European Journal of Operational Research 274 (2019) 1112–1124

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

A Bayesian approach to modeling mortgage default and prepayment

Arnab Bhattacharya[a,*], Simon P. Wilson[a], Refik Soyer[b]

[a] Department of Statistics, School of Computer Science and Statistics, Trinity College, Dublin 02, Ireland
[b] Department of Decision Sciences, The George Washington University, Washington, DC 20052, USA

ARTICLE INFO

Article history:
Received 12 May 2017
Accepted 28 October 2018
Available online 1 November 2018

Keywords:
Reliability
Proportional hazards model
Competing risks
MCMC

ABSTRACT

In this paper we present a Bayesian competing risk proportional hazards model to describe mortgage defaults and prepayments. We develop Bayesian inference for the model using Markov chain Monte Carlo methods. Implementation of the model is illustrated using actual default/prepayment data and additional insights that can be obtained from the Bayesian analysis are discussed.

1. Introduction and Overview

From a legal point of view mortgage default is defined as "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"; see Gilberto and Houston Jr. (1989). However, as noted by Ambrose and Capone (1998), it is common in the literature to define default as being delinquent in mortgage payment for ninety days.

As noted by Soyer and Xu (2010), due to major costs resulting from default to all involved parties, such as mortgage lenders, mortgagors, investors of mortgage backed securities (MBS) and the guarantors of MBS, assessment and management of the default risk is a major concern for financial institutions, and policy makers. As a result, there exists a rich literature on modeling mortgage default risk; see for example, Quercia and Stegman (1992) and Leece (2004). An important 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 if the market value of the house exceeds the 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, Keenan, III, and Epperson (1990). Thus, under the option theoretic approach, other factors, such as the transaction costs, borrower 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 and evidence against the validity of the assumption has been presented by many authors. Furthermore, as pointed out by Soyer and Xu (2010), 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 risks 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. The competing risks models have been considered by many such as Deng and Order (2000); Deng, Quigley, and Order (1996), Deng (1997), and Calhoun and Deng (2002). These can be considered as the competing risks versions of proportional hazards and multinomial logit models. The competing risks version of the PHM suggested by Deng (1997) involves evaluating hazard rates under the prepayment and default options. The author refers to these as prepayment and default risks. The competing risks approach is found to be useful in explaining the prepayment and default behaviors and improving the prediction of mortgage terminations. Application of these models to commercial mortgages can be found in Ciochetti, Deng, Gao, and Yao (2002) and in the more recent work by Deng and Haghani (2018).

It is important to note that these class of models, focusing on assessment of time to default, differ from the classification type approaches that are typically used to assess whether a loan

* Corresponding author.
E-mail addresses: bhattaca@tcd.ie (A. Bhattacharya), Simon.Wilson@tcd.ie (S.P. Wilson), soyer@email.gwu.edu (R. Soyer).

https://doi.org/10.1016/j.ejor.2018.10.047
0377-2217/

defaults or not. A recent review by Lessmann, Baesens, Seow, and Thomas (2015) discuss classification methods and algorithms that are used for credit scoring. An empirical comparison of classification algorithms for prediction of mortgage defaults can be found in Fitzpatrick and Mues (2016) where authors consider standard methods such as logistic regression as well as decision tree-based approaches such as random forests. Liu, Hua, and Lim (2015) note the potential limitations of classification models in dealing with censored data and propose hierarchical mixture models as an extension of the work of Tong, Mues, and Thomas (2012).

Most of the above models use classical methods for estimation and as a result they do not provide probabilistic inferences. Some exceptions to these are the Bayesian work by Popova, Popova, and George (2008) who proposed Bayesian methods for forecasting mortgage prepayment rates, Soyer and Xu (2010) who considered Bayesian mixtures of proportional hazards models for describing time to default and Kiefer (2010) who proposed an Bayesian approach for default estimation using expert information. More recently, Bayesian time series models have been considered in Aktekin, Soyer, and Xu (2013) and Lee, Rösch, and Scheule (2016). Bayesian mixture and segmentation models have been considered in Galloway, Johnson, and Shemyakin (2017) and Bayesian mixture cure models are discussed by Liu et al. (2015). Our approach differs from the previous in that we consider Bayesian competing risk proportional hazards models and in so doing we use both default and prepayment data. Bayesian analysis of competing risks models has been considered by Sun and Berger (1993) in reliability analysis and semiparametric Bayesian proportional hazards competing risk models have been introduced by Gelfand and Mallick (1995) in survival analysis. Our work differ from these both in terms of the application and the specific approach taken. Furthermore, our focus is on time to default/prepayment since assessment of default time is important for financial institutions who offer loan modification and loss mitigation programs which are available in US as well as in Europe; see Olrich (2006) and Andritzky (2014).

In this paper we consider modeling duration of single-home mortgages. In doing so, we model default and prepayment probabilities simultaneously using competing risks proportional hazards models. We include both individual and aggregate level covariates in our model. We adopt the Bayesian viewpoint in the analysis and develop posterior and predictive inferences by using Markov chain Monte Carlo (MCMC) methods. In addition to providing a formalism to incorporate prior opinion into the analysis, the Bayesian approach enables us to describe all our inferences probabilistically and provides additional insights from the analysis. In what follows, we first introduce the competing risks proportional hazards models in Section 2. The Bayesian inference is presented in Section 3 where posterior and predictive analyses are developed. In Section 4 we illustrate implementation of our model and Bayesian methods using simulated data. Concluding remarks follow in Section 5.

2. Competing risk proportional hazards model

To introduce some notation let L denote the mortgage lifetime and T_M denote the maturity date of the mortgage loan. Note that if a mortgage loan is not defaulted or prepaid then $L = T_M$. If we let T_D and T_P denote time to default and time to prepayment for a mortgage loan, respectively, then $L = T_M$ if $(T_D > T_M)$ and $(T_P > T_M)$. Fig. 1 illustrates the relationship between T_M, T_D and T_P. If both T_D and T_P are larger than T_M then the mortgage will be paid on time. For a given mortgage loan it is of interest to infer events of "full payment", default and prepayment. In other words, we are interested in computing probability statements such as $P(T_D > T_M, T_P > T_M)$, $P(T_D < T_P \,|\, T_D < T_M)$ or $P(L > t \,|\, L < T_M)$.

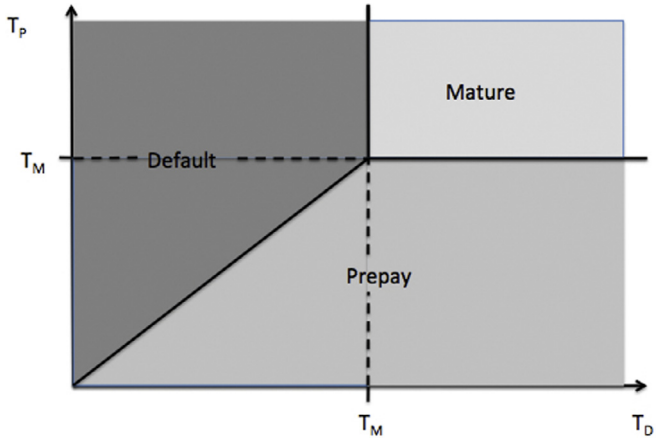

Fig. 1. Competing risk representation of a mortgage that can default, prepay or mature at time T_M.

In view of the above, we can write

$$L = min(T_D, T_P, T_M),$$

where both T_D and T_P are random variables. We will model T_D and T_P separately as proportional hazards models (PHMs) as in Cox (1972). We denote the hazard (failure) rate for default and for prepayment as $\lambda_D(t)$ and $\lambda_P(t)$, respectively. We will refer to $\lambda_D(t)$ as the *default rate* and to $\lambda_P(t)$ as the *prepayment rate*. We model the default rate as

$$\lambda_D(t \,|\, X_D(t)) = r_D(t \,|\, \psi_D)\, \exp(\theta_D' X_D(t)), \tag{1}$$

where $r_D(t|\,\psi)$ is the baseline default rate, ψ_D is vector of parameters, $X_D(t)$ is a vector of time dependent covariates specific to default mortgages and θ_D is a vector of regression parameters. Similarly, the prepayment rate is modeled as

$$\lambda_P(t \,|\, X_P(t)) = r_P(t \,|\, \psi_P)\, \exp(\theta_P' X_P(t)). \tag{2}$$

Note that for ease of notation, we use the same notation for all covariates, i.e. $X_D(t) \equiv X_P(t) \equiv X(t)$.

We assume that default and prepayment are "competing risks", so that we only observe the first of them to occur. The observation of one at time t implies that the other is right-censored at t. We assume T_D and T_P to be independent, conditional on the baseline rate, parameters and set of covariates. Thus, the joint survival function of T_D and T_P is given by

$$\begin{aligned}
&\mathbb{P}(T_D > t_D, T_P > t_P \,|\, r_D(), r_S(), \theta_D, \theta_P, X^*) \\
&\quad = \mathbb{P}(T_D > t_D \,|\, r_D(), \theta_D, X^*)\mathbb{P}(T_P > t_P \,|\, r_S(), \theta_P, X^*) \\
&\quad = \exp\left(- \int_0^{t_D} \lambda_D(w)\, dw \; - \int_0^{t_P} \lambda_P(w)\, dw \right),
\end{aligned}$$

where $X^* = \{X(w) \,|\, 0 \le w \le \max(t_D, t_P)\}$. This standard assumption of conditional independence of the time to each risk occurring means that, if the model parameter values were known in addition to the value of the covariates, then the occurrence of one of the risks does not change the distribution of the time to the other. This assumption allows for a considerable simplification in the computation of the inference procedure and yet does still permit some dependence between the times because they share common covariates.

An active mortgage lifetime observed as t implies that neither a default nor a prepayment occurs by time t, that is, both default and prepayment are right-censored at t. This includes the event that the mortgage matures at time T.

Empirical evidence suggests that the default rate is non-monotonic. As discussed by Soyer and Xu (2010), it is reasonable to expect that the default rate is first increasing and then decreasing.

