Stockholms universitet

# Research Report

## Department of Statistics


No. 2015:3

# DYNAMIC MIXTURE-OF-EXPERTS MODELS FOR LONGITUDINAL AND DISCRETE-TIME SURVIVAL DATA

Matias Quiroz
Mattias Villani

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

# DYNAMIC MIXTURE-OF-EXPERTS MODELS FOR LONGITUDINAL AND DISCRETE-TIME SURVIVAL DATA

MATIAS QUIROZ AND MATTIAS VILLANI

ABSTRACT. We propose a general class of flexible models for longitudinal data with special emphasis on discrete-time survival data. The model is a finite mixture model where the subjects are allowed to move between components through time. The time-varying probabilities of component memberships are modeled as a function of subject-specific time-varying covariates. This allows for interesting within-subject dynamics and manageable computations even with a large number of subjects. Each parameter in the component densities and in the mixing function is connected to its own set of covariates through a link function. The models are estimated using a Bayesian approach via a highly efficient Markov Chain Monte Carlo (MCMC) algorithm with tailored proposals and variable selection in all sets of covariates. The focus of the paper is on models for discrete-time survival data with an application to bankruptcy prediction for Swedish firms, using both exponential and Weibull mixture components. The dynamic mixture-of-experts models are shown to have an interesting interpretation and to dramatically improve the out-of-sample predictive density forecasts compared to models with time-invariant mixture probabilities.

KEYWORDS: Bayesian inference, Markov Chain Monte Carlo, Bayesian variable selection, Survival Analysis, Mixture-of-experts.

## 1. INTRODUCTION

We propose a finite mixture model for flexible modeling of longitudinal data. Our model belongs to the mixture-of-expert type of models first proposed by Jacobs et al. (1991) and Jordan and Jacobs (1994). In particular, we extend the class of *Generalized Smooth Mixture* (GSM) models presented in Villani et al. (2009) and Villani et al. (2012) to a longitudinal

Quiroz: *Research Division, Sveriges Riksbank, SE-103 37 Stockholm, Sweden* and *Department of Statistics, Stockholm University. E-mail: quiroz.matias@gmail.com.* Villani: *Division of Statistics, Department of Computer and Information Science of Statistics, Linkoping University.* E-mail: mattias.villani@liu.se. The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Executive Board of Sveriges Riksbank.

data setting. Villani et al. (2012) generalizes the *Smoothly Mixing Regression* (SMR) model in Geweke and Keane (2007). The key features of our approach are: i) subjects are allowed to move between mixture components over time, ii) the within-subject dynamics is modeled by letting the component membership probabilities be functions of subject-specific time-varying covariates, and iii) an efficient Bayesian inference methodology using MCMC with variable selection.

Finite mixtures are useful for modeling unobserved heterogeneity in many fields, see Frühwirth-Schnatter (2006) for a general introduction to finite mixture models. Given the longitudinal dimension of our data this paper is closely related to clustering panel data in form of relatively short time series, see Frühwirth-Schnatter (2011) for a recent survey. A main difference between this literature and our approach is that we allow for the possibility of subjects to change clusters over time.

Our main focus in this paper is on using dynamic mixture-of-experts models for analyzing survival data, see Miller et al. (1981) and Ibrahim et al. (2005) for general introductions to survival analysis. The most widely used model for survival data is the *Proportional Hazards*, or *Cox regression model* introduced in Cox (1972). The restrictiveness of the proportionality assumption and the inability to capture unobserved heterogeneity has lead researches to develop more flexible models. A popular model extension is to multiply the hazard with a subject-specific random effect, often called a frailty; Mosler (2003) surveys the theory and applications of these models in econometrics. The frailty can be continuous from a parametric distribution (Lancaster, 1979 and Vaupel et al., 1979) or be modelled by a finite mixture (Huynh and Voia, 2009) to capture a wide variety of functional shapes. Alternatively, finite mixture models offers a rich model class where some of the restrictive assumptions in the traditional survival models can be relaxed. McLachlan et al. (1994) provides a survey on the role of finite mixture models in survival analysis. Finite mixture of survival models are closely related to frailty models which is most easily seen when the distribution of the frailty is discrete and finite. The intuitive interpretation of finite mixtures combined with

the capability of modeling frailties makes it an interesting framework for analyzing complex data structures in survival analysis.

In most economics and social sciences applications, time is measured discretely (Allison, 1982). Examples include labor economics when studying the duration of individual unemployment measured e.g. in weeks (Carling et al., 1996), or educational research where the data is often recorded in school years (Singer and Willett, 1993). In our application we model time to bankruptcy (in years) for nascent firms. Heterogeneity has not been explored as much in the discrete-time framework. Notable exceptions are the continuous frailties in Xue and Brookmeyer (1997) and the finite mixture approach in Muthén and Masyn (2005).

Our article extends Muthén and Masyn (2005) in the following directions. First, we allow subjects to be classified to potentially different mixture components at each time period (dynamic mixture) while Muthén and Masyn (2005) restrict each subject to belong to one and only one mixture component during its exposure time (static mixture). Second, we use the Bayesian paradigm and Markov Chain Monte Carlo (MCMC) to estimate the model. This allows us to use Bayesian variable selection to obtain model parsimony and give insights on importance of covariates in different parts of the model. Our approach can also be straightforwardly extended to include the general latent variable (factor analysis) part in Muthén and Masyn (2005).

This paper is organized as follows. Section 2 presents the longitudinal mixture-of-experts models in a general setting. Section 3 applies the framework in Section 2 to discrete-time survival models and introduces the mixture component models. Section 4 presents the inference methodology and the general MCMC algorithm with variable selection. Section 5 illustrates the methodology by modeling the bankruptcy risk for nascent Swedish firms. Section 6 discusses future research and concludes. In Appendix A we state and prove a theorem on the flexibility of the proposed model.

