty$. Thus, the generalized gamma model is more general than the generalized Weibull model proposed by Lambrecht *et al.* (1997).

The cumulative distribution of T_i is given by

$$F(t \mid \alpha, \gamma, \lambda) = \frac{\Gamma_{\lambda t^\gamma}(\alpha)}{\Gamma(\alpha)}, \tag{3.2}$$

where

$$\Gamma_{\lambda t^\gamma}(\alpha) = \int_0^{\lambda t^\gamma} u^{\alpha - 1} exp(-u) \, du. \tag{3.3}$$

Thus, the baseline default rate can be obtained as

$$\lambda_0(t; \alpha, \gamma, \lambda) = \frac{\gamma \lambda^\alpha t^{\alpha\gamma - 1} exp\{-\lambda t^\gamma\}}{\Gamma(\alpha) - \Gamma_{\lambda t^\gamma}(\alpha)}. \tag{3.4}$$

We note that the numerator of (3.4) is quite similar to the baseline default rate considered by Lambrecht *et al.* (1997) in (2.4). The default rate $\lambda_0(t; \alpha, \gamma, \lambda)$ in (3.4) is not necessarily monotonic and is capable of representing a wide variety of failure rate behavior. It is shown by Pham and Almhana (1995) that for $\gamma \neq 1$, if

$$\frac{(1 - \alpha\gamma)}{\gamma(\gamma - 1)} > 0 \text{ and } 0 < \gamma < 1 \tag{3.5}$$

then the failure rate will be first increasing and then decreasing. If $\gamma > 1$ in (3.5) then the failure rate takes a *bathub shape*. If we specify $\alpha = 1.95$, $\gamma = 0.55$ and $\lambda = 0.1$, then the default (failure) rate takes the nonmomotonic behavior is shown in Figure 1 below.

The effect of other determinants of default rate can be incorporated into the model and a PHM type representation can be obtained for the generalized gamma model. If we write the default rate as

$$\lambda_i(t, \boldsymbol{X_i}) = \lambda_0(t;\ \alpha, \gamma, \lambda)\ exp(\boldsymbol{\beta' X_i}). \tag{3.6}$$

where $\lambda_0(t;\ \alpha, \gamma, \lambda)$ is given by (3.4), then it can be easily shown that the density function is

$$f(t \mid \alpha, \gamma, \lambda, \boldsymbol{\beta}, \boldsymbol{X_i}) = \ \frac{\gamma \lambda^\alpha\ t^{\alpha\gamma - 1}}{\Gamma(\alpha)} exp(\alpha\boldsymbol{\beta' X_i})\ exp[(-\ \lambda t^\gamma exp(\boldsymbol{\beta' X_i})]. \tag{3.7}$$
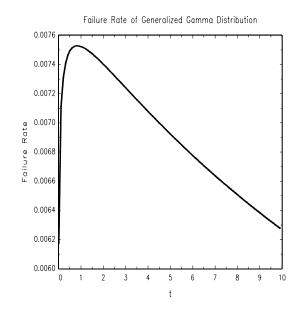

Figure 1. Failure Rate of Generalized Gamma Distribution for Specified Parameters.

3.2 Mixture Models

An alternate strategy to generalize the duration models is to use finite mixture models. As noted by Dieboldt and Robert (1994), "Mixture models provide an interesting

8

alternative to nonparametric modeling, while being less restrictive than the usual distributional assumptions." The mixture models are commonly used in the reliability literature and more specifically in *burn-in* testing where the population is assumed to consist of two subpopulations referred to as *weak* and *strong*; see for example, Lynn and Singpurwalla (1997).

Relevance of mixture models in mortgage default is due to the recent increase in subprime[2] mortgages in the US market. Krinsman (2007) points out that as a result of this increase, *early payment defaults* have become common in the market. Early payment defaults (EPDs) are usually characterized as those loans that defaulted within 12 months of their origination. The weak and strong composition which is seen in burn-in testing is also applicable to EPDs. Figure 2 below illustrates the histogram and the density plot of time to default for a randomly selected sample of EPDs originated during 2001. From the figure, we can clearly see the presence of mixtures in the data.
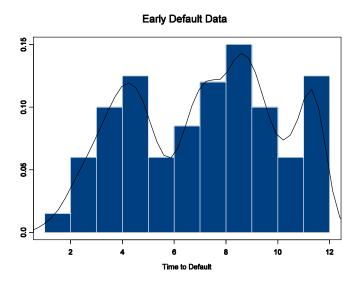

Figure 2. Distribution of Early Default Data from 2001.