A lifetime model having such hazard rate behavior is the lognormal model, other alternatives being generalised Gamma or log-logistic distribution which can also be entertained in our framework. Thus, we assume that baseline time to default T_D follows a lognormal model with probability density function

$$p(t_D \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2} t_D} \exp\left(-\frac{1}{2\sigma^2}(\log(t_D) - \mu)^2\right), \; t > 0.$$

Since it is not unreasonable to expect a similar behavior in the prepayment rate, we also assume that the baseline distribution of T_P is also lognormal. Thus, the baseline model for T_D will be lognormal with parameters μ_D and σ_D^2, and for prepayment with parameters μ_P and σ_P^2.

The failure rate of the lognormal distribution can be written in terms of the standard Gaussian distribution function Φ. In fact the failure rates for T_D and T_P then take the form:

$$\lambda_D(t \mid X_D(t)) = \frac{(2\pi\sigma_D^2)^{-1/2} t^{-1} \exp\left(-0.5(\log(t) - \mu_D)^2/\sigma_D^2\right)}{1 - \Phi((\log(t) - \mu_D)/\sigma_D)} \exp\left(\theta_D' X_D(t)\right) \tag{3}$$

and

$$\lambda_P(t \mid X_P(t)) = \frac{(2\pi\sigma_P^2)^{-1/2} t^{-1} \exp\left(-0.5(\log(t) - \mu_P)^2/\sigma_P^2\right)}{1 - \Phi((\log(t) - \mu_P)/\sigma_P)} \exp\left(\theta_P' X_P(t)\right); \tag{4}$$

see the Appendix for details on derivation of 3 and 4.

3. Bayesian analysis of the competing risk PHM

We assume that data on N mortgages are available. From these, n_D have defaulted, n_P have prepayed and $n_C = N - n_D - n_P$ are still active, including those that have matured successfully. The N mortgages are indexed $i = 1, \ldots, n_D$ for the defaulted mortgages, $i = n_D + 1, \ldots, n_D + n_P$ for prepaid and $i = n_D + n_P + 1, \ldots, N$ for active. Let $\boldsymbol{t_D} = \{t_1^D, \ldots, t_{n_D}^D\}$ be the times of default and $\boldsymbol{t_P} = \{t_{n_D+1}^P, \ldots, t_{n_D+n_P}^P\}$ be the times of prepayment. For the n_C mortgages that are still active, let $\boldsymbol{t_C} = \{t_{n_D+n_P+1}^C, \ldots, t_N^C\}$ be the times since the initiation of mortgages; $t_i^C = T_M$ for those that have matured.

Also observed are the covariates. Let $\boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{im}(t))$ be the vector of covariates for mortgage i at time t. Some of these are common covariates e.g. interest rates, while others are mortgage specific e.g. mortgage size or credit score. We assume that they are observed at a known set of times $\tau_1 < \tau_2 < \cdots < \tau_m$ and that they are piecewise constant on intervals for which these times are the mid-points. Hence

$$\boldsymbol{X}_i(t) = \boldsymbol{X}_i(\tau_j), \; s_{j-1} < t \le s_j, \tag{5}$$

for $j = 1, \ldots, m$, where $s_0 = 0$, $s_j = 0.5(\tau_j + \tau_{j+1})$ for $j = 1, \ldots, m-1$ and $s_m = \infty$. We let $\boldsymbol{X}_i = \{\boldsymbol{X}_i(\tau_1), \ldots, \boldsymbol{X}_i(\tau_m)\}$ be the observed covariates for mortgage i and $\boldsymbol{\mathcal{X}} = \{\boldsymbol{X}_i(\tau_k) \mid i = 1, \ldots, N; \; k = 1, \ldots, m\}$ be the set of all observed covariates. We would like to reiterate here that the set of covariates could be specific to mortgage type (default/prepay).

The unknown quantities in this model are the regression parameters θ_D and θ_P, and the baseline failure rate parameters $\psi = (\mu_D, \sigma_D^2, \mu_P, \sigma_P^2)$. The required posterior distribution is therefore:

$$p(\theta_D, \theta_P, \psi \mid \boldsymbol{t_D}, \boldsymbol{t_P}, \boldsymbol{t_C}, \boldsymbol{\mathcal{X}}) \propto p(\boldsymbol{t_D}, \boldsymbol{t_P}, \boldsymbol{t_C} \mid \theta_D, \theta_P, \psi, \boldsymbol{\mathcal{X}}) \; p(\theta_D) p(\theta_P) p(\psi). \tag{6}$$

For the likelihood term $P(\boldsymbol{t_D}, \boldsymbol{t_P}, \boldsymbol{t_C} \mid \theta_D, \theta_P, \psi, \boldsymbol{\mathcal{X}})$, we assume observations are conditionally independent, given the parameters. From the competing risks assumption, an observation t_i^D is an exact observation of T_D and a right-censored observation of T_P; it is vice versa for t_i^P. Finally, t_i^C is a right-censored observation of both T_D and T_P. Hence:

$$\begin{aligned}
&p(\boldsymbol{t_D}, \boldsymbol{t_P}, \boldsymbol{t_C} \mid \theta_D, \theta_P, \psi, \boldsymbol{\mathcal{X}}) \\
&= \left(\prod_{i=1}^{n_D} p(t_i^D \mid \theta_D, \boldsymbol{X}_i) \, P(T_P > t_i^D \mid \theta_P, \boldsymbol{X}_i)\right) \\
&\quad \times \left(\prod_{i=n_D+1}^{n_D+n_P} p(t_i^P \mid \theta_P, \boldsymbol{X}_i) \, P(T_D > t_i^P \mid \theta_D, \boldsymbol{X}_i)\right) \\
&\quad \times \left(\prod_{i=n_D+n_P+1}^{N} P(T_D > t_i^C \mid \theta_D, \boldsymbol{X}_i) \, P(T_P > t_i^C \mid \theta_P, \boldsymbol{X}_i)\right) \\
&= \left(\prod_{i=1}^{n_D} \lambda_D(t_i^D \mid \boldsymbol{X}_i(t_i^D)) \exp \right. \\
&\quad \left. \times \left(-\int_0^{t_i^D} \lambda_D(w \mid \boldsymbol{X}_i(w)) + \lambda_P(w \mid \boldsymbol{X}_i(w)) \, dw\right)\right) \\
&\quad \times \left(\prod_{i=n_D+1}^{n_D+n_P} \lambda_P(t_i^P \mid \boldsymbol{X}_i(t_i^P)) \exp \right. \\
&\quad \left. \times \left(-\int_0^{t_i^P} \lambda_D(w \mid \boldsymbol{X}_i(w)) + \lambda_P(w \mid \boldsymbol{X}_i(w)) \, dw\right)\right) \\
&\quad \times \left(\prod_{i=n_D+n_P+1}^{N} \exp\left(-\int_0^{t_i^C} \lambda_D(w \mid \boldsymbol{X}_i(w)) + \lambda_P(w \mid \boldsymbol{X}_i(w)) \, dw\right)\right)
\end{aligned} \tag{7}$$

where $\lambda_D(t \mid \boldsymbol{X}_i(t))$ and $\lambda_P(t \mid \boldsymbol{X}_i(t))$ are given by Eqs. (3) and (4), $\boldsymbol{X}_i(t)$ is given in Eq. (5) and a formula for the integrals is given in Eq. A.4 of the Appendix. The formula for the integrals is more complex for time varying covariates, as noted in Cox and Oakes (1984).

An independent zero-mean normal prior is assumed for each component of θ_D and θ_P, as well as μ_D and μ_P. For σ_D and σ_P, since no prior provides us with known full conditionals, an exponential prior is assumed. It is noted that a lognormal or gamma prior could also have been chosen.

The model above is such that the parameters are not identifiable without further assumptions. For a Bayesian analysis, such as ours, that means whether the data can inform well enough about all the parameters. Identifiability issues can be overcome via specific prior specification or model dimension reduction; see Gelfand and Mallick (1995). For example, if posterior distributions of parameters are different from their priors, the problem of identifiability is resolved. For those parameters where the priors and the posteriors are the same, the problem remains and can be solved using an improper prior (if it results in a proper posterior) or reducing dimension. We will see in the results that for the default model, identifiability is present. It is attributable to the low number of default mortgages under which parameter learning becomes very difficult; no such issue exists for the parameters under the prepaid model.

An MCMC procedure, based on the Metropolis within Gibb's sampler (Tierney, 1994), has been implemented to sample from $p(\theta_D, \theta_P, \psi \mid \boldsymbol{t_D}, \boldsymbol{t_P}, \boldsymbol{t_C}, \boldsymbol{\mathcal{X}})$. The covariate coefficient vectors θ_D and θ_P are updated as blocks from their full conditional distributions with a Gaussian random walk proposal, while each component of ψ is updated separately. The Appendix contains the details of the algorithm.

The MCMC output is a set of samples of all the unknowns from the posterior distribution. Let the number of samples be G, and

let $\theta_D^{(g)}$, $\theta_P^{(g)}$ and $\psi^{(g)}$ denote the g^{th} samples of ψ, θ_D and θ_P, respectively.

The MCMC output can be used to compute many quantities of interest. With the posterior samples, one can compute for a mortgage with a known set of covariates $\boldsymbol{X} = \{\boldsymbol{X}(w) \mid w \geq 0\}$. The posterior predictive reliability function of the time to default is approximated by

$$P(T_D > t \mid \boldsymbol{t_D}, \boldsymbol{t_P}, \boldsymbol{t_C}, \boldsymbol{\mathcal{X}}, \boldsymbol{X}) \approx \frac{1}{G}\sum_{g=1}^{G} \exp\left(-\int_0^t \lambda_D^{(g)}(w)\, dw\right), \quad (8)$$

and the time to prepayment is approximated by

$$P(T_P > t \mid \boldsymbol{t_D}, \boldsymbol{t_P}, \boldsymbol{t_C}, \boldsymbol{\mathcal{X}}, \boldsymbol{X}) \approx \frac{1}{G}\sum_{g=1}^{G} \exp\left(-\int_0^t \lambda_P^{(g)}(w)\, dw\right), \quad (9)$$

where $\lambda_D^{(g)}(w) = r_D^{(g)}(w \mid \psi)\exp(\theta_D^{(g)\prime}\boldsymbol{X}(w))$ and $\lambda_P^{(g)}(w) = r_P^{(g)}(w \mid \psi)\exp(\theta_P^{(g)\prime}\boldsymbol{X}(w))$, the values of $r_D^{(g)}(w \mid \psi)$ and $r_P^{(g)}(w \mid \psi)$ are given by Eq. (A.1), using the parameter values in $\psi^{(g)}$, and a formula for the integrals is given by Eq. (A.4) of the Appendix.