## 2. Mixture-of-experts models for Longitudinal data

In the standard cross-sectional framework a smooth finite mixture density with $K$ components can be formulated as

$$p(y_i|x_i,\beta,\gamma) = \sum_{k=1}^{K} w_k(z_i|\gamma_k)p_k(y_i|x_i,\beta_k), \quad i=1,...,n, \tag{2.1}$$

where $w_k(z_i|\gamma_k)$ denotes the $i$th observation's mixing probability and can be interpreted as its prior probability of belonging to the $k$th component density $p_k(y_i|x_i,\beta_k)$. We often set $z=x$, but they can differ. To simplify inference with the Gibbs sampler, augmented data $s_1, s_2, \ldots s_n$ is introduced so that $s_i = k$ means that the $i$th observation belongs to the $k$th component. The model in Equation (2.1) can then be formulated as

$$\begin{aligned} y_i|(s_i=k,x_i,\beta_k) &\sim p_k(y_i|x_i,\beta_k) \\ P(s_i=k|z_i,\gamma_k) &= w_k(z_i|\gamma_k). \end{aligned}$$

To extend to a longitudinal mixture the following notation is introduced. Assume subject $i$ has been observed over $n_i$ time periods. Let $y_{1:n_i} = (y_{i1},\ldots,y_{in_i})^T \in \mathbb{R}^{n_i\times 1}$, $x_{1:n_i} = (x_{i1},\ldots,x_{in_i})^T \in \mathbb{R}^{n_i\times p_x}$ and $z_{1:n_i} = (z_{i1},\ldots,z_{in_i})^T \in \mathbb{R}^{n_i\times p_z}$. Let $v_i \in \mathbb{R}^{p_v\times 1}$ denote the time-invariant predictors and $s_{1:n_i} \in \{1,\ldots,K\}^{n_i}$ where $s_{ij}=k$ if the $i$th subject belongs to component $k$ at time period $j$. The longitudinal dimension allows for two main specifications of $s$: $s_{ij}=k$ for all $j$ or $s_{ij}=k_j$ where $k_j \in \{1,2\ldots K\}$. We refer to the former as a *static mixture* and the latter as a *dynamic mixture*. We say that a model is a $p$-lag longitudinal model if the joint distribution factorizes as

$$p(y_{1:n_i}|x_{1:n_i},\beta) = \prod_{j=1}^{n_i} p(y_{ij}|y_{j-p:j-1},x_{ij},\beta), \tag{2.2}$$

under the common assumption that $p$ pre-sample observations $y_{i0}, y_{i,-1}, ..., y_{i,-(p-1)}$ are available when $p$ lags of the response is used in the model.

*Static mixture.* Let $v_i$ be a vector with time invariant covariates. The static mixture model is a finite mixture for the joint distribution of the $p$-lag model in Equation (2.2), i.e.

$$
\begin{aligned}
p(y_{1:n_i}|x_{1:n_i}) &= \sum_{k=1}^{K} w_k(v_i) p_k(y_{1:n_i}|x_{1:n_i}) \\
&= \sum_{k=1}^{K} w_k(v_i) \left( \prod_{j=1}^{n_i} p_k(y_{ij}|y_{j-p:j-1}, x_{ij}, \beta) \right) \qquad (2.3)
\end{aligned}
$$

where the dependence on parameters is suppressed. Note that the covariates $x_{ij}$ and $v_i$ can enter in the component models, while the mixing function is only a function of time-invariant predictors $v_i$. This is because the static mixture has by definition the same mixture probabilities $w_k$ for all observations in the sequence $y_{1:n_i}$, and therefore it is not possible to have time-varying covariates in the mixing function as the subject would then for example be allocated to a component at time $j = 1$ based on future information ($j = 2, 3, \dots$) not available at that time. To avoid notational clutter, we often suppress the dependence on $v_i$ in the overall mixture and in the components. The mixing probabilities are modeled with the multinomial logit

$$
w_k(v_i) = \frac{\exp(v_i^T \gamma_k)}{\sum_{l=1}^{K} \exp(v_i^T \gamma_l)} \qquad (2.4)
$$

where $\gamma_k \in \mathbb{R}^{p_v \times 1}$ with $\gamma_1 = 0$ for identification.

The latent variable formulation of the model in Equation (2.3) is

$$
\begin{aligned}
y_{1:n_i}|s_i = k, x_{1:n_i} &\sim \prod_{j=1}^{n_i} p_k(y_{ij}|y_{j-p:j-1}, x_{ij}, \beta) \\
P(s_i = k|v_i) &= \frac{\exp(v_i^T \gamma_k)}{\sum_{l=1}^{K} \exp(v_i^T \gamma_l)}. \qquad (2.5)
\end{aligned}
$$

The model in Equation (2.3) expresses the *joint* distribution of the finite mixture. It is straightforward to show that the density at period $t$ conditional on previous values is given by

$$
p(y_{ij}|y_{i(j-p:j-1)}, x_{ij}) = \sum_{k=1}^{K} \tilde{w}_{ij}^{k} \cdot p_k(y_{ij}|y_{j-p:j-1}, x_{ij}), \qquad (2.6)
$$

where

$$\tilde{w}_{ij}^k = w_k \cdot \frac{\prod_{j'=1}^{j-1} p_k(y_{ij'}|y_{i(j'-p):i(j'-1)}, x_{ij'}, \beta)}{\sum_{l=1}^{K} w_l \left(\prod_{j'=1}^{j-1} p_l(y_{ij'}|y_{i(j'-p):i(j'-1)}, x_{ij'}, \beta)\right)}. \tag{2.7}$$

Note that the *conditional* distribution of the static model actually has time-varying mixture weights, but in a highly restrictive form where the weight for component $k$ at time $j$ is a function of the probability of the observed data up to time $j-1$ given that component. This means that component $k$ cannot obtain a large weight $\tilde{w}_{ij}^k$ at time $j$ unless it assigns a high joint probability to the complete history $y_{1:j-1}$. Since $w_k$ is constant through time and therefore determined from all observation jointly, the flexibility from the static mixture is very limited.

*Dynamic mixture.* Restricting a subject to a single component over time may not be realistic in some situations because individual behavior may not be homogeneous over time. To exemplify, consider the modeling of firm bankruptcy. If the components can be interpreted as high versus low risk for bankruptcy, the assumption of being constantly a risky or a safe firm is unrealistic. The economic surrounding and individual financial variables do change over time which is likely to make the firm more or less risky.

The obvious approach to a dynamic mixture is to let $s_{1:n_i}$ follow a (hidden) Markov model, see Baum and Petrie (1966) and Kim and Nelson (2003). The posterior sampling of $s_{1:n_i}$ is then performed sequentially from the conditional distribution at each time point using e.g. forward filtering-backward sampling algorithm for Gaussian models (Carter and Kohn, 1994 and Frühwirth-Schnatter, 1994) or Sequential Monte Carlo (SMC) for non-Gaussian models (Doucet et al., 2000). Such an approach is computationally infeasible in many longitudinal applications since the SMC would have to be performed for each of the subjects, which is clearly not an option in data sets with a large number of subjects, such as the one in our application to firm bankruptcy. An alternative approach is to sample directly from the joint distribution for each $s_{1:n_i}$ sequence (Franzén, 2008), but the sample space of $s_{1:n_i}$ grows

dramatically with $K$ and the number of time periods $n_i$ so $P(s_{1:n_i}|y_{1:n_i}, x_{1:n_i}, z_{1:n_i})$ quickly becomes computationally intractable.