---

[2]Subprime mortgages are loans given to borrowers with bad credit scores.

By using mixtures of increasing failure rate (IFR) distributions one can reflect nonmonotonic failure rates. Mi (1993) notes that mixtures of IFR distributions can be used to measure what he refers to *upside down* failure rate, which seems to be the expected behavior of mortgage default rate.

For example, we can define a two component mixture model for T_i as

$$f_{T_i}(t) = \pi f_1(t) + (1 - \pi) f_2(t) \tag{3.8}$$

where π is the mixing probability (or the mixing weight). If we choose $\pi = 0.5$ and both $f_1(t)$ and $f_2(t)$ as Weibull distributions with parameters ($\alpha_1 = 1$, $\gamma_1 = 1$) and ($\alpha_2 = 1$, $\gamma_2 = 2$) respectively, then we can obtain a nonmonotonic failure (or default) rate as shown in Figure 3.

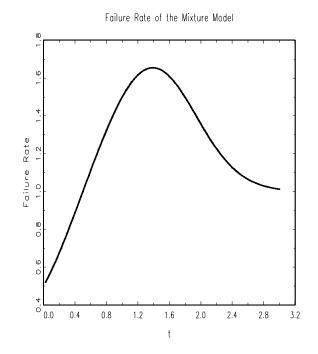Figure 3. Failure Rate of Mixture of Two Weibull Densities

In general we can define a $K -$ component mixture model for T_i as

$$f(t) = \sum_{k=1}^{K} \pi_k f_k(t|\phi_k) \tag{3.9}$$

10

where $\sum_{k=1}^{K} \pi_k = 1$ and ϕ_k are the parameters of the kth component in the mixture. For example, for mixture of Weibull densities

$$f_k(t|\phi_k) = \alpha_k \gamma_k t^{\gamma_k - 1} exp[-\alpha_k t^{\gamma_k}] \tag{3.10}$$

we have $\phi_k = (\alpha_k, \gamma_k)$. It is important to note that when we are given default times from n different loans, we do not know from which distribution ith default time t_i is coming from. Thus, we can consider this as a missing data problem and introduce latent variables S_{ik} for $k = 1, \ldots, K$ of each observation such that

$$S_{ik} = \begin{cases} 1 & \text{if } t_i \sim f_k(t|\phi_k) \\ 0 & \text{otherwise,} \end{cases} \tag{3.11}$$

and for each i we have $\sum_{k=1}^{K} S_{ik} = 1$.

If we define the latent vector for the ith observation as $\boldsymbol{S}_i = (S_{i1} \ldots S_{iK})$, then given the above setup, we can assume that $\boldsymbol{S}_i$'s are independent multinomially distributed vectors denoted as

$$\boldsymbol{S}_i|\pi_1, \ldots, \pi_K \sim Mult(1; \pi_1, \ldots, \pi_K). \tag{3.12}$$

Note that (3.12) implies that only one of the components of $\boldsymbol{S}_i$ is 1 and the remaining are 0's. It follows from the above that

$$t_i|\boldsymbol{S}_i, \boldsymbol{\phi} \sim f_k(t|\phi_k) \tag{3.13}$$

where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$.

A PHM type of mixture model can also be developed. The PHM setup is conceptually similar to mixtures of normal regression models that are considered in Hurn, Justel and Robert (2003) and mixtures Weibull regression models that are used by Attardi, Guida and Pulcini (2005). In our development we will consider mixtures of Weibull PHMs. In so doing, we take the failure rate of the kth Weibull density component (3.10), that is, $\alpha_k \gamma_k t^{\gamma_k - 1}$ and rewrite it as

$$\lambda_k(t, \boldsymbol{X}) = \alpha_k \gamma_k t^{\gamma_k - 1} \, exp(\boldsymbol{\beta_k'} \boldsymbol{X}). \tag{3.14}$$

Thus, we can write the mixture model (3.9) as

$$f(t) = \sum_{k=1}^{K} \pi_k f_k(t | \phi_k, \boldsymbol{\beta_k}, \boldsymbol{X}), \tag{3.15}$$

where the kth component density is given by

$$f_k(t | \phi_k, \boldsymbol{\beta_k}, \boldsymbol{X}) = \alpha_k \gamma_k t^{\gamma_k - 1} exp(\boldsymbol{\beta_k'} \boldsymbol{X}) exp(-\alpha_k t^{\gamma_k} exp(\boldsymbol{\beta_k'} \boldsymbol{X})). \tag{3.16}$$

Similar to (3.13) we can write

$$t_i | \boldsymbol{S}_i, \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{X_i} \sim f_k(t | \phi_k, \boldsymbol{\beta_k}, \boldsymbol{X_i}) \tag{3.17}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$.