Eqs. (8) and (9) allows us to determine, by simulation, the probability that a mortgage will default, prepay or mature with a given set of covariates $\boldsymbol{X}$. The inverse distribution method can be used to simulate independently many values pairs (t_D, t_P) from these reliability functions e.g. for t_D, generate a random number u and then solve $u = P(T_D > t \mid \boldsymbol{t_D}, \boldsymbol{t_P}, \boldsymbol{t_C}, \boldsymbol{\mathcal{X}}, \boldsymbol{X})$ for t, an easy numerical exercise. This further means we can compute predictive densities $P(T_D \mid \boldsymbol{t_D}, \boldsymbol{t_P}, \boldsymbol{t_C}, \boldsymbol{\mathcal{X}}, \boldsymbol{X})$ and $P(T_P \mid \boldsymbol{t_D}, \boldsymbol{t_P}, \boldsymbol{t_C}, \boldsymbol{\mathcal{X}}, \boldsymbol{X})$ for each mortgage as well. Furthermore to this, the probabilities that a loan defaults, prepays or matures are approximated by the proportion of simulated pairs (t_D, t_P) that lie in their respective regions as defined in Fig. 1:

Defaults $\Leftrightarrow t_D < T_M$ and $t_D < t_P$;

Prepays $\Leftrightarrow t_P < T_M$ and $t_P < t_D$;

Matures $\Leftrightarrow t_D \geq T_M$ and $t_P \geq T_M$.

Since the marginal density is

$$f(t) = \lambda(t)\, R(t),$$

where $\lambda(t)$ is failure rate and $R(t)$ is reliability function as defined in Eqs. (8) and (9), so the posterior density function for T_D is approximated by

$$f_D(t) \approx \frac{1}{G}\sum_{g=1}^{G} \lambda_D^{(g)}(t)\exp\left(-\int_0^t \lambda_D^{(g)}(w)\, dw\right).$$

A similar expression holds for $f_P(t)$.

4. The Freddie Mac single family loan dataset

The Federal Home Loan Mortgage Corporation (FHLMC), known as Freddie Mac, is a public company that is sponsored by the United States government. It was formed in 1970 to expand the secondary market for mortgages in the US. It has provided a dataset about single family loan-level credit performance data on a portion of fully amortizing fixed-rate mortgages that the company purchased or guaranteed. The dataset contains information about approximately 21.5 million fixed-rate mortgages that originated between January 1, 1999, and December 31, 2014. The dataset can be downloaded from the Freddie Mac website and is organised as two files for each quarter:

1. the origination data file that contains data concerning the set up of the loan;
2. the monthly performance data file that contains the monthly performance of each loan e.g. amount repaid, the outstanding principal, whether it is in default, etc.

There is also a smaller sample data set that contains a simple random sample of 50,000 loans selected from each year and a proportionate number of loans from subsequent years (the actual definition is 50,000 loans selected from each full vintage year and a proportionate number of loans from each partial vintage year of the full single family loan-level data set). The sample data set also has an origination and monthly performance file for each year

Some processing of the raw data was needed to transform it into a format that can be be analysed by this model. Each loan was tracked through the data to categorize it as active, defaulted or prepaid. Since loans in the dataset originated in 1999 and are valid for 30 years, there were no loans classified as mature and so this category could be ignored.

The data is highly unbalanced. This is well noted in the literature, with default rates typically staying around 1% to 2.5% for conventional mortgage loans whereas for subprime loans default rates rise over 14% in some years, see for example Danis and Pennington-Cross (2008).

4.1. Loan categorization

Four fields in the data were used to categorize each loan as default, prepay or active, and to define the observed time:

- `zero_balance` defines whether a particular loan's balance has reduced to 0 or not, and has the following codes:
  01 Prepaid or Matured (voluntary payoff);
  03 Foreclosure Alternative Group (Short Sale, Third Party Sale, Charge Off or Note Sale);
  06 Repurchase prior to Property Disposition;
  09 REO Disposition; and
  empty Not Applicable.
- `delinquency` provides a value corresponding to the number of days the borrower has not paid the loan, according to the due date of last paid installment, or if a loan is acquired by REO, coded as:
  0 Current, or less than 30 days past due;
  1 30–59 days delinquent;
  2 50–89 days delinquent;
  3 90–119 days delinquent, etc.;
  R REO acquisition;
  empty Unavailable.
- `reporting_date` is the month that the observation is made in.
- `months_remain` is the number of months until the legal maturity of the loan.

Then the loan status was defined as:

- *Prepaid* if there exists a month where `zero_balance` = 01 AND `repurchase` = "N". In this case, the prepaid time t_P is the time from loan origination to the `reporting_date` where this first happens.
- *Default* if there exists a month where `zero_balance` = 03, 06 or 09. In this case, the default time t_D is the time from loan origination to the `reporting_date` where this first happens.
- *Active* if the loan could not be classified as Prepaid or Active AND the latest `reporting_date` corresponding to the loan is later than 01/01/2014 AND `zero_balance` is empty at that latest date AND `delinquency` is not equal to R at that latest date. The active time is the time from loan origination to the `reporting_date` where this happens.

These definitions are not exhaustive; there are loans in the dataset that are discontinued without any clear information and such loans have been excluded from our analysis.

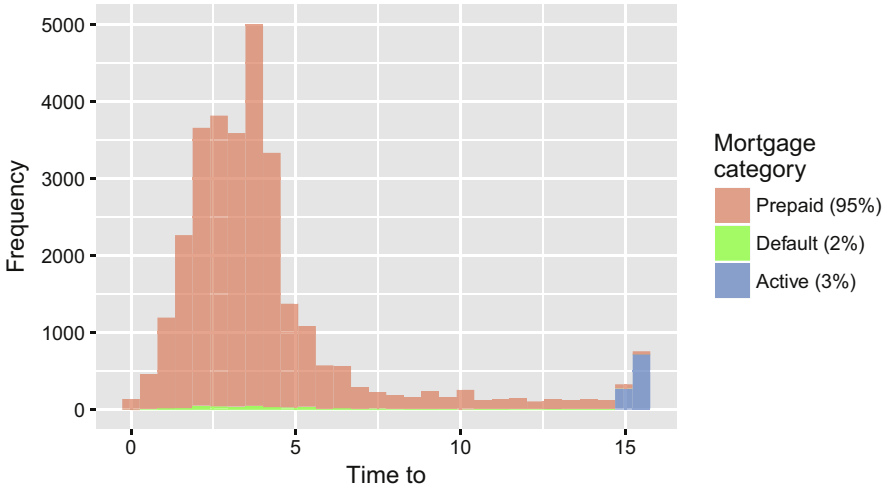

Fig. 2. Histogram of time to default, prepaid and active times for each category.

4.2. Covariates

The following covariates (fixed term) are available in the dataset: credit score, mortgage insurance percentage (MI), number of units, combined loan-to-value (CLTV), debt-to-income (DTI), unpaid principal balance (UPB), original interest rate, number of borrowers, first homebuyer, occupancy status, property type, property state (state in which property resides) and current interest rate. Original interest rate refers to the rate at loan initiation while current interest rate contains monthly interest rates since loan start time. So the latter is the only variable whose values changes with time. Out of the rest, first homebuyer, occupancy status, property type and property state are categorical variables and have been converted to indicator variables. The covariate property state has been re-categorized into judicial or non-judicial state (renamed *Foreclosure state*), where in a judicial state, the lender needs to go through the court system for the foreclosure process. For the rest of the categorical variables, some of categories were of low frequency. For example, there are 6 categories in variable `property_type`, of which 81% were `single family home` and some categories like `leasehold` accounting for as low as 0.0003%. It was decided to group categories with extremely low frequencies for all the categorical variables. All the quantitative variables have been standardized. Furthermore strong correlation have been found between `mortgage insurance percentage` and `combined loan-to-value`, and between original and current `interest rates` which led us to drop the latter in both the cases. Since `current interest rate` has been dropped we do not have to work with any time dependent covariate.

5. Analysis of the data

The data set comprised of 672208 mortgages originating in the year 1999. This data set is extremely unbalanced with 95% of the mortgages being `prepaid`, 3% being `active` and the only about 1.6% belonging to `default` category. This huge imbalance is evident in Fig. 2.

Both θ_D and θ_P had Normal prior with zero mean and standard deviation 100. The mean parameters μ_D and μ_P have zero mean Normal priors with standard deviation 10, while an exponential prior with mean 100 was assumed for the standard deviation parameters σ_D and σ_P. The starting value of each of the parameter chains were randomly selected using Normal and inverse gamma distributions for mean and standard deviation parameters respectively. The standard deviation for the proposal distributions, for example $s^2_{\theta,D}$ or $s^2_{\mu,D}$ have also been generated randomly from inverse gamma distributions to provide more diversity in the chains. The scale parameter (a) used in proposal for σ_D and σ_P was generated from uniform distribution.

Table 1
Summary of the marginal posterior distributions of the baseline default and prepay rates.

Parameter	Median	95% Prob. Interval
μ_D	2.817	(2.631, 3.077)
σ_D	0.963	(0.916, 1.028)
μ_P	1.578	(1.566, 1.591)
σ_P	0.717	(0.713, 0.721)

Rcpp (Eddelbuettel & François, 2011) has been used to construct the MCMC algorithm. This has greatly improved the speed of the algorithm given that the data set is extremely large. The MCMC procedure was run in 50 chains for 75,000 iterations each. We set MCMC burn-in at 60000 and thinned the remaining by selecting every 50*th* sample. Trace plots, provided in the Appendix, for all the parameters show good mixing for all the covariates implying convergence. We provide the density plot constructed by combining the thinned chains for a subset of covariates in Fig. 3. The skew in the density of μ_D is caused by the slow convergence of a single chain. We assume that this single chain suffers from slow convergence due to the problem of identifiability in proportional hazards models. The problem is further enhanced by the fact that the number of default mortgages is very low, hence making sufficient learning difficult.