To overcome these problems we suggest the following approach. Let $s_{1:n_i}$ be an independent sequence *conditional* on the path of time-varying covariates $z_{1:n_i}$, i.e.

$$P(s_{1:n_i} = k_{1:n_i}|z_{1:n_i}) = \prod_{j=1}^{n_i} P(s_{ij} = k_j|z_{ij}), \tag{2.8}$$

with $k_{1:n_i} = (k_1, \ldots, k_{n_i})$ and $1 \leq k_j \leq K$ for $j = 1, \ldots, n_i$. The temporal dependence of the time series $s_{1:n_i}$ is thus induced by the path of the time series for the covariates in $z_{1:n_i}$; note that lagged values of the response may be included in $z$. The strength of this approach is that given the time path of the covariates (and other model parameters) the component allocations can be sampled independently for all subjects and time periods in the Gibbs sampler, see Section 4.2.

The dynamic mixture of the $p$-lag model in Equation (2.2) is a finite mixture on each conditional distribution, i.e.

$$p(y_{ij}|y_{j-p:j-1}, x_{ij}) = \sum_{k=1}^{K} w_{ij}^k(z_{ij}) p_k(y_{ij}|y_{j-p:j-1}, x_{ij}),\ j = 1, ..., n_i \tag{2.9}$$

where

$$w_{ij}^k(z_{ij}) = \frac{\exp(z_{ij}^T \gamma_k)}{\sum_{l=1}^{K} \exp(z_{ij}^T \gamma_l)}, \quad \text{with } z_{ij} = (x_{ij}, y_{j-p:j-1})^T \tag{2.10}$$

and $\gamma_k \in \mathbb{R}^{p_z \times 1}$ with $\gamma_1 = 0$ for identification. The joint distribution for the $i$th subject becomes

$$p(y_{1:n_i}|x_{1:n_i}) = \prod_{j=1}^{n_i} \left( \sum_{k=1}^{K} w_{ij}^k p_k(y_{ij}|y_{j-p:j-1}, x_{ij}) \right). \tag{2.11}$$

Time invariant predictors $v_i$ may also be included in $z_i$. Rather than modeling the $x$-process directly, persistence in component allocations over time can be achieved by defining $z_{ij}$ as

an exponential moving average of the time dependent covariates $x_{ij}$

$$z_{ij} = \alpha x_{ij} + (1-\alpha) z_{i(j-1)},$$

and $z_{i1} = x_{i1}$ (may be a moving average of data before the start of the analyzed sample), where $0 \leq \alpha \leq 1$, and $\alpha = 1$ corresponds to no smoothing. Persistence prevents a sudden change in the explanatory variables to trigger an immediate reallocation of the subject; a sudden decrease in a firm's profits may not immediately make it a high risk firm, but several consecutive years of losses might. Lagged values of $y$ may be included directly in $z$, but may also be transformed by exponential moving averages.

Jiang and Tanner (1999) prove that standard (non-longitudinal) mixture-of-experts, with sufficiently many exponential family regression models with generalized linear mean functions, can approximate any density in the exponential family with an essentially arbitrarily non-linear predictor. In Appendix A we build on that result and show that the proposed dynamic mixture model approximates a longitudinal generalization of the target class in Jiang and Tanner (1999) arbitrarily well as the number of components increase.

The latent variable formulation of Equation (2.9) is

$$\begin{aligned} y_{ij}|s_{ij} = k, x_{ij} &\sim p_k(y_{ij}|y_{j-p:j-1}, x_{ij}) \\ P(s_{ij} = k|z_{ij}) &= \frac{\exp(z_{ij}^T \gamma_k)}{\sum_{l=1}^K \exp(z_{ij}^T \gamma_l)}. \end{aligned} \tag{2.12}$$

The dynamic mixture has the interpretation that the *conditional* density at a given time period is a mixture of densities with weights that are directly modeled as function of the covariates $z$. This is in sharp contrast to the static mixture where the mixture weights in the conditional distributions are history dependent without a natural interpretation, see Equation (2.7) and the ensuing discussion. Note that one can easily add lags of the $x$'s in the component models. Another route to generate temporal dynamics is to add random effects, which would be straightforward to include as an additional updating step in our MCMC scheme.

## 3. Mixture-of-experts model for survival data

This section presents the survival models that will be used as the components in the finite mixture.

3.1. **Discrete-time survival data.** Let the random variable $T^c$ denote the time to some unrepeatable event. Survival data are often observed in discrete time, for example monthly or yearly, see e.g. Allison (1982) and Singer and Willett (1993). Assume that a study is observed over $J$ periods which can be divided as $(0, t_1], (t_1, t_2], \ldots (t_{J-1}, t_J]$. Let $T \in \{1, 2, ...\}$ be the discrete random variable recording the time period where the event occurs, i.e. $T = j$ if $T^c \in (t_{j-1}, t_j]$. It is convenient to express the joint likelihood of the data in terms of the hazard, which in discrete time is the conditional probability $h_j = P(T = j | T \geq j)$. Let the $i$th subjects' hazard probability at period $j$ be denoted $h_{ij} = h(x_{ij})$. Assuming $n$ independent subjects, the likelihood is expressed as

$$L = \prod_{i=1}^{n} \prod_{j=1}^{n_i} h(x_{ij})^{y_{ij}} (1 - h(x_{ij}))^{1-y_{ij}}, \tag{3.1}$$

where

$$y_{ij} = \begin{cases} 0 \text{ if subject } i \text{ does not experience the event at period } j, \\ 1 \text{ if subject } i \text{ does experience the event at period } j. \end{cases}$$

Singer and Willett (1993) and Shumway (2001) note that this likelihood has the same form as regression for binary data with $h^{-1}$ as the link function.