Note that we can consider different type of mixture models by assuming some of the elements of $(\phi_k, \boldsymbol{\beta_k})$ are common to all components. Such alternative models will be analyzed in Section 5 for EPD data.

4. Bayesian Inference for Time to Default Models

We next present Bayesian inference for the generalized gamma PHM and the mixtures of Weibull PHM using time to default data on n loans. In the case of the mixture models we assume the data is on EPDs. Since in the time to default data of Section 5 all covariates are at the time of initiation of the mortgage, we denote the covariates by $\boldsymbol{X_i}$ in our development.

4.1 Bayesian Analysis of the Generalized Gamma PHM

Bayesian analysis of PHMs using MCMC methods has been considered by Dellaportes and Smith (1993) for the Weibull failure model. A similar approach can be used for the generalized gamma PHM.

Given time to default data and covariate information $D = (t_1, \ldots, t_n; \boldsymbol{X}_1, \ldots, \boldsymbol{X_n})$ on n loans, the joint likelihood function of $(\alpha, \gamma, \lambda, \boldsymbol{\beta})$ can be written as

$$L(\alpha, \gamma, \lambda, \boldsymbol{\beta}; D) = \prod_{i=1}^{n} \frac{\gamma \lambda^{\alpha} t_i^{\alpha\gamma-1}}{\Gamma(\alpha)} exp(\alpha \boldsymbol{\beta}' \boldsymbol{X_i}) \, exp[(- \lambda t_i^{\gamma} exp(\boldsymbol{\beta}' \boldsymbol{X_i})]. \quad (4.1)$$

For any choice of a prior, the joint posterior distribution $p(\alpha, \gamma, \lambda, \boldsymbol{\beta}|D)$ can not be obtained analytically, but a Gibbs sampler, similar to considered by Dellaportes and Smith (1993), can be used for developing posterior and predictive inferences. It can be shown that assuming independent logconcave priors for α, γ, and λ and a multivariate normal prior for $\boldsymbol{\beta}$ independently, all the full conditionals are logconcave and therefore the adaptive rejection sampling method can be used to draw from the full conditional densities. In our analysis we specify gamma priors on α and γ and a lognormal prior on λ.

Given posterior samples $\left\{ \alpha^{(g)}, \gamma^{(g)}, \lambda^{(g)}, \boldsymbol{\beta}^{(g)} \right\}_{g=1}^{G}$ from the joint posterior distribution, the posterior predictive distributions for default rates and time to default, can be obtained. Note that the baseline default rate function for the generalized gamma PHM is given by (3.4) which is a function of α, γ, and λ. As pointed out by Lynn and Singpurwalla (1997), one can <u>not</u> obtain the posterior predictive baseline default rate $\lambda_0(t|D)$ by integrating out (3.4) using $p(\alpha, \gamma, \lambda|D)$. The posterior predictive baseline default rate is given by

$$\lambda_0(t|D) = \frac{\int f(t|\alpha, \gamma, \lambda) p(\alpha, \gamma, \lambda|D) d\alpha \, d\gamma \, d\lambda}{\int [1 - F(t|\alpha, \gamma, \lambda)] \, p(\alpha, \gamma, \lambda|D) d\alpha \, d\gamma \, d\lambda}, \quad (4.2)$$

where $f(t|\alpha, \gamma, \lambda)$ and $F(t|\alpha, \gamma, \lambda)$ are given by (3.1) and (3.2), respectively. Note that due to use of time independent covariates, (4.2) does not depend on the covariate terms. Using the posterior samples from the joint distribution, we can approximate the posterior predictive default rate (4.2) as

$$\lambda_0(t|D) \simeq \frac{\frac{1}{G}\sum_{g=1}^{G} f(t|\ \alpha^g, \gamma^g, \lambda^g)}{\frac{1}{G}\sum_{g=1}^{G}[1 - F(t|\ \alpha^g, \gamma^g, \lambda^g)]}. \tag{4.3}$$