Tables 1 and 2 are summaries of the marginal posterior distributions of the model parameters, based on the 15,000 combined samples of the MCMC. We see in table 2 that nearly all the covariates, with the exception of `no. of units`, turn out to be significant. `Credit score`, UPB, `no. of units`, `type of property` and `no. of borrowers` have opposite effects on default and prepay rates. Default rate is found to decrease with `credit score`, UPB etc, as it should be, and prepay rate increases for the same. Other variables, for example, DTI, `mortgage insurance %`, `original interest rate`, `first time homebuyer`, `occupancy status` and `foreclosure state` have the same signs of coefficients for both of default and prepay, which is consistent with other findings from the literature; see for example Deng and Haghani (2018). Thus we can see that for `mortgage insurance %` both default and prepay rate increase, whereas for `first time homebuyer`

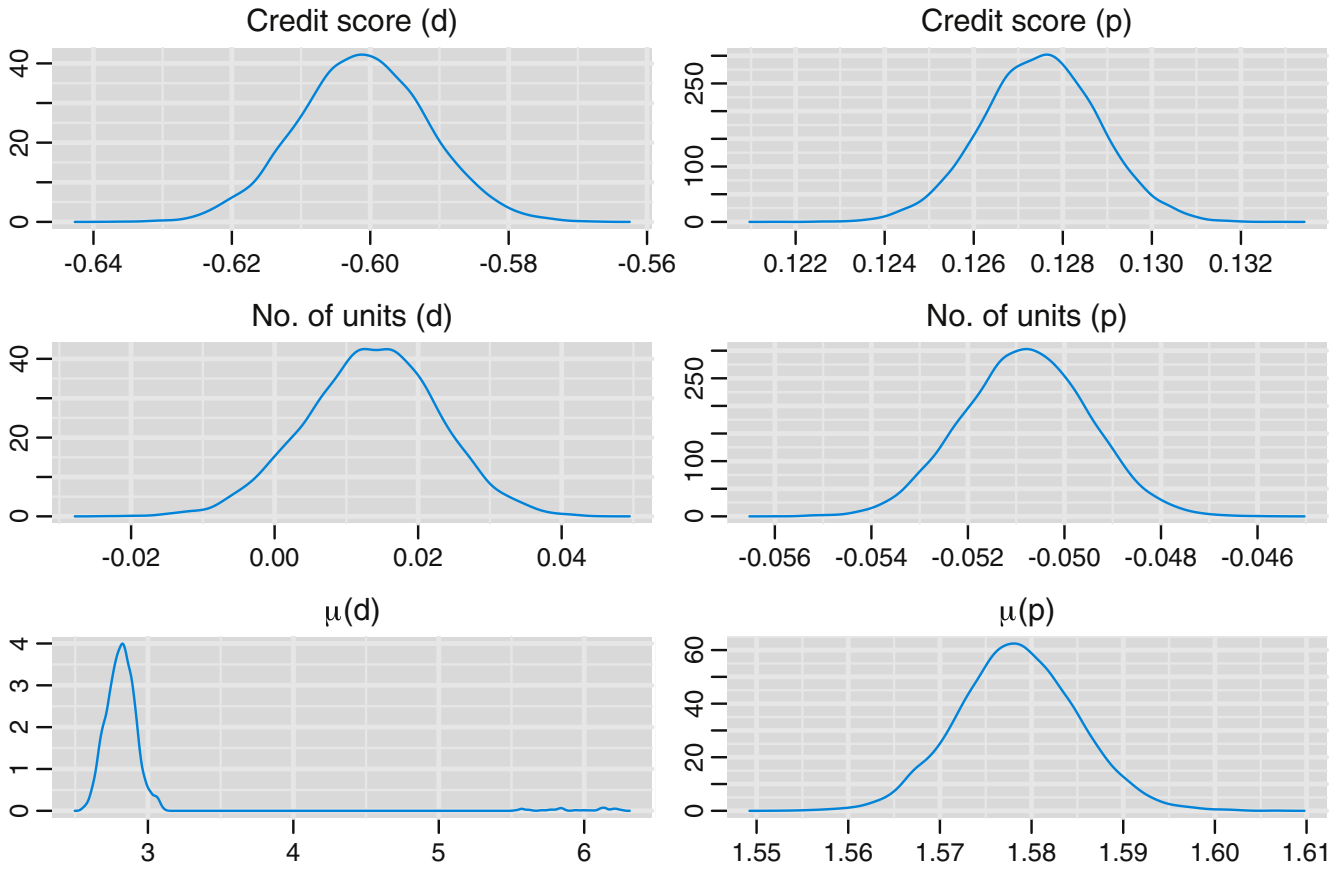

Fig. 3. Density plot of combined samples from merging all the chains for credit score, number of units and μ, both for default (d) and prepaid (p) times. The long tail corresponding to μ(d) can be attributed to a single chain which is slow in converging.

Table 2
Summary of the marginal posterior distributions of θ_D and θ_P. The 95% probability intervals are the 2.5 – 97.5 percentiles of the sampled parameter values. Number of units under default mortgages is the only covariate that can be termed as not-significant, since the CI contain 0.

Covariate	Default		Prepay	
	Median	95% Prob. Interval	Mean	95% Prob. Interval
Credit score	−0.601	(−0.620, −0.583)	0.128	(0.125, 0.130)
Mortgage insurance %	0.395	(0.376, 0.415)	0.068	(0.065, 0.070)
Number of units	0.014	(−0.005, 0.031)	−0.051	(−0.053, −0.048)
Original DTI	0.124	(0.103, 0.146)	0.020	(0.018, 0.023)
UPB	−0.069	(−0.093, −0.046)	0.305	(0.302, 0.307)
Original interest rate	0.412	(0.396, 0.429)	0.376	(0.374, 0.379)
No. of borrowers	−0.296	(−0.316, −0.276)	0.055	(0.052, 0.058)
Intercept	−3.090	(−3.356, −2.694)	0.182	(0.158, 0.207)
First time home-buyer	−0.244	(−0.293, −0.194)	−0.009	(−0.016, −0.003)
Occupancy status	0.460	(0.342, 0.575)	0.249	(0.237, 0.261)
Foreclosure state	−0.110	(−0.149, −0.071)	−0.080	(−0.085, −0.075)
Property type	0.304	(0.249, 0.362)	−0.061	(−0.67, −0.054)

*Covariate *Foreclosure state* refers to whether the state is judicial or non-judicial.

both the rates decrease. Also note that the estimated mean parameter (as also for the standard deviation parameter) of the baseline default rate is substantially greater than that of prepay.

The posterior predictive densities for time to default (or prepay) corresponding to each mortgage can be computed using Eqs. (8) and (9) (see discussion following these two equations). Fig. 4 shows predictive posterior distribution of time to default for 3 randomly chosen mortgages in the top panel. A similar plot corresponding to time to prepay has been provided in the lower panel. The posterior predictive densities for time to default are generally found to be flat, while those for prepay have variable shapes. Note that the area under the densities do not necessarily sum to 1 in these plots since we have truncated them at 2029, the maturity date of the mortgages.

The mortgages and their covariate values are provided in Table 3.

6. Model assessment

The suitability of the model is assessed by deriving, for each loan in the data:

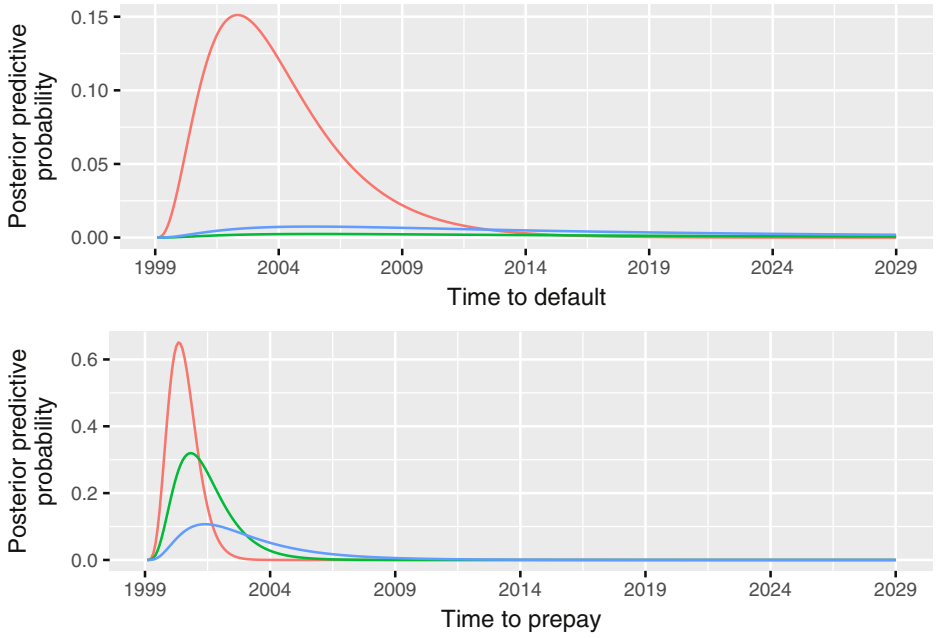

Fig. 4. The two panels provide posterior predictive distributions of time to default and prepay, respectively. The flat posterior predictive distribution of time to default is very common in almost all mortgages, while for prepay the shape of the distributions are quite varying. Note that some of the distributions are truncated at 2029, the year the mortgages end.

Table 3
The values of the covariates for the 6 mortgages that have been used for computing the posterior predictive densities in Fig. 4 is provided here. Abbreviations used are: "Owner" - "Owner occupied", "Non-Jud"/"Jud" - "Non-Judicial"/"Judicial" and "SF" - "Single family".