In this paper two different models are considered. The first, the *exponential* model, is derived by assuming that $T^c \sim Exp(\lambda)$ and using a log-link $g(\lambda) = \log(\lambda)$. Then the discrete-time hazard is easily shown to be

$$h(x_{ij}) = 1 - \exp(-\exp(\alpha + x_{ij}^T \beta)(t_{ij} - t_{i(j-1)})).$$

The second, the *Weibull* model, is derived by assuming the Weibull distribution for $T^c$, parametrized by $f(t|\lambda, \rho) = \rho \lambda t^{\rho-1} \exp(-\lambda t^\rho)$, which implies

$$h(\lambda_{ij}, \rho_{ij}) = 1 - \exp(-\lambda_{ij}(t_{ij}^{\rho_{ij}} - t_{i(j-1)}^{\rho_{ij}})).$$

Because both $\lambda$ and $\rho$ are positive the dependence on the covariates are modeled through

$$\begin{aligned} \log(\lambda_{ij}) &= \alpha_\lambda + x_{\lambda_{ij}}^T \beta_\lambda \\ \log(\rho_{ij}) &= \alpha_\rho + x_{\rho_{ij}}^T \beta_\rho. \end{aligned}$$

Both these models can easily be extended with a flexible baseline hazard, see our application in Section 5.

Discrete-time survival data is recorded as the binary vector $y_{1:n_i} = (0, 0, \ldots, c_i)$ where $c_i \in \{0, 1\}$ is the censor indicator such that $c_i = 0$ means that the $i$th subject did not experience the event in the study period. The joint distribution is

$$p(y_{1:n_i}|x_{1:n_i}) = \left( \prod_{j=1}^{n_i-1} p(y_{ij} = 0|y_{i(j-1)} = 0, x_{ij}) \right) p(y_{in_i} = c_i|y_{i(n_i-1)} = 0, x_{in_i}),$$

so discrete-time survival models are 1-lag longitudinal using the terminology in Section 2.

3.2. **Smooth mixtures of survival models.** We characterize the distribution by the hazard probability. The hazard probability will depend on a set of model parameters $\phi_1, \ldots, \phi_L$. As in Villani et al. (2012) each parameter depends on a set of predictors through link functions $g_l(\phi_l) = x_l^T \beta_l$. For example, in the Weibull model we have $\phi_1 = \lambda$, $\phi_2 = \rho$ and both links are logs. The likelihood for a given mixture component is

$$L(\beta_1, ..., \beta_L) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} h(x_{ij}|\phi_1, \ldots \phi_L)^{y_{ij}} (1 - h(x_{ij}|\phi_1, \ldots \phi_L))^{1-y_{ij}} \tag{3.2}$$

where $\phi_l = g_l^{-1}(x_l^T \beta_l)$.

*Static mixture.* The general expression for this model is given in Equation (2.3). This is the latent class model considered in Muthén and Masyn (2005), but without the general latent variable part and not restricted to the logit hazard model. The interpretation is that the mixture is on the joint distribution of $y_i$, i.e.

$$p(y_{1:n_i}|x_{1:n_i}) = \sum_{k=1}^{K} w_k p_k(y_{1:n_i}|x_{1:n_i}) = \sum_{k=1}^{K} w_k \left( \prod_{j=1}^{n_i-c_i} (1 - h_{ij}^k) \right) (h_{in_i}^k)^{c_i} \tag{3.3}$$

and the dependence on covariates and parameters is suppressed everywhere to save space. In the component model $h_{ij}^k = h^k(x_{ij}, v_i)$ while the mixing function is $w_k = w_k(v_i)$. The mixing probabilities are modeled with the multinomial logit as in Equation (2.4).

The hazard probability at period $t$ is the equivalent of the conditional density in Equation (2.6), i.e.

$$p(y_{it} = 1 | y_{i(t-1)} = 0) = \sum_{k=1}^{K} \tilde{w}_{it}^k \cdot h_{it}^k,$$

where

$$\tilde{w}_{it}^k = w_k \cdot \frac{\prod_{j=1}^{t-1}(1 - h_{ij}^k)}{\sum_{l=1}^{K} w_l \left( \prod_{j=1}^{t-1}(1 - h_{ij}^l) \right)}.$$

As discussed in Section 2, although the mixture weights $\tilde{w}_{it}^k$ are time-varying in the hazard for the static mixture, it is important to remember that the static mixture is a mixture for the joint distribution with time invariant weights. The form of the conditional weights $\tilde{w}_{it}^k$ is very hard to interpret and do not allow for flexible time-varying hazards. This is in contrast with the dynamic mixture where the hazard probabilities are by construction a flexible and highly interpretable mixture of component hazards.

The latent variable formulation of the model in Equation (3.3) is

$$\begin{aligned} y_{1:n_i} | s_i = k, x_{1:n_i} &\sim \left( \prod_{j=1}^{n_i - c_i} (1 - h_{ij}^k(x_{ij})) \right) (h_{in_i}^k(x_{ij}))^{c_i} \\ P(s_i = k | v_i) &= \frac{\exp(v_i^T \gamma_k)}{\sum_{l=1}^{K} \exp(v_i^T \gamma_l)}. \end{aligned} \tag{3.4}$$

*Dynamic mixture.* The general dynamic mixture model in Equation (2.9) can be formulated in terms of hazards as

$$p(y_{1:n_i} | x_{1:n_i}) = \left( \prod_{j=1}^{n_i - 1} \left( \sum_{k=1}^{K} w_{ij}^k (1 - h_{ij}^k) \right) \right) \left( \sum_{k=1}^{K} w_{in_i}^k (h_{in_i}^k)^{c_i} (1 - h_{in_i}^k)^{1 - c_i} \right) \tag{3.5}$$

where $h_{ij}^k = h^k(x_{ij}, v_i)$ and $w_{ij}^k$ follows the multinomial model in Equation (2.10). Note that the hazard of the dynamic mixture at any given time period is a smooth mixture of hazards, i.e. a mixture-of-experts model.

We prove in Appendix A that the dynamic mixture of longitudinal experts is arbitrarily flexible as $K$ increases. The proof builds on a result in Jiang and Tanner (1999) (for non-longitudinal mixtures) that applies when the components (and also the target class, see Appendix A and Jiang and Tanner (1999, p. 992)) belong to a one parameter exponential family, i.e.

$$p(y|x;g(\cdot)) = \exp\left(a(g(x))y + b(g(x)) + c(y)\right). \tag{3.6}$$

Note that our components can be written in the form of a Bernoulli model, i.e.

$$p_k(y_t = y|y_{t-1} = 0) = \theta_k^y(1-\theta_k)^{1-y}$$

with $y \in \{0,1\}$ and $\theta_k = g(x) = h(x_t)$ is a smooth function of the covariates. For the exponential model it can easily be verified that it is of the form in Equation (3.6). The Weibull model has two parameters, and is therefore outside the Jiang-Tanner target class, but it includes the exponential model as a special case ($\rho = 1$), and therefore it is more flexible for a given number of mixture components. This extra flexibility is shown to be empirically important in our application in Section 5.