The posterior predictive distribution for time to default of any loan i with covariate information $\boldsymbol{X_i}$ is given by

$$f(t_i|D, \boldsymbol{X_i}) = \int f(t\ |\ \alpha, \gamma, \lambda, \boldsymbol{\beta}, \boldsymbol{X_i})\, dP(\alpha, \gamma, \lambda, \boldsymbol{\beta}|D), \tag{4.4}$$

where $f(t\ |\ \alpha, \gamma, \lambda, \boldsymbol{\beta}, \boldsymbol{X_i})$ is given by (3.7). We can approximate the above using the Monte Carlo average

$$f(t_i|D, \boldsymbol{X_i}) \simeq \frac{1}{G}\sum_{g=1}^{G} f(t_i|\alpha^g, \gamma^g, \lambda^g, \boldsymbol{\beta}^g, \boldsymbol{X_i}). \tag{4.5}$$

4.2 Bayesian Analysis of the Mixtures of Weibull PHM

As in the previous case we assume that we have time to default data and covariate information D on n EPD loans. Following Section 3.2, we consider mixtures of Weibull PHMs where each component of the mixture has the density (3.16). Bayesian analysis of mixtures of Weibull distributions has been considered in Tsionas (2002), but, to the best of our knowledge, Bayesian analysis of finite mixtures of Weibull PHMs has not been developed.

Bayesian modeling of mixtures requires specification of prior distributions for all the unknown parameters. The mixing probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ can be assumed to have a Dirichlet prior distribution as

$$p(\boldsymbol{\pi}) \propto \prod_{k=1}^{K} \pi_k^{\psi_k - 1} \tag{4.6}$$

14

with specified parameters ψ_k's. We can assume independent priors for elements of $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ vectors, and also assume that $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ are independent of each other as well as of $\boldsymbol{\pi}$. Thus, we can write

$$p(\boldsymbol{\phi}, \boldsymbol{\beta}) = \prod_{k=1}^{K} p(\alpha_k, \gamma_k) p(\boldsymbol{\beta_k}). \tag{4.7}$$

Furthermore, we specify independent gamma priors for α_k and γ_k and a multivariate normal prior for $\boldsymbol{\beta_k}$ in (4.7).

Given the data D, the joint likelihood function of $(\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{S}, \boldsymbol{\pi})$, where $\boldsymbol{S} = \{\boldsymbol{S}_i;\ i = 1, \dots n\}$, is given by

$$L(\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{S}, \boldsymbol{\pi}; D) \propto \prod_{i=1}^{n} \prod_{k=1}^{K} [f(t_i | S_{ik}, \phi_k, \boldsymbol{\beta_k}, \boldsymbol{X_i})]^{S_{ik}} (\pi_k)^{S_{ik}}. \tag{4.8}$$

The joint posterior distribution of $p(\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{S}, \boldsymbol{\pi} | D)$ can not be obtained analytically, but a fully Bayesian analysis can be developed using MCMC. Following a development similar to what is presented in Diebolt and Robert (1994) and Hurn, Justel and Robert (2003), we can use a Gibbs sampler.

Using independent gamma priors for α_k's with parameters a_k and b_k and defining $M_k = \sum_{i=1}^{n} \mathbf{1}(S_{ik} = 1)$, the full conditional distribution of α_k can be obtained as a gamma distribution with parameters $(M_k + a_k)$ and $[t_i^{\gamma_k} exp(\boldsymbol{\beta'_k X_i}) + b_k]$. The full conditional distribution of mixing probability vector $\boldsymbol{\pi}$ can be obtained as a Dirichlet distribution with parameters $\{M_k + \psi_k : k = 1, ..., K\}$. The full conditionals of $\boldsymbol{S}_i$'s can be shown to be multinomial as $Mult(1; \pi_1(t_i), \dots, \pi_K(t_i))$, where

$$\pi_k(t_i) = \frac{\pi_k\, f(t_i | \phi_k, \boldsymbol{\beta_k}, \boldsymbol{X_i})}{\sum_{j=1}^{K} \pi_j\, f(t_i | \phi_j, \boldsymbol{\beta_j}, \boldsymbol{X_i})}. \tag{4.9}$$

The full conditionals of γ_k's and $\boldsymbol{\beta}_k$'s are not of known forms, but they can be shown to be log-concave and therefore adaptive rejection sampling methods can be used to draw from these distributions.