Covariate	Default			Prepay		
Mortgage number	1	2	3	1	2	3
Credit score	724	541	750	787	668	619
Mortgage insurance %	12	30	0	0	0	30
Number of units	1	1	1	1	1	1
Original DTI	16	27	23	39	34	44
UPB	73,000	112,000	83,000	37,000	312,000	204,000
Original interest rate	6.875	10	8	6.875	7	9.625
No. of borrowers	2	2	2	1	2	2
First time home-buyer	No	No	No	No	No	No
Occupancy status	Owner	Owner	Owner	Owner	Owner	Owner
Foreclosure state	Non-Jud	Jud	Non-Jud	Non-Jud	Non-Jud	Non-Jud
Property type	SF	SF	SF	SF	SF	SF

- The probabilities that the loan defaults, prepays or remains active up to the end of the data, following the method in Section 3, which can be compared to the actual outcome;
- If the mortgage defaulted then the predicted reliability function of the default time can also be computed from Eq. (8), and hence the quantile of the observed time. A standardised residual can also be computed e.g. $(t_D - E(t_D))/sd(t_D)$, where t_D is the observed default time, $E(t_D)$ and $sd(t_D)$ are the mean and standard deviation of the posterior default time, derived from the predicted reliability function.
- Similarly, if the mortgage was prepaid then the predicted reliability function of the prepay time can be computed from Eq. (9). The quantile of the observed prepay time and a standardised residual can be derived.

Active loans are right-censored observations of both the default and prepay times. The competing hazards model implies that default times are also right-censored observations of a prepay time, and vice versa.

We assessed the fitted model on the sample data set in the year 1999, which has 30,755 mortgages. Fig. 5 shows a box plot of standardised residuals (as explained above) for all the default and the prepaid mortgages and is found to be centered around 0. If we isolate the defaulted mortgages we find that the corresponding residuals are biased away from 0. Identifying mortgages that defaulted is found to be difficult form that data we have since they constitute less than 2% of the whole set.

Separate box plots of standardised residuals for mortgages show a good fit for prepay, where residuals are slightly biased away from 0. However, for default mortgages almost all residuals are negative, which shows that the estimated mean time to default is higher than what was observed. We think that this can be largely attributed to huge contrast in proportion of each type of mortgages in the data set.

The central 95% posterior prediction interval for the time to prepayment or default showed good coverage properties for what was observed, at 93% e.g. 93% of observed prepayment or default times lie within their 95% posterior prediction interval. When split by the type of event, the observed prepaid mortgages had 95% coverage but the default mortgages had rather poor coverage at only 50%. Although the objective of our proposed model is not classification, we have compared our coverage percentages with classification performances of machine learning (ML) methods such as

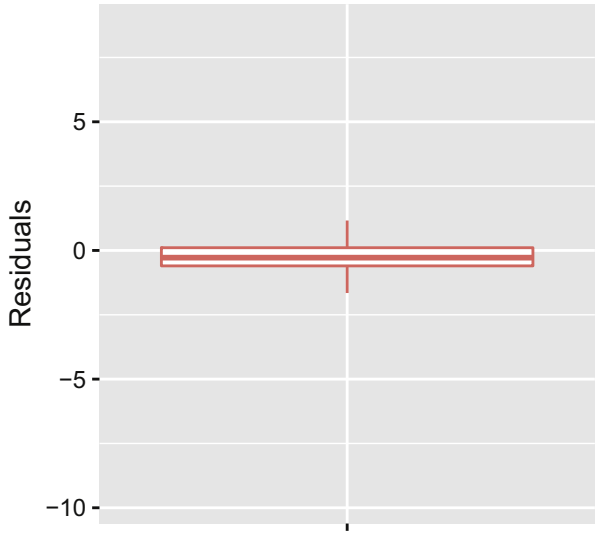

Fig. 5. Residuals of all the default and prepay mortgages combined. Note the slight bias below 0 which is caused by the default mortgages.

random forest (RF) and logistic regression with lasso (see Appendix for results). We found that RF classified defaulted mortgages correctly 27.8% of the time which is lower than our coverage probability of 50%. The prepaid mortgages were classified correctly 100% of the time by the RF compared to our coverage probability of 95%. Logistic regression with lasso performed even poorly than RF for default mortgages at 0.0034%, while prepaid success prediction was at par. In other words, the proposed Bayesian model, in addition to providing more insights, has given better results in predicting defaults and almost as good results in predicting prepayments compared to RF.

The analysis suggests that we appear to be under-estimating the uncertainty in the default times. This can be explained by the imbalance in the data, with poor learning because of a much smaller number of observations of default, or by missing important covariates such as the location of the mortgaged property, local conditions and regulations as well as other borrower characteristics; see for example Goodman and Smith (2010). Also, as pointed out by a reviewer, the fact that default and prepayment behaviours are motivated by different factors may contribute to this. The model can possibly be improved by using behaviour specific covariates. As for this idea, while our method and logistic regression allows a different set of borrower covariates for default/prepay behaviour, RF provides a single unified set.

The contrast between default and prepaid mortgages is also evident when we calculated the predicted reliability function at the observed default or prepayment time. The median predicted reliability function for default mortgages is found to be 0.976 and (2.5, 97.5) quantiles being (0.740, 0.999), while those for prepaid are 0.524(0.041, 0.981). This confirms that in general the estimation of prepayment time performs well but that we over-estimate the default time. The box plots in Fig. 7 demonstrates the range of posterior reliability corresponding to both default and prepaid mortgages. Reliability is computed at the time to default or prepay.

7. Conclusion and future work

In this paper we have introduced a model for the time to mortgage prepayment or default as a function of mortgage covariates. The proposed competing risks model allows one to take account of the fact that an observation of a mortgage prepay is also a censored observation of a default, and vice versa; hence observation of one does contain information about the other that should be used in inference. Model inference can be done even for quite large data sets, as has been illustrated here for a set of single family loan data from Freddie Mac, where the relative effects of the different covariates on eventual prepayment or default have been quantified. Some difficulties with the inference were encountered, particularly for the defaults that were only a small percentage of the data. In particular, the identifiability issues with this model can cause some convergence issues with the MCMC implementation of the inference.

A natural extension of this work is to consider heterogeneity between mortgages, which would allow one to explore the between-mortgage variability and identify clusters of mortgages with similar properties. A random effects model is a simple way to allow for this e.g. the failure rates for mortgage i are now

$$\lambda_{D,i}(t, \ |X_{D,i}(t)) = r_D(t)\exp(\theta_D^T X_{D,i}(t) + \phi_{D,i}) \text{ and}$$
$$\lambda_{P,i}(t, \ |X_{D,i}(t)) = r_P(t)\exp(\theta_P^T X_{D,i}(t) + \phi_{P,i}),$$

where $\phi_{D,\,i}$ and $\phi_{P,\,i}$ have a zero-mean prior distribution such as a Gaussian with variances that are either known or are also specified by prior distributions. The $\phi_{D,\,i}$ and $\phi_{P,\,i}$ quantify the differences

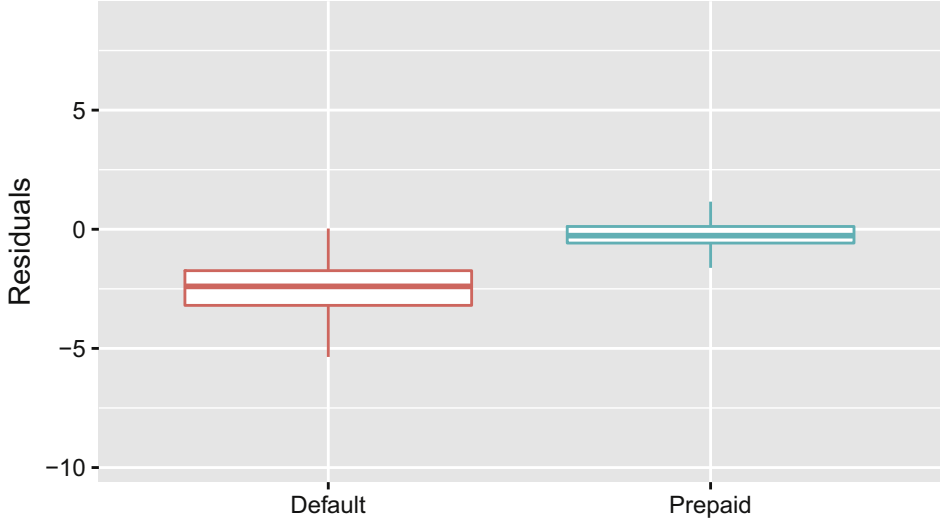

Fig. 6. Separate box plots of residuals corresponding to default and prepay mortgages. The residuals for default mortgages show clear bias below 0.

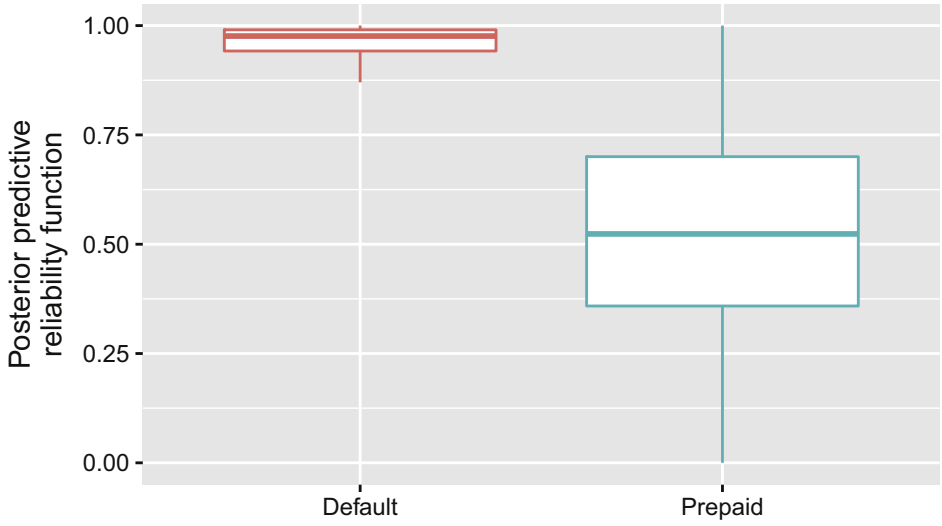

Fig. 7. Box plots of posterior predictive reliability function computed for mortgages at time to default and prepay.

between default and prepay times from the average behaviour that is specified through θ_D and θ_P. In terms of implementing inference for his model, the full conditional distributions of θ_D and θ_P, as needed for the MCMC, remain accessible to the Metropolis algorithm with minor modifications. The full conditional distributions for the $\phi_{D,i}$ and $\phi_{P,i}$ are also accessible; the main difficulty is the very large increase in the number of parameters to be inferred as we have added 2 for every mortgage in the data set, with a resulting slowdown in the computation time.