The marginal effect of covariate $x_t$, on the hazard when $z_t = \alpha x_t + (1-\alpha)z_{t-1}$, is given by

$$\frac{d}{dx_t}h_t(x_t) = \sum_{k=1}^{K} \alpha \frac{d}{dz_t}\left(w_k(z_t)\right) h_t^k(x_t) + w_k(z_t)\frac{d}{dx_t}h_t^k(x_t),$$

where the derivative of the multinomial logit is

$$\frac{d}{dz_t}\left(w_k(z_t)\right) = w_k(z_t)\left[\gamma_k - \sum_{l=1}^{K} w_k(z_t)\gamma_l\right].$$

The latent variable formulation of Equation (3.5) is

$$y_{ij}|s_{ij} = k, x_{ij} \sim \begin{cases} 1 - h_{ij}^k(x_{ij}), & y_{ij} = 0 \\ h_{ij}^k(x_{ij}), & y_{ij} = 1 \end{cases}$$

$$P(s_{ij} = k|z_{ij}) = \frac{\exp(z_{ij}^T\gamma_k)}{\sum_{l=1}^{K}\exp(z_{ij}^T\gamma_l)}. \tag{3.7}$$

## 4. Inference

We adopt a Bayesian approach to inference and use a Metropolis-within-Gibbs sampler with variable selection to sample from the posterior distribution. The sampler utilizes the gradient and Hessian of the full conditional posterior to construct tailored proposals.

This section is organized as follows. First, prior distributions are introduced in all parts of the model. These priors are simple and the user only needs to specify prior beliefs about scalar parameters. Then the general MCMC scheme is illustrated, followed by a section describing the algorithm that construct tailored proposals for efficient inference. Finally, the method for choosing the number of components is explained.

### 4.1. **Prior Elicitation.**

4.1.1. *Components.* We use the prior construction initially developed in Ntzoufras et al. (2003) for the Generalized Linear Model (GLM) and subsequently refined and extended in Villani et al. (2012) to GSM models. Assume a component model with a single model parameter $\lambda$ and a link function $g$ such that $g(\lambda) = \alpha_\lambda + x^T\beta_\lambda$. We first discuss the prior on the intercept. Start by standardizing the covariates to have mean zero and unit standard deviation. The intercept $\alpha_\lambda$ is then $g(\lambda)$ at the mean of the original covariates. Assume that $\alpha_\lambda \sim N(m_\lambda, s_\lambda^2)$ and the task is to find $m_\lambda$ and $s_\lambda^2$ by eliciting a suitable prior on the model parameter $\lambda$ with mean and variance specified by the user, say $E(\lambda) = m_\lambda^*$ and $V(\lambda) = s_\lambda^{*2}$. In the simplest example, with the identity link, $\lambda \sim N(m_\lambda^*, s_\lambda^{*2})$ transforms directly to $\alpha_\lambda \sim N(m_\lambda, s_\lambda^2)$ with $m_\lambda = m_\lambda^*$ and $s_\lambda^2 = s_\lambda^{*2}$. In the case with a log-link used in this paper, a suitable prior on $\lambda$ is the log-normal density with mean $m_\lambda^*$ and variance $s_\lambda^{*2}$ which transforms to $\alpha_\lambda \sim N(m_\lambda, s_\lambda^2)$ with $s_\lambda^2 = \log\left[\left(\frac{s_\lambda^*}{m_\lambda^*}\right)^2 + 1\right]$ and $m_\lambda = \log(m_\lambda^*) - s_\lambda^2/2$.

The regression coefficients in $\beta_\lambda$ are assumed to be a priori independent of $\alpha_\lambda$ with $\beta_\lambda \sim N(0, c_\lambda\Sigma_\lambda)$. Here $\Sigma_\lambda = (W^T\hat{D}_\lambda W)^{-1}$, where $W$ is the matrix of covariates excluding the intercept and $\hat{D}_\lambda$ is the conditional Fisher information for $\lambda$ evaluated at the prior modes of $\alpha_\lambda$ and $\beta_\lambda$, which is the vector $\hat{\beta}_\lambda = (m_\lambda, \mathbf{0}^T)^T$. Thus $\hat{D}_\lambda$ depends only on the constant $m_\lambda$.

The conditional Fisher information for $\lambda = (\lambda_1, \dots \lambda_n)^T$ is a diagonal matrix with elements

$$-E\left[\frac{\partial^2 \log p(y_i|\lambda_i)}{\partial \lambda_i^2}\right] g'_\lambda(\lambda_i)^{-2}.$$

Setting $c_\lambda = n$ gives a unit information prior, i.e. a prior that carries the information equivalent to a single subject from the model. For the models in our framework $\hat{D}_\lambda$ can not be obtained analytically but is easily computed by simulation. It is straightforward to extend the argument to elicit priors for more than one model parameter. For details and examples see Villani et al. (2012).

We allow for variable selection in all covariate sets in the model. For a given component let the indicator variable $\mathcal{I} = \{I_1, \dots I_{p_x}\}$ be defined such that $I_j = 0$ means that the $j$th element in $\beta$ is zero and the corresponding covariate drops out. Let $\beta_\mathcal{I}$ be the vector of non-zero coefficients, and for any $\mathcal{I}$ let $\mathcal{I}^c$ denote its complement. We make the assumption that the intercept is always in the model. Let $\beta \sim N(0, c\Sigma)$ as discussed above for the regression coefficients. Conditioning on the variables that are in the model we obtain

$$\beta_\mathcal{I}|\mathcal{I} \quad \sim \quad N\left[0, c(\Sigma_{\mathcal{I},\mathcal{I}} - \Sigma_{\mathcal{I},\mathcal{I}^c}\Sigma^{-1}_{\mathcal{I}^c,\mathcal{I}^c}\Sigma^T_{\mathcal{I}^c,\mathcal{I}})\right]$$

and $\beta_{\mathcal{I}^c}|\mathcal{I}$ is identically zero.

4.1.2. *Mixing function.* For the vector $\gamma = (\gamma_2^T, \dots \gamma_K^T)^T$ (recall that $\gamma_1 = 0$) we assume $\gamma \sim N(0, c_\gamma I)$. It is also possible to use a prior with non-diagonal structure as above but this is not pursued here. Variable selection is done similarly as above by introducing the indicator $\mathcal{I}_Z$ for $\gamma$.

4.1.3. *Variable selection indicators.* For both the component and the mixing parts of the model the indicators are assumed to be a priori independent and Bernoulli distributed, i.e $P(I_i = 1) = \pi$, $0 \leq \pi \leq 1$ and $\pi$ is allowed to be different for each model parameter. It is straightforward to let $\pi$ be unknown and estimate it in a separate updating step as in Kohn et al. (2001).