Once posterior samples are obtained for all unknown quantities we can obtain posterior predictive densities for time to default as in the generalized gamma PHM. To obtain the predictive distribution $f(t|D),$ for each posterior sample $\boldsymbol{\pi}^{(g)}$ of $\boldsymbol{\pi}$ we generate the vector $\{S_k^{(g)} ; k = 1, \ldots, K\}$ which has only one component with value 1 and the rest with 0's. After repeating this G times we compute

$$f(t|D) \approx \frac{1}{G}\Big(\sum_{k=1}^{K} 1(S_k)^g f_k(t|D)\Big) \qquad (4.10)$$

where

$$f_k(t|D) \approx \frac{1}{G}\sum_{r=1}^{G} f_k(t|\phi_k^g, \boldsymbol{\beta_k^g}, \boldsymbol{X}). \qquad (4.11)$$

4.3 Bayesian Model Comparison and Fit Measures

Assume that we are considering two alternative models M_1 and M_2 for time to default data. Computation of Bayes factor (BF)

$$BF = \frac{p(D|M_1)}{p(D|M_2)} \qquad (4.12)$$

is a challenging task since the marginal likelihood $p(D|M_i)$ for model i can not be obtained analytically in our case. One alternative comparison measure is the "*posterior Bayes factors*" suggested by Aitkin (1991) which evaluates marginal likelihood using the posterior distribution instead of the prior distribution. This provides us with the posterior mean of the marginal likelihood term which has been suggested as a fit measure by others in the literature including Dempster (1974). The posterior mean of the marginal likelihood term can be evaluated using Monte Carlo method by drawing samples from the posterior. More specifically, we can approximate

$$p(D|M_i) \approx \frac{1}{G} \sum_{g=1}^{G} p(D|\Theta^g, M_i), \tag{4.13}$$

where Θ is a generic parameter vector and $\{\Theta^g\}_{g=1}^{G}$ are samples from $p(\Theta|D)$. We note that (4.13) provides us with a retrospective measure and therefore we will refer to the posterior Bayes factors as *retrospective Bayes factors* (RBF) in Section 5. The posterior mean of the marginal likelihood term is also related to the deviance concept [see for example, Spiegelhalter *et al.* (2002)].

An alternative comparison can be based on predictive performance of the models. If data D can be decomposed into two parts $D = (D_0, D_F)$ then original data D_0 can be used to update the parameters of each model and future data D_F can be used to evaluate the posterior predictive density. In other words, for each model we can obtain the marginal likelihood as

$$p(D_F|D_0, M_i) = \int p(D_F|\Theta, M_i) p(\Theta|D_0) \, d\Theta, \tag{4.14}$$

which we can approximate using a Monte Carlo average as

$$p(D_F|D_0, M_i) \approx \frac{1}{G} \sum_{g=1}^{G} p(D_F|\Theta^g, M_i), \tag{4.15}$$

where $\{\Theta^g\}_{g=1}^{G}$ are samples from $p(\Theta|D_0)$. Note that we can compute Bayes factors using predictive marginal likelihoods based on (4.15). We will refer to these as the *predictive Bayes factors* (PBF) in Section 5.

5. Illustrations Using Actual Time to Default Data

Mortgage default data used in this section is from FHA's regional office at Atlanta. The data in Section 5.1 consists of default times on FRM 30-year loans originated during 1994 period. Besides loan endorsement and default dates, other information provided in the dataset includes loan amount, interest rate, borrower's

effective household income, loan to value ratio (LTV), borrower's marital status and age, all recorded at loan origination time.

For analysis of generalized gamma PHM we randomly select 400 samples from defaulted mortgage loans originated in 1994. Our data is similar to what is considered by Lambrecht, Perraudin and Satchell (1997) in that we have only defaulted loans and all covariates are at the time of initiation of the mortgage.

The data that we use in Section 5.2 for the analysis of mixtures of PHMs consists of default times on FRM 30-year EPD loans originated during 2001. In this case we randomly select 200 sample loanss from all EPD loans originated in 2001. Again we have the same covariate information available on each EPD loan.