Acknowledgment

This research was partly supported by Science Foundation Ireland (SFI) under Grant number SFI/12/RC/2289 (The INSIGHT Centre for Data Analytics).

Appendix

Deriving the failure rate of the lognormal distribution

Let T be a lognormally distributed random variable with parameters μ and σ^2 and density function

$$f(t\,|\,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}\,t}\exp\left(-\frac{1}{2\sigma^2}(\log(t)-\mu)^2\right).$$

The failure rate is defined as

$$r(t) = \frac{f(t\,|\,\mu,\sigma^2)}{P(T>t\,|\,\mu,\sigma^2)} = \frac{f(t\,|\,\mu,\sigma^2)}{\int_t^\infty f(s\,|\,\mu,\sigma_D)\,ds}.$$

The lognormal failure rate can be calculated in terms of the normal cdf because T has the property that $\log(T)$ is normally distributed. Therefore

$$\begin{aligned}\int_t^\infty f(s\,|\,\mu,\sigma^2)\,ds &= 1-P(T<t) = 1-P(\log(T)<\log(t))\\ &= 1-\Phi((\log(t)-\mu)/\sigma),\end{aligned}$$

where Φ is the standard normal cdf. Hence

$$r(t) = \frac{(2\pi\sigma^2)^{-1/2}t^{-1}\exp\left(-0.5(\log(t)-\mu)^2/\sigma^2\right)}{1-\Phi((\log(t)-\mu)/\sigma)}. \tag{A.1}$$

Computing the integral of the failure rate function

The integral of the failure rate function appears in the likelihood function. It is assumed that the covariates $\boldsymbol{X}(t)$ vary piecewise constantly on intervals with mid-points $\tau_1<\tau_2<\cdots<\tau_m$. So $\boldsymbol{X}(t)=\boldsymbol{X}(\tau_j)$ for $s_{j-1}<t\le s_j$, with interval end-points $s_0=0$ and $s_j=0.5(\tau_j+\tau_{j+1})$, $j=1,\ldots,m$, with $\tau_{m+1}=\infty$.

Let $m'=\max\{j\,|\,\tau_j<t_D\}$. The integral of the failure rate, needed in the specification of the distribution of T_D, is then:

$$\begin{aligned}\int_0^{t_D}\lambda_D(w\,|\,\boldsymbol{X}_D(w))\,dw = &\sum_{j=1}^{m'}\exp(\theta_D'\boldsymbol{X}_D(\tau_j))\int_{s_{j-1}}^{s_j} r_D(w)\,dw\\ &+\exp(\theta_D'\boldsymbol{X}_D(\tau_j))\int_{s_{m'}}^{t_D} r_D(w)\,dw\end{aligned} \tag{A.2}$$

The integral of the lognormal failure rate can be calculated in a closed form expression, using the fact that $\log(T)$ is Gaussian, and that

$$-\log(P(T>t)) = \int_0^t r(s)\,ds$$

holds for any failure rate, so that:

$$\begin{aligned}\int_{t_a}^{t_b} r(s)\,ds &= \int_0^{t_b} r(s)\,ds - \int_0^{t_a} r(s)\,ds\\ &= -\log(P(T>t_b)) + \log(P(T>t_a))\\ &= -\log(1-\Phi[(\log(t_b)-\mu)/\sigma])\\ &\quad +\log(1-\Phi[(\log(t_a)-\mu)/\sigma]).\end{aligned} \tag{A.3}$$

Substituting Eq. (A.3) into Eq. (A.2) gives:

$$\begin{aligned}&\int_0^{t_D}\lambda_D(w\,|\,\boldsymbol{X}_D(w))\,dw\\ &= \sum_{j=1}^{m'}\exp(\theta_D'\boldsymbol{X}_D(\tau_j))\Big[-\log(1-\Phi[(\log(s_j)-\mu_D)/\sigma_D])\\ &\quad +\log(1-\Phi[(\log(s_{j-1})-\mu_D)/\sigma_D])\Big]\\ &\quad +\exp(\theta_D'\boldsymbol{X}_D(\tau_{m'}))\Big[-\log(1-\Phi[(\log(t_D)-\mu_D)/\sigma_D])\\ &\quad +\log(1-\Phi[(\log(s_{m'})-\mu_D)/\sigma_D])\Big]\end{aligned} \tag{A.4}$$

The integral for T_P, $\int_0^{t_P}\lambda_P(w\,|\,\boldsymbol{X}_P(w))\,dw$ is also given by Eq. (A.4) with t_D, θ_D, μ_D and σ_D replaced by t_P, θ_P, μ_P and σ_P respectively.

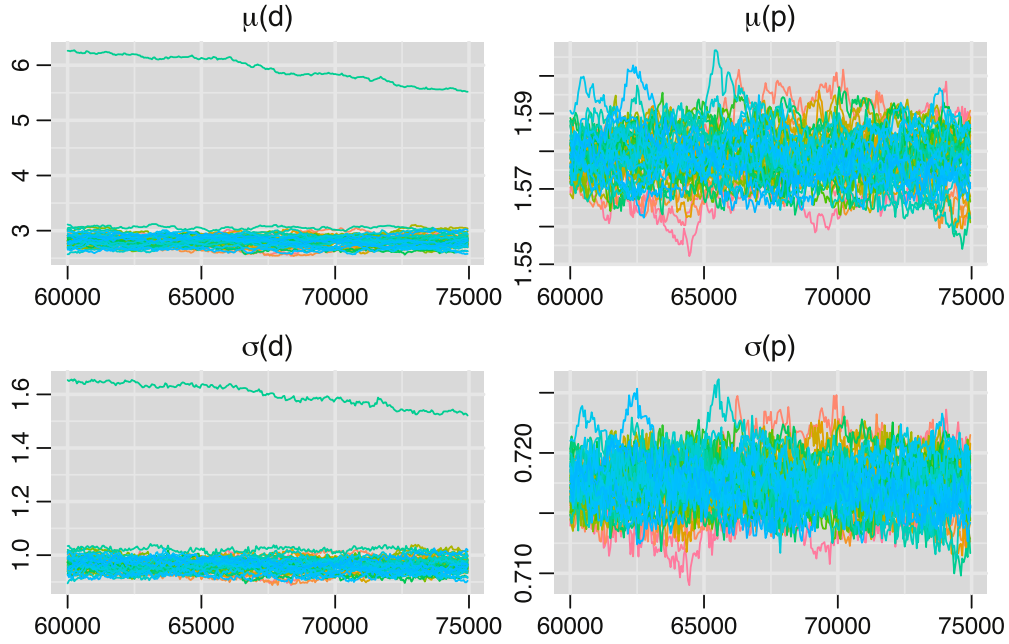

Fig. 8. Trace plot of all distributional parameters. For parameters in default category, the single slow converging chain is visible here as well.

Details of the MCMC Algorithm for the homogeneous model

Sampling of the posterior distribution of Eq. (6), with likelihood given by Eq. (7), is done by a Metropolis within Gibbs algorithm. Each block of parameters are sampled from their full conditional distribution, with those samples obtained through a Metropolis proposal, as follows:

Sample $\boldsymbol{\theta_D}$. From a current value $\boldsymbol{\theta_D}$, a random walk proposal θ_D^* is made from a Gaussian with mean θ_D and variance $s^2_{\theta,D} I_{m\times m}$, where $I_{m\times m}$ is the identity matrix of dimension m and s^2_D is tuned to provide a reasonable acceptance rate. The proposal is accepted with probability

$$\min\left\{1, \frac{p(\boldsymbol{t}_D, \boldsymbol{t}_P, \boldsymbol{t}_C \,|\, \theta_D^*, \theta_P, \psi, \mathcal{X})\, p(\theta_D^*)}{p(\boldsymbol{t}_D, \boldsymbol{t}_P, \boldsymbol{t}_C \,|\, \theta_D, \theta_P, \psi, \mathcal{X})\, p(\theta_D)}\right\}$$
$$= \min\left\{1, \frac{p(\theta_D^*)\prod_{i=1}^{n_D} \lambda_D^*(t_i^D \,|\, \boldsymbol{X}_i(t_i^D))}{p(\theta_D)\prod_{i=1}^{n_D} \lambda_D(t_i^D \,|\, \boldsymbol{X}_i(t_i^D))} \frac{\exp(-A^* - B^* - C^*)}{\exp(-A - B - C)}\right\},$$

where

$$A^* = \sum_{i=1}^{n_D} \int_0^{t_i^D} \lambda_D^*(w \,|\, \boldsymbol{X}_i(w))\, dw,$$

$$B^* = \sum_{i=n_D+1}^{n_D+n_P} \int_0^{t_i^P} \lambda_D^*(w \,|\, \boldsymbol{X}_i(w))\, dw \text{ and}$$

$$C^* = \sum_{i=n_D+n_P+1}^{N} \int_0^{t_i^C} \lambda_D^*(w \,|\, \boldsymbol{X}_i(w))\, dw,$$

and,

$$A = \sum_{i=1}^{n_D} \int_0^{t_i^D} \lambda_D(w \,|\, \boldsymbol{X}_i(w))\, dw,$$

$$B = \sum_{i=n_D+1}^{n_D+n_P} \int_0^{t_i^P} \lambda_D(w \,|\, \boldsymbol{X}_i(w))\, dw \text{ and}$$

$$C = \sum_{i=n_D+n_P+1}^{N} \int_0^{t_i^C} \lambda_D(w \,|\, \boldsymbol{X}_i(w))\, dw.$$

$\lambda_D^*(t \,|\, \boldsymbol{X}(t))$ is given by Eq. (3) with $\theta_D = \theta_D^*$, $\boldsymbol{X}(t)$ is given by Eq. (5) and $\int_0^t \lambda_D(w \,|\, X(w))\, dw$ is given by Eq. (A.4).