4.2. **General MCMC scheme.** Villani et al. (2009) experimented with different algorithms for finite mixture models in a related setting. Their preferred algorithm is the one used in this paper. The algorithm is a Metropolis-within-Gibbs sampler that draws the regression parameters and variable selection indicators jointly. Assume a component density with $L$ different model parameters and $K$ components. The following three blocks are sampled

(1) $s$

(2) $\gamma, \mathcal{I}_Z$

(3) $\{(\beta_1, \mathcal{I}_1), \ldots, (\beta_L, \mathcal{I}_L)\}_{k=1}^K$.

How to sample $s$ depends if it is a static or dynamic mixture. For the static mixture

$$P(s_i = k|x_i, v_i, y_i) \propto \left(\prod_{j=1}^{n_i - c_i} (1 - h^k(x_{ij}))\right) (h^k(x_{in_i}))^{c_i} \frac{\exp(v_i^T \gamma_k)}{\sum_{l=1}^K \exp(v_i^T \gamma_l)} \tag{4.1}$$

independently for $i = 1, \ldots, N$. For the dynamic mixture, the full conditional of $s_{ij}$ is independent of all other $s_{ij}$, $i = 1, ..., n$ and $j = 1, ..., n_i$, and is of the form

$$P(s_{ij} = k|x_i, z_i, y_i) \propto \begin{cases} h_{ij}^k \frac{\exp(z_{ij}^T \gamma_k)}{\sum_{l=1}^K \exp(z_{ij}^T \gamma_l)} & \text{if } c_i = 1 \text{ and } j = n_i \\ (1 - h_{ij}^k) \frac{\exp(z_{ij}^T \gamma_k)}{\sum_{l=1}^K \exp(z_{ij}^T \gamma_l)} & \text{otherwise.} \end{cases} \tag{4.2}$$

Note that this allows us to sample $s_{ij}$ independently for all $i$ and $j$ so this updating step is very fast in comparison with Markov models of $s_{ij}$.

Conditional on $s$, Step 2 is a multinomial logistic regression with variable selection. It is possible to apply a generalization of the algorithm described in the next section to handle this updating step efficiently, see Villani et al. (2009) for details.

4.3. **Variable-dimension finite step Newton proposals.** This section presents how to construct the tailored proposals for dynamic mixtures based on the algorithm in Villani et al. (2009) and Villani et al. (2012), which generalizes earlier algorithms in Gamerman (1997), Qi and Minka (2002) and Nott and Leonte (2004). For clarity, the algorithm is first presented in the case with no variable selection and then extended. The only requirement is that the

likelihood part of the posterior can be factorized as

$$p(\beta|y) \quad = \quad \prod_{i=1}^{N} p(y_i|\phi_i)p(\beta) \tag{4.3}$$

where $\phi_i = g^{-1}(x_i^T\beta)$. Note that there can be more than one model parameter and then $p(\beta|y)$ is a full conditional posterior distribution and the algorithm can be used as a Metropolis-within-Gibbs step. After a proper relabeling of the product in the likelihood in Equation (3.2) it has the same form as the likelihood part in Equation (4.3). The proposal distribution is tailored using an approximate posterior mode and the curvature around that mode. The approximate mode is found by taking a few steps with Newton's algorithm. To implement the algorithm we need the following results from Lemma 1 in Villani et al. (2012)

$$\frac{\partial \log p(y|\beta)}{\partial \beta} \quad = \quad X^T\tilde{g} \tag{4.4}$$

where $X$ is the covariate matrix, $\tilde{g} = (\tilde{g}_1, \ldots, \tilde{g}_n)^T$,

$$\tilde{g}_i \quad = \quad \frac{\partial \log p(y|\phi_i)}{\partial \phi_i} g'(\phi_i)^{-1}.$$

The outer-product approximation of the Hessian is

$$\frac{\partial^2 \log p(\beta|y)}{\partial \beta \partial \beta^T} \quad \approx \quad X^T D X, \tag{4.5}$$

where $D = \text{diag}(\tilde{g}_i^2)$. Villani et al. (2012) also derives expression for the exact Hessian but we have found the outer-product approximation to be more numerically stable for our problem. Note that the lemma only requires derivatives for the scalar parameters of the log-likelihood. Newton's algorithm is

$$\beta_{r+1} \quad = \quad \beta_r - A_r^{-1}s_r, \; r = 0, \ldots, R \tag{4.6}$$

where $s_r$ and $A_r$ is the gradient and Hessian of the log posterior, respectively. Using the results above we have

$$\begin{aligned} s_r &= X^T\tilde{g} + \frac{\partial \log p(\beta)}{\partial \beta} \\ A_r &= X^T DX + \frac{\partial^2 \log p(\beta)}{\partial \beta \partial \beta^T}. \end{aligned}$$

Start with $\beta_0 = \beta_c$ and let $\hat{\beta}$ be the vector obtained after $R$ Newton steps. This is not necessarily the mode but is often close because the previously accepted draw is used as initial value. Setting $R = 1, 2$ or 3 is usually sufficient. Let $\beta_c \in \mathbb{R}^{p_x \times 1}$ denote the current and $\beta_p \in \mathbb{R}^{p_x \times 1}$ the proposed posterior draw. The proposal distribution is a multivariate t-distribution with $\nu \geq 2$ degrees of freedom, i.e

$$\beta_p|\beta_c \sim t_\nu \left[\hat{\beta}, -\left(\frac{\partial^2 \log p(\beta|y)}{\partial \beta \partial \beta^T}\right)\Big|_{\beta=\hat{\beta}}\right].$$

To extend the algorithm to variable selection the pair $(\beta, \mathcal{I})$ is proposed jointly conditional on the previously accepted parameter and indicator. This proposal can be factorized as

$$J(\beta_p, \mathcal{I}_p|\beta_c, \mathcal{I}_c) = J_1(\beta_p|\mathcal{I}_p, \beta_c)J_2(\mathcal{I}_p|\beta_c, \mathcal{I}_c) \tag{4.7}$$

$J_1$ is a generalization of the proposal for $\beta_p$ above and $J_2$ is the proposal for the indicators. Consider first the $\beta$ proposal. Since $\beta_c$ and $\beta_p$ may be of different dimensions we use the following generalized Newton algorithm from Villani et al. (2012)

$$\beta_{r+1} = A_r^{-1}(B_r\beta_r - s_r),\ r = 0, \ldots, R \tag{4.8}$$

with

$$\begin{aligned} s_r &= X_{r+1}^T\tilde{g} + \frac{\partial \log p(\beta)}{\partial \beta} \\ A_r &= X_{r+1}^T DX_{r+1} + \frac{\partial^2 \log p(\beta)}{\partial \beta \partial \beta^T} \\ B_r &= X_{r+1}^T DX_r + \frac{\partial^2 \log p(\beta)}{\partial \beta \partial \beta^T}, \end{aligned}$$

where $X_r$ is the matrix with columns corresponding to the non-zero coefficients in $\beta_r$, and the likelihood part of the expressions are evaluated at $\beta = \beta_r$. The prior parts are evaluated at the entire vector $\beta$ (including the zero parameters) and then the sub-vector conformable with $\beta_{r+1}$ is extracted from the result. Note that after the first step the parameter no longer changes dimension and the generalized Newton algorithm reduces to the usual Newton algorithm.