5.1 Analysis of Regular Default Data

In our analysis of the sample of 400 loans, we consider the generalized gamma PHM and we compare its performance with the Weibull PHM. In so doing, we will use the retrospective and predictive Bayes factors of Section 4.3. We will be using proper but diffused priors for all parameters in the analysis.

In Table 1 we present posterior summaries of the parameters of the model including the 95% Bayesian central credibility intervals (CCI). We note from the table that the effects of mortgage interest rate and loan amount on failure rate are consistent with what is expected, that is, higher initial interest rate and higher loan amount would leave borrower with heavier financial burden, resulting in a higher propensity to default. Furthermore, income, marital status and LTV do not seem to have clear effects on the default. This may due to the fact that the default risk is more influenced by changes in these covariates rather their initial levels at loan origination. Similar findings are obtained for the Weibull PHM.

Parameter	Mean	StDev	95% CCI
α	1.318	0.444	(0.718, 2.392)
γ	1.422	0.260	(0.953,1.981)
$\beta_{\text{interest-rate}}$	1.449	0.531	(0.526, 2.624)
$\beta_{\text{loan-amount}}$	0.846	0.264	(0.391,1.423)
$\beta_{\text{annual-income}}$	-0.454	0.435	(-1.373, 0.3284)
$\beta_{\text{marital-status}}$	0.146	0.487	(-0.845, 1.077)
$\beta_{\text{borrower-age}}$	-0.820	0.467	(-1.782,0.026)
β_{LTV}	0.031	0.419	(-0.791,0.850)

Table 1. Posterior Summaries from the Generalized Gamma Model

We can also see from Table 1 that most values of α are greater than 1 which suggests that the failure rate behaves different than that of the Weibull model. We can compute the posterior probability that α is greater than 1 as 0.762. We can also see the potential deviation in the data from the Weibull model by looking at the posterior failure rate of the model. In Figure 4 we present the plot of expected baseline failure rate from the generalized gamma model based on (4.3). The plot provides evidence in favor of a nonmonotonic default behavior.
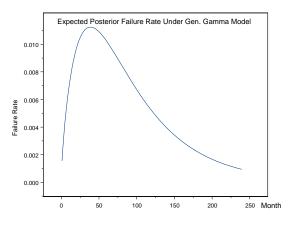

Figure 4. Expected Posterior Baseline Failure rate under the Gamma Model.

Comparison of the generalized gamma PHM with the Weibull PHM using the retrospective BF suggests a 1.7 value in favor of the generalized gamma model. Note that this value is not considered as an indication of strong evidence; see Kass and Raftery

(1995). To assess the out of sample predictive performance of the models we next compared them using predictive BFs. In so doing, we used 100 randomly selected samples of size 10 from the defaulted loans and compute PBFs. Out of 100 samples 63 of them favored the generalized gamma PHM based on the PBF.

5.2 Analysis of EPD Loan Data

In our analysis of the sample of 200 EPD loans from 2001 we will use the mixtures of Weibull PHMs discussed in Sections 3.2 and 4.2. In our analysis we define EPD's as loans defaulted within 12 months of their origination. We have presented the probability histogram and the density plot of the sample in Figure 2 of Section 3.2.

In what follows, we will fit two-component mixtures of Weibull PHMs. In so doing, we will consider different types of mixture models. Our first model considers a mixture with different shape parameters but common scale and covariate coefficients. In other words, in section 3.2, we have $(\alpha, \boldsymbol{\beta}, \gamma_1)$ and $(\alpha, \boldsymbol{\beta}, \gamma_2)$ as the parameters of the two components.

In our analysis we use diffused but proper priors for all parameters of the model. In Figure 5 we present the posterior distributions of covariate coefficients. We note that the initial interest rate of the loan causes an increase in the default rate. For all other cases except the marital status the posterior distributions are assigning a high density value to zero. Again, this may be due to the fact the covariates represent values at the loan origination. In Figure 6 we present the posterior distributions of shape parameters γ_1 and γ_2. We clearly see from the figure that the first posterior distribution has a mean around 6 whereas the second one has mean around 4. Based on the joint posterior distribution we can compute $Prob(\gamma_1 > \gamma_2|D) \approx 1$.