Sample $\boldsymbol{\theta_P}$. This is identical to sampling from θ_D, with $\lambda_D(t \,|\, \boldsymbol{X}(t))$ replaced by $\lambda_P(t \,|\, \boldsymbol{X}(t))$ throughout.

Sample $\boldsymbol{\mu_D}$. From a current value μ_D, a random walk proposal μ_D^* is made from a Gaussian with mean μ_D and variance $s^2_{\mu,D}$, where $s^2_{\mu,D}$ is tuned to provide a reasonable acceptance rate. The proposal is accepted with probability

$$\min\left\{1, \frac{p(\boldsymbol{t}_D, \boldsymbol{t}_P, \boldsymbol{t}_C \,|\, \theta_D, \theta_P, \psi^*, \mathcal{X})\, p(\mu_D^*)}{p(\boldsymbol{t}_D, \boldsymbol{t}_P, \boldsymbol{t}_C \,|\, \theta_D, \theta_P, \psi, \mathcal{X})\, p(\mu_D)}\right\}$$
$$= \min\left\{1, \frac{p(\mu_D^*)\prod_{i=1}^{n_D} \lambda_D^*(t_i^D \,|\, \boldsymbol{X}_i(t_i^D))}{p(\mu_D)\prod_{i=1}^{n_D} \lambda_D(t_i^D \,|\, \boldsymbol{X}_i(t_i^D))} \times \frac{\exp(-A^* - B^* - C^*)}{\exp(-A - B - C)}\right\},$$

where A^*, B^*, C^*, A, B and C have already been defined earlier. $\psi^* = (\mu_D^*, \sigma_D^2, \mu_P, \sigma_P^2)$, $\lambda_D^*(t \,|\, \boldsymbol{X}(t))$ is given by Eq. (3) with $\mu_D = \mu_D^*$, $\boldsymbol{X}(t)$ is given by Eq. (5) and $\int_0^t \lambda_D(w \,|\, X(w))\, dw$ is given by Eq. (A.4).

Sample $\boldsymbol{\mu_P}$. This is identical to sampling from μ_D, with $\lambda_D(t \,|\, \boldsymbol{X}(t))$ replaced by $\lambda_P(t \,|\, \boldsymbol{X}(t))$ throughout and $\psi^* = (\mu_D, \sigma_D^2, \mu_P^*, \sigma_P^2)$.

Sample $\boldsymbol{\sigma_D^2}$. From a current value σ_D^2, a proposal $\sigma_D^{2,*}$ is generated from a uniform distribution on the interval $(a\sigma_D^2, \sigma_D^2/a)$, where $a \in (0, 1)$ is tuned to provide a reasonable acceptance rate. The proposal is accepted with probability

$$\min\left\{1, \frac{p(\boldsymbol{t}_D, \boldsymbol{t}_P, \boldsymbol{t}_C \,|\, \theta_D, \theta_P, \psi^*, \mathcal{X})\, p(\sigma_D^{2,*})\, p(\sigma_D^2 \,|\, \sigma_D^{2,*})}{p(\boldsymbol{t}_D, \boldsymbol{t}_P, \boldsymbol{t}_C \,|\, \theta_D, \theta_P, \psi, \mathcal{X})\, p(\sigma_D^2)\, p(\sigma_D^{2,*} \,|\, \sigma_D^2)}\right\}$$
$$= \min\left\{1, \frac{\sigma_D^2\, p(\sigma_D^{2,*})\prod_{i=1}^{n_D} \lambda_D^*(t_i^D \,|\, \boldsymbol{X}_i(t_i^D))}{\sigma_D^{2,*} p(\sigma_D^2)\prod_{i=1}^{n_D} \lambda_D(t_i^D \,|\, \boldsymbol{X}_i(t_i^D))} \times \frac{\exp(-A^* - B^* - C^*)}{\exp(-A - B - C)}\right\},$$

where: $\psi^* = (\mu_D, \sigma_D^{2,*}, \mu_P, \sigma_P^2)$, $\lambda_D^*(t \,|\, \boldsymbol{X}(t))$ is given by Eq. (3) with $\sigma_D^2 = \sigma_D^{2,*}$, $\boldsymbol{X}(t)$ is given by Eq. (5) and $\int_0^t \lambda_D(w \,|\, X(w))\, dw$ is given by Eq. (A.4). Terms in the second multiplicand e.g. $A^*, A, \ldots$ etc have already been defined earlier.

Sample $\boldsymbol{\sigma_P^2}$. This is identical to sampling from σ_D^2, with $\lambda_D(t \,|\, \boldsymbol{X}(t))$ replaced by $\lambda_P(t \,|\, \boldsymbol{X}(t))$ throughout and $\psi^* = (\mu_D, \sigma_D^2, \mu_P, \sigma_P^{2,*})$.

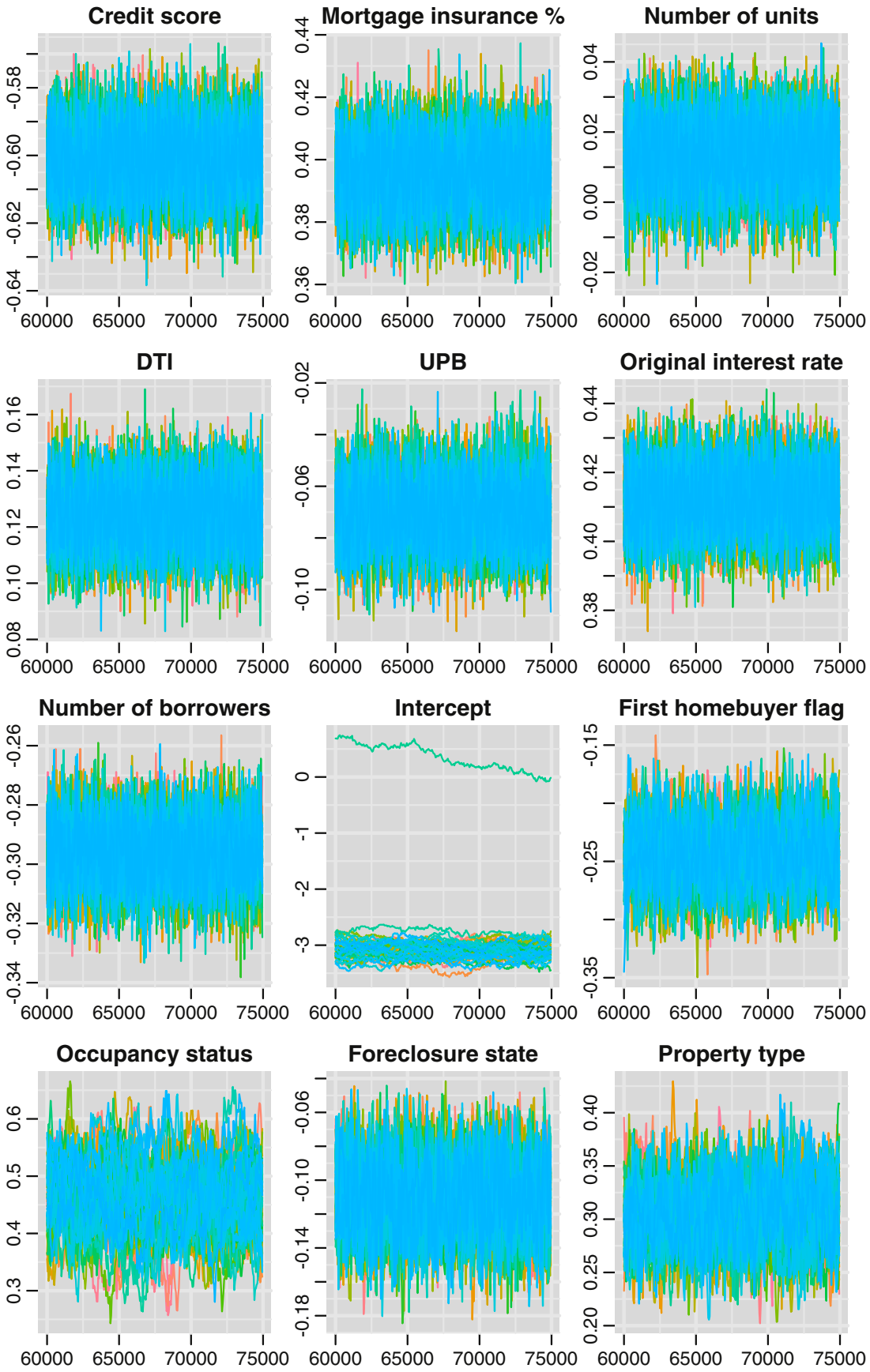

Fig. 9. Trace plot of all parameters associated with covariates for default category. A single chain for intercept parameter is found to converge much more slowly than the others.