Following Villani et al. (2009) and Villani et al. (2012) we choose a simple proposal of $\mathcal{I}$ where a subset of the indicators is randomly selected and a change of the selected indicators is proposed, one variable at a time.

With these proposals the acceptance probability in the Metropolis Hastings algorithm is

$$\alpha[(\beta_c, \mathcal{I}_c) \to (\beta_p, \mathcal{I}_p)] = \min\left(1, \frac{p(y|\beta_p, \mathcal{I}_p)p(\beta_p|\mathcal{I}_p)p(\mathcal{I}_p)/J_1(\beta_p|\mathcal{I}_p, \beta_c)J_2(\mathcal{I}_p|\beta_c, \mathcal{I}_c)}{p(y|\beta_c, \mathcal{I}_c)p(\beta_c|\mathcal{I}_c)p(\mathcal{I}_c)/J_1(\beta_c|\mathcal{I}_c, \beta_p)J_2(\mathcal{I}_c|\beta_p, \mathcal{I}_p)}\right).$$

The proposal density for $\beta$ at the proposed point $J_1(\beta_p|\mathcal{I}_p, \beta_c)$ is the multivariate t-density with mode $\hat{\beta}$ and covariance matrix evaluated at $\hat{\beta}$, where $\hat{\beta}$ is obtained by iterating Equation (4.8) with $\beta_0 = \beta_c$. The proposal density at the current point $J_1(\beta_c|\mathcal{I}_c, \beta_p)$ is also a multivariate t density but with mode $\tilde{\beta}$ and covariance matrix evaluated at $\tilde{\beta}$, obtained from the same iteration scheme but this time from initial value $\beta_0 = \beta_p$. The proposal density for $\mathcal{I}$ at the current and proposed is the same for this simple proposal.

It is well-known that finite mixtures have identification problems because the likelihood is invariant with respect to permutations of the components. This is referred to as the label switching problem, see Frühwirth-Schnatter (2006) and Jasra et al. (2005). When estimating the predictive density this is not a problem (Geweke, 2007) but if the model is used for model based clustering one needs to proceed with caution. Plotting the MCMC samples may reveal if there was a problem with switching labels. Order conditions on the parameter space may be imposed to avoid the identification problem, see Jasra et al. (2005).

Mixture models with flexible components can have many minor local modes. It is therefore important to use a rapidly mixing MCMC scheme that avoids getting stuck in local modes. As documented in Villani et al. (2009, Section 3.3), algorithms based on variable-dimension

finite step Newton proposals are rapidly mixing, do not get stuck in local modes, and are extremely quick to localize areas of high posterior density. We have verified that our results and model evaluation (log predictive scores) do not depend on the choice of intial values in the MCMC.

4.4. **Selecting number of components.** The key quantity for selecting models in the Bayesian framework is the marginal likelihood which allows to compute Bayes factors and determine the plausibility of one model against another. However, the marginal likelihood may be sensitive to the choice of prior distribution, especially when the prior information is vague. For a general discussion see Kass (1993) and Richardson and Green (2002) in the context of mixture models.

Following Geweke and Keane (2007) and Villani et al. (2009) we therefore choose models based on the log predictive score (LPS). The LPS removes most of the dependence on the prior by sacrificing a subset of the data to train the prior to get a posterior based on the training data. If $y_{test}$ denotes the test data and $y_{train}$ the training data then the LPS is

$$p(y_{test}|y_{train}) = \int p(y_{test}|\theta)p(\theta|y_{train})\mathrm{d}\theta$$

if the test and training data are independent conditional on $\theta$, which is the case in our longitudinal setting since the entire time series for a single subject belongs to either the test or training set. To deal with the arbitrary division into training and test data, a cross validated version of the LPS is used

$$LPS = \frac{1}{B}\sum_{b=1}^{B}\log p(\tilde{y}_b|\tilde{y}_{-b}, x),$$

where $\tilde{y}_b$ is the test data in the $b$th test sample and $\tilde{y}_{-b}$ denotes the training data. Since subjects are independent conditional on the parameters

$$p(\tilde{y}_b|\tilde{y}_{-b}, x) = \int \prod_{i\in\tau_b} p(y_i|\theta, x_i)p(\theta|\tilde{y}_{-b})\mathrm{d}\theta,$$

where $\tau_b$ contains the index set of the observations in the test data for the $b$th sample. $p(\tilde{y}_b|\tilde{y}_{-b}, x)$ is easily computed by averaging $\prod_{i\in\tau_b} p(y_i|\theta)$ over the posterior draws $p(\theta|\tilde{y}_{-b})$. This requires sampling from $B$ posterior distributions based on different training data but can be done independently for each data set so computer parallelism may be exploited.

## 5. Application: modeling Firm Bankruptcy risk

5.1. **Data.** Our data set contains yearly observations for Swedish firms in the time period 1991-2008 on bankruptcy status, firm-specific variables and two macro variables. This data set has been analyzed in Jacobson et al. (2011) and Giordani et al. (2013). Jacobson et al. (2011) uses a similar approach as Shumway (2001) with a multi-period logit model extended with macro economic variables. Giordani et al. (2013) extend by modeling the log odds of the firm failure probability as a non-linear function of covariates by introducing spline functions. They show substantial improvements in predictive power as a result of accounting for nonlinearities. The present paper considers the same predictors as in Giordani et al. (2013). These are three financial ratios, two firm-specific control variables and two macroeconomic variables. The financial ratios are: *EBIT/TA* - earnings before interest and taxes over total assets (earnings ratio); *TL/TA* - total liabilities over total assets (leverage ratio) and *CH/TL* - cash and liquid assets over total liabilities (cash ratio). The control variables are: *logTS* - logarithm of deflated total sales and *logAge* - logarithm of firm age in years since first registered as a corporate. Finally the macroeconomic variables included are: *GDPG* - yearly GDP-growth rate and *Repo* - the interest rate set by Sveriges Riksbank (the Central bank of Sweden). For a thorough description of the data set, definition of bankruptcy, and other details see Giordani et al. (2013).