The posterior distribution of the mixing probability π is given in Figure 7. We can see from the figure that the expected value of π is in the vicinity of 0.2, implying that on

average 20% of the time we will observe defaults coming from the first component of the mixture.
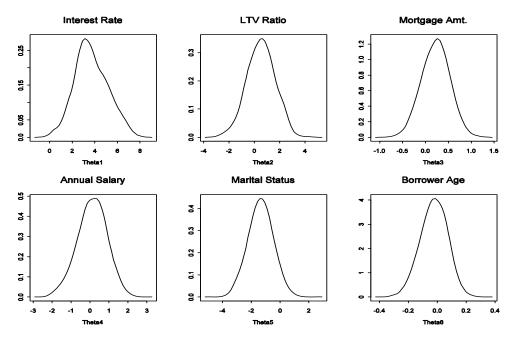

Figure 5. Posterior Distributions of Covariate Parameters in the Mixture Model.
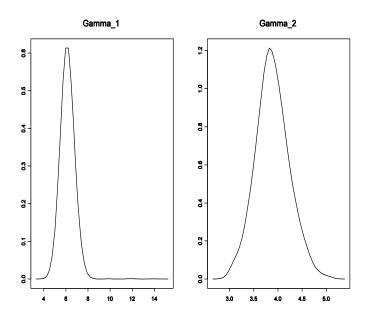

Figure 6. Posterior Distributions of the Shape Parameters

We also analyzed the early default data using a Weibull PHM without a mixture. We have found that the fit provided by the mixture model is lot superior to the Weibull PHM. The retrospective *BF* in favor of the mixture model was much larger than 150. Thus, the evidence in favor of the mixture model is very strong for the early default data.

Posterior Distribution of pi

Figure 7 Posterior Distribution of Mixing Probability π for EPD Data.

In our second model we considered a mixture with different covariate coefficients but common shape and scale parameters for the components. Thus, in section 3.2, we have $(\alpha, \gamma, \boldsymbol{\beta_1})$ and $(\alpha, \gamma, \boldsymbol{\beta_2})$ as the parameters of the two components. Again we used diffused but proper priors for all coefficients. The posterior summaries including 95% Bayesian central credibility intervals for γ, π, $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are shown in Table 2. From the table we see that there is evidence of two components with regards to the effect of interest rate on default rate. In both cases higher mortgage interest rate implies higher probability of default but in one group this effect is stronger. Also, in one of the groups there is evidence of the negative effect of borrower's age on default rate, that is, the older the person the less likely the default. Similarly, for one of the groups, the marital status has a clear negative effect on the default rate.

22

Parameter	Mean	StDev	95% CCI
$\beta_{\text{interest-rate, 1}}$	8.720	3.425	(2.569, 15.820)
$\beta_{\text{interest-rate, 2}}$	3.096	1.626	(0.005, 6.387)
$\beta_{\text{loan-amount,1}}$	0.415	0.917	(-1.163, 2.437)
$\beta_{\text{loan-amount,2}}$	0.185	0.343	(-0.475, 0.859)
$\beta_{\text{annual-income,1}}$	2.247	2.040	(-2.116, 5.882)
$\beta_{\text{annual-income,2}}$	-0.603	0.906	(-2.401, 1.151)
$\beta_{\text{marital-status,1}}$	2.405	3.322	(-4.844, 8.433)
$\beta_{\text{marital-status,2}}$	-1.88	0.955	(-3.681, -0.050)
$\beta_{\text{borrower-age,1}}$	-4.047	2.134	(-8.028, 0.381)
$\beta_{\text{borrower-age,2}}$	0.422	1.058	(-1.702, 2.431)
$\beta_{\text{LTV,1}}$	-0.948	3.42	(-7.725, 5.903)
$\beta_{\text{LTV,2}}$	0.662	1.231	(-1.722, 3.120)
γ	4.787	0.393	(4.001, 5.548)
π	0.285	0.047	(0.196, 0.379)

Table 2. Posterior Summaries from the Mixtures of Weibull PHM.

We made a comparison of the two mixture models, that is, the model with different shape parameters but common covariate coefficients (Model 1) and the model with different covariate coefficients but common shapes (Model 2). The retrospective BF in favor of Model 2 was much larger than 150. We also obtained predictive BF based on a randomly selected additional 20 EPD loans. This was 28 to 1 also in favor of Model 2. When we analyzed the data using a mixture model where both shape parameters and covariate coefficients were different, the results were very similar to those obtained under Model 2 with shape parameters did not seem to differ between the two groups. Thus, the evidence suggests that the mixture may be due to the different covariate effects on default rate in the two groups.