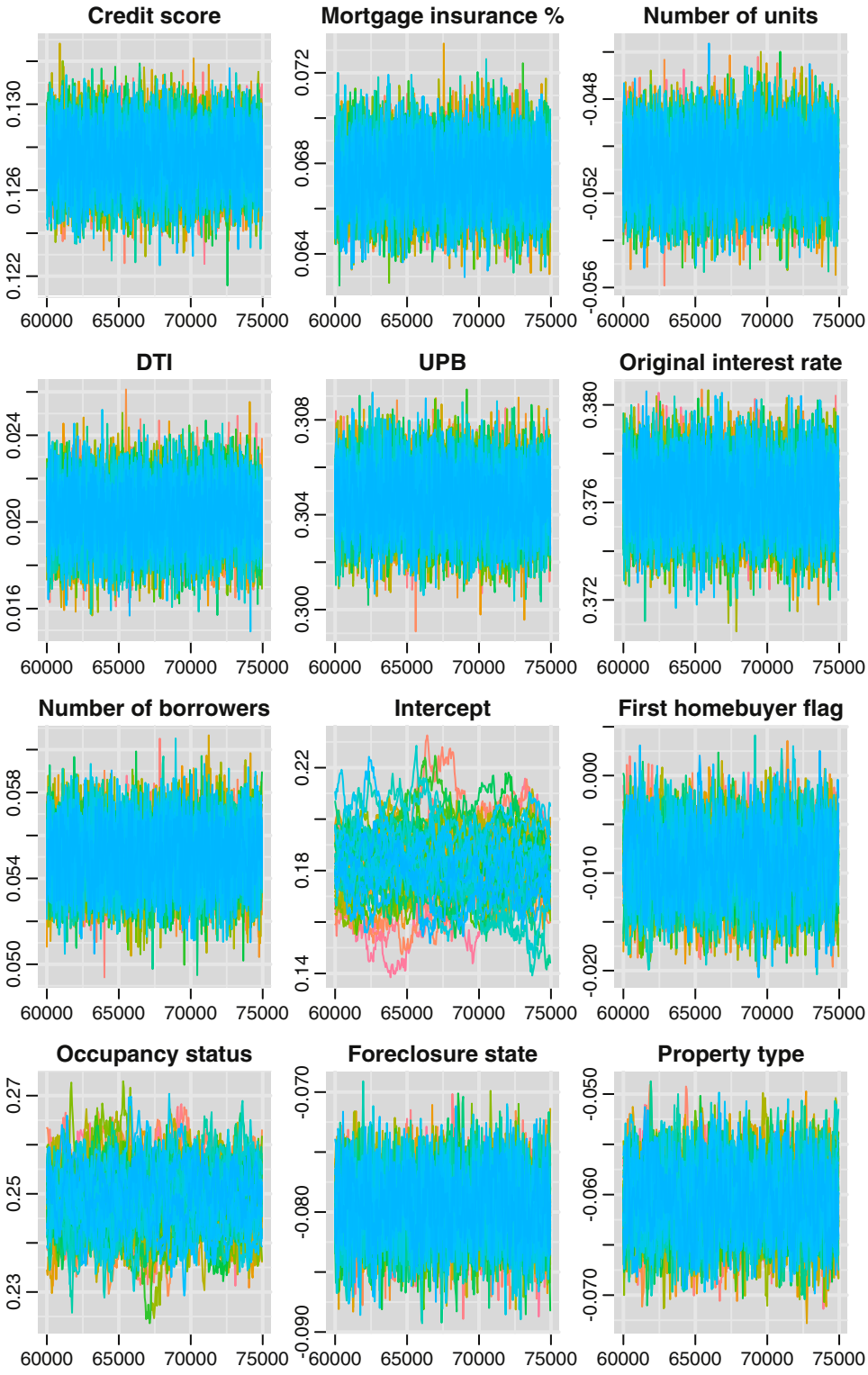

Fig. 10. Trace plot of all parameters associated with covariates for prepaid category.

MCMC output plots

Subsequent figures provide trace plots of all the variables and distributional parameters from both default and prepaid models. The problem with convergence of a single chain is noticeable in the intercept (β_0), μ and σ parameters in the default category. This can possibly be attributed to the twin problem of identifiability and low number of mortgages in this category. A larger proportion of default mortgages data and/or a longer run of the chain would have prevented this problem.

Trace plots of distributional parameters μ_d, μ_p, σ_d, σ_p are provided in Fig. 8. Note the single slow converging chain in the default category.

The trace plots for all the variables for category default are provided in Fig. 9 which seem to indicate towards fair convergence.

Table 4
Confusion matrix of classification results for the test set, for both types of lasso penalties applied to logistic regression. Nearly all mortgages have been classified as *Prepaid*.

	Ungrouped	
	Predicted default	Predicted prepaid
Default	0 (0.00%)	510 (100.00%)
Prepaid	1 (0.0034%)	29252 (99.9966%)
	Grouped	
	Predicted default	Predicted prepaid
Default	0 (0.00%)	510 (100.00%)
Prepaid	1 (0.0034%)	29251 (99.9932%)

Table 5
Confusion matrix of classification results for the test set from using random forest. The default classification success rate is 28%, much lower than that of our model.

	Predicted default	Predicted prepaid
Default	143 (28.04%)	367 (71.96%)
Prepaid	0 (0.00%)	29253 (100.00%)

The problem with a single chain is again noticeable in the intercept (β_0^{d}) trace.

Trace plots of parameters associated with category prepaid are provided in Fig. 10. The traces converge well and seem to have identified the posteriors satisfactorily.

Machine learning output

The ML outputs - logistic regression+lasso and Random Forest are provided in this section.

Logistic regression with lasso. Table 4 shows the confusion matrix corresponding to logistic regression + lasso with both ungrouped and grouped penalties on the coefficients. The latter ensures that a grouped lasso penalty is applied on the coefficients, such that they remain or are dropped together for all categories in the multinomial. The confusion matrix only contain classification performance of default and prepayment categories.

Random Forest. Table 5 is the confusion matrix for the random forest implementation to our data. Performance is much better that lasso, however default detection is lower than our model.

References

Aktekin, T., Soyer, R., & Xu, F. (2013). Assessment of mortgage default risk via Bayesian state space models. *The Annals of Applied Statistics, 7*(3), 1450–1473.

Ambrose, B. W., & Capone, C. A. (1998). Modeling the conditional probability of foreclosure in the context of single-family mortgage default resolutions. *Real Estate Economics, 26*(3), 391–429.

Andritzky, J. R. (2014). Resolving residential mortgage distress: Time to modify?IMF Working Paper, WP/14/226., IMF.

Calhoun, C. A., & Deng, Y. (2002). A dynamic analysis of fixed and adjustable-rate mortgage terminations. *The Journal of Real Estate Finance and Economics, 24*(1), 9–33.

Ciochetti, B. A., Deng, Y., Gao, Y., & Yao, R. (2002). The termination of commercial mortgage contracts through prepayment and default: A proportional hazard approach with competing risks. *Real Estate Economics, 30*(4), 595–633.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B, 34*(2), 187–220.

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. CRC Press.

Danis, M. A., & Pennington-Cross, A. (2008). The delinquency of subprime mortgages. *Journal of Economics and Business, 60*(1–2), 67–90.

Deng, R., & Haghani, S. (2018). FHA Loans in foreclosure proceedings: Distinguishing sources of interdependence in competing risks. *Journal of Risk and Financial Management, 11*(1), 1911–8074.

Deng, Y. (1997). Mortgage termination: An empirical hazard model with a stochastic term structure. *The Journal of Real Estate Finance and Economics, 14*(3), 309–331.

Deng, Y., & Order, R. V. (2000). Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica, 68*(2), 275–307.

Deng, Y., Quigley, J. M., & Order, R. V. (1996). Mortgage default and low down payment loans: The costs of public subsidy. *Regional Science and Urban Economics, 26*(3), 263–285.

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software, 40*(8), 1–18.

Fitzpatrick, T., & Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. *European Journal of Operational Research, 249*(2), 427–439.

Galloway, M., Johnson, A., & Shemyakin, A. (2017). Time-to-default analysis of mortgage portfolios. *Model Assisted Statistics and Applications, 12*(4), 359–367.

Gelfand, A. E., & Mallick, B. K. (1995). Bayesian analysis of proportional hazards models built from monotone functions. *Biometrics, 51*(3), 843–852.

Gilberto, S. M., & Houston Jr. , A. L. (1989). Relocation opportunities and mortgage default. *Real Estate Economics, 17*(1), 55–69.

Goodman, A. C., & Smith, B. C. (2010). Residential mortgage default: Theory works and so does policy. *Journal of Housing Economics, 19*(4), 280–294.

Kau, J. B., Keenan, D. C., III, W. J. M., & Epperson, J. F. (1990). Pricing commercial mortgages and their mortgage-backed securities. *The Journal of Real Estate Finance and Economics, 3*(4), 333–356.

Kiefer, N. M. (2010). Default estimation and expert information. *Journal of Business and Economic Statistics, 28*(2), 320–328.

Lambrecht, B. M., Perraudin, W. R. M., & Satchell, S. (1997). Time to default in the UK mortgage market. *Economic Modelling, 14*(4), 485–499.

Lambrecht, B. M., Perraudin, W. R. M., & Satchell, S. (2003). Mortgage default and possession under recourse: A competing hazards approach. *Journal of Money, Credit and Banking, 35*(3), 425–442.

Lee, Y., Rösch, D., & Scheule, H. (2016). Accuracy of mortgage portfolio risk forecasts during financial crises. *European Journal of Operational Research, 249*(2), 440–456.

Leece, D. (2004). *Economics of the mortgage market: Perspectives on household decision making*. Wiley-Blackwell.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136.

Liu, F., Hua, Z., & Lim, A. (2015). Identifying future defaulters: A hierarchical Bayesian method. *European Journal of Operational Research, 241*(1), 202–211.

Olrich, D. (2006). A new era for default management. *Mortgage Banking, 66*(6), 127–128.

Popova, I., Popova, E., & George, E. I. (2008). Bayesian forecasting of prepayment rates for individual pools of mortgages. *Bayesian Analysis, 3*(2), 393–426.

Quercia, R. G., & Stegman, M. A. (1992). Residential mortgage default: A review of the literature. *Journal of Housing Research, 3*(2), 341–379.

Soyer, R., & Xu, F. (2010). Assessment of mortgage default risk via Bayesian reliability models. *Applied Stochastic Models in Business and Industry, 26*(3), 308–330.

Sun, D., & Berger, J. O. (1993). Recent developments in Bayesian sequential reliability demonstration tests. In A. P. Basu (Ed.), *Advances in reliability*. Amsterdam: North-Holland.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics, 22*(4), 1701–1728.

Tong, E. N., Mues, C., & Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research, 218*(1), 132–139.