5.2. **Models.** Although the spline model accounts for nonlinearities in a flexible way it has some drawbacks. First, the model assumes additivity, i.e. it rules out interactions between the covariates, and the extension to spline surface models with interactions is not computationally realistic for a data set of our size. Second, it can be hard to interpret spline models as the nonlinearities are not themselves explained by other covariates. Third, it

cannot account for heterogeneity coming from missing explanatory variables. Fourth, it can be computationally demanding for moderate to large data sets when doing Bayesian inference via MCMC. This is because the dimension of the covariate space can increase dramatically after expanding in basis functions. Variable selection can be used to keep the number of effective parameters at a minimum, but increases the computational burden.

We propose to analyze bankruptcy data for Swedish firms with a finite mixture of survival models. Such models can not only account for heterogeneity and nonlinearities, but also gives an interpretation of these features in terms of covariates. A mixture model can also be used for model based clustering which gives insights about firm dynamics. The use of covariates in the mixing function is extremely useful for understanding the role of the different mixture components. Many models in the bankruptcy literature are special cases of our model. For example the models in Shumway (2001) and Jacobson et al. (2011) are obtained with $K = 1$ and $h(x_{ij}) = \frac{\exp x_{ij}^T \beta}{1+\exp x_{ij}^T \beta}$. Likewise, the model in Giordani et al. (2013) has the same structure but in addition $x$ is expanded using spline functions. It is even possible to have $K > 1$ and use splines simultaneously as in Villani et al. (2009) for the case of heteroscedastic Gaussian regression. This paper omits splines to stress the fact that the finite mixture itself can capture the non-monotonic relationships. Adding spline terms in the mixture components would also increase the computing time dramatically.

We want each firm to have a sample space $t = \{1, 2, \dots\}$. This requires covariates for each observed time period, so we are restricted to consider firms with start-up year 1991 at the earliest. The analysis can be broaden to other type of firms but then one has to consider missing data issues so this is not pursued here. Thus the population studied in the present paper consist of Swedish firms that enter the sample in the period 1991-2008. The dataset is huge with a total of $228,589$ firms with $1,670,781$ firm-year observations, on average 7.3 time-periods per firm. To speed up computing times, we shall here analyze a randomly selected subset of $11,317$ firms with $82,831$ firm-year observations, on average 7.3 time-period per firm. We are currently working on an extension of the MCMC methods

with the potential of handling essentially arbitrarily large data sets, but this will be reported elsewhere.

We estimate and compare both static and dynamic mixtures and also a one-component model with flexible baseline hazards. Two different distributions for the survival time are considered: exponential and Weibull as described in Section 3. The Weibull models are used with and without covariates in the shape parameter $\rho$. Weibull models with covariates in $\rho$ seem to be novel in the literature.

In all dynamic mixtures, exponential moving average covariates have been used to achieve persistence in component allocations over time, as described in Section 2. The choice $\alpha = 0.3$ was justified by computing for a range of values for $\alpha$ and then choose the one with highest in-sample LPS score. The choice of $\alpha$ does not affect the relative comparison between the dynamic and static models. It is also possible to estimate $\alpha$ in a separate Gibbs step, but this is not pursued here.

5.3. **Priors.** The prior for $\lambda$ is log-Normal with $E(\lambda) = 0.01405$ (the empirical hazard for another subset of the data) and $V(\lambda) = 0.05^2$ for both the exponential and Weibull model. The additional parameter $\rho$ in the Weibull model is also assigned a log-Normal prior with $E(\rho) = 1$ and $V(\rho) = 5^2$. Note that $\rho = 1$ gives the exponential model. Both priors are rather non-informative considering the scale and the log-link. The prior utilizing the Fisher information described in Section 4.1.1 is not needed in this particular example because of the enormous amounts of data, and we therefore assume prior independence between the regression coefficients for simplicity. For the mixing function the shrinkage factor $c_\gamma = 10$ gives a non-informative prior. The prior inclusion probability was set to 0.5 for each variable and in all parts of the model.

5.4. **Algorithmic considerations.** We use the Metropolis-within-Gibbs algorithm with tailored proposals and variable selection to sample from the posterior. The number of steps in the variable dimension Newton algorithm $R$ is set to 1 for the component model in all parameters and 3 for the mixing function. The degrees of freedom in the multivariate $t$ proposal is set to 10, for both the component and the mixing part of the model. Each

variable selection indicator is proposed to change with probability 0.2 in each iteration of the algorithm.

For all combinations of models in Section 5.2, $20,000$ iterations with the MCMC algorithm where performed and $5,000$ of them discarded as burn-in period, leaving $15,000$ draws from the posterior distribution. The efficiency of the sampler is measured by the inefficiency factor, which is defined as

$$IF \quad = \quad 1 + 2\sum_{l=1}^{L} \rho_l,$$

where $\rho_l$ is the autocorrelation at the $l$th lag in the MCMC chain and $L$ is an upper limit such that $\rho_l \approx 0$ when $l > L$. IF-values near 1 suggests a very efficient algorithm. We monitor convergence and measure performance using the cumulative means and IFs for the predictive mean $E(y|x)$ over a grid of $x$-values. The LPS was computed using $B = 4$ folds of the data.

5.5. **Results.** As a first attempt to investigate the fit of the models, Figure 5.1 compares the models' implied hazard function $h_t(x_t)$ as function of time to the empirical hazard rate. The models' hazard probabilities are computed for each of the firms in the panel and then averaged across all firms. The posterior uncertainty regarding the hazard is illustrated with a box plot computed from the MCMC draws. In the case of the exponential model (left column), it is clear that the one-component model gives a very poor fit to the empirical hazard, but then quickly improves as more components are added to the model. A two-component exponential model gives a similar estimated hazard as a one-component model with flexible baseline hazards (top right). The one-component Weibull model without covariates in the shape parameter $\rho$ produces a similar hazard as the one-component exponential model, but by adding covariates in $\rho$ the Weibull model can capture the non-monotonic relationship of the empirical hazard fairly well.

The assessment of model fit in Figure 5.1 is visually appealing, but is very much a rather limited marginal view of the data. Table 1 reports the LPS for static and dynamic mixtures, using either exponential or Weibull components, with and without covariates in the Weibull