References

Ambrose, B. W., and C. A. Capone, Jr. (1998). Modeling the Conditional Probability of Foreclosure in the Context of Single-Family Mortgage Default Resolutions. *Real Estate Economics* 26 (3): 391-429.

Aitkin, M. (1991). Posterior Bayes Factors (with discussion). *Journal of the Royal Statistical Society, Ser. B* 53: 111-142.

Attardi, L., M.Guida, and G. Pulcini (2005). A mixed-Weibull regression model for the analysis of automotive warranty data. *Reliability Engineering & System Safety*, 87 (2): 265-273.

Cox, D. R. (1972). Regression Models and Life-tables (with Discussion). *Journal of the Royal Statistical Society*, Series B, 34 (2): 187-220.

Dadpay, A., E. S. Soofi, and R. Soyer (2007). Information Measures for Generalized Gamma Family. *Journal of Econometrics*, 138: 568-85.

Dellaportes, P., and A.F.M. Smith (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics* 42: 443-459.

Dempster, A. P. (1974). The Direct use of Likelihood for Significance Testing. in *Proceedings of Conference on Foundational Questions in Statistical Inference*, O. Barndorff-Nielsen, P. Blaesild and G. Schou (Eds.), 335-352.

Diebolt, J. and C. P. Robert (1994). Estimation of Finite Mixture Distributions through Bayesian Sampling. *Journal of the Royal Statistical Society, Ser. B*, 56: 363-75.

Giliberto, S. M., and A. L. Houston (1989). Relocation Opportunities and Mortgage Default. *AREUEA Journal* 17(1): 55-69.

Gurland, G. and J. Sethuraman (1994). Reversal of Increasing Failure Rates When Pooling Failure Data. *Technometrics* 36: 416-418.

Hurn, M., A. Justel and, C. P. Robert (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics* 12: 1-25.

Kass, R., and A. Raftery (1995). Bayes Factors. *Journal of American statistical Association* 90: 773-795.

Kau, J. B., D. C. Keenan, W. J. Muller, and J. F. Epperson (1990). Pricing Commercial Mortgages and Their Mortgage-Backed Securities. *The Journal of Real Estate Finance and Economics* 3: 333-36.

Krinsman, A. N. (2007). Subprime mortgage meltdown: How did it happen and how will it end ? *The Journal of Structured Finance*, 13 (2):1-9.

Lambrecht, B., W. Perraudin, and S. Satchell (1997). Time to default in the UK mortgage market. *Economic Modelling* 14: 485-99.

Lambrecht, B., W. Perraudin, and S. Satchell (2003). Mortgage default and possession under recourse: A competing hazards approach. *Journal of Money, Credit and Banking* 35: 425-442.

Leece, D. (2004). *Economics of the Mortgage Market: Perspectives on Household Decision Making.* Blackwell Publishing Ltd.

Lynn, N. J. and N. D. Singpurwalla (1997)."Burn-In" Makes us Feel Good. *Statistical Science* 12: 13-19.

Lynn, N. J. (2004). The Price of Failure. in *Mathematical Reliability: An Expository Perspective,* R. Soyer, T. A. Mazzuchi, and N. D. Singpurwalla, Eds., 303-16.

Mi, J. (1993). Discrete Bathtub Failure rate and Upside Down Failure Rate. *Naval Research Logistics* 40, 361-371.

Quercia, R. G., and M. A. Stegman (1992). Residential Mortgage Default: A Review of the Literature. *Journal of Housing Research* 3 (2): 341-79.

Popova, I., Popova, E., and George, E. I. (2008). Bayesian forecasting of prepayment rates for individual pools of mortgages. *Bayesian Analysis*, 3:393-426.

Pham, T., and J. Almhana (1995). The generalized gamma distribution: Its hazard rate and stress-strength model. *IEEE Transactions on Reliability*, 44 (3): 393-397.

Singpurwalla, N. D. (2007). Reliability and survival in financial risk. *Institute of Mathematical Statistics: Festschrift for K. Doksum.*

Spiegelhalter, D., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64: 1-34.

Stacy, E. W. (1962). A Generalization of the Gamma Distribution. *The Annals of Mathematical Statistics*, 33 (3): 1187-92.

Tsionas, E. (2002). Bayesian analysis of finite mixtures of Weibull distributions. *Communications in Statistics: Theory and Methods* 31(1): 37-48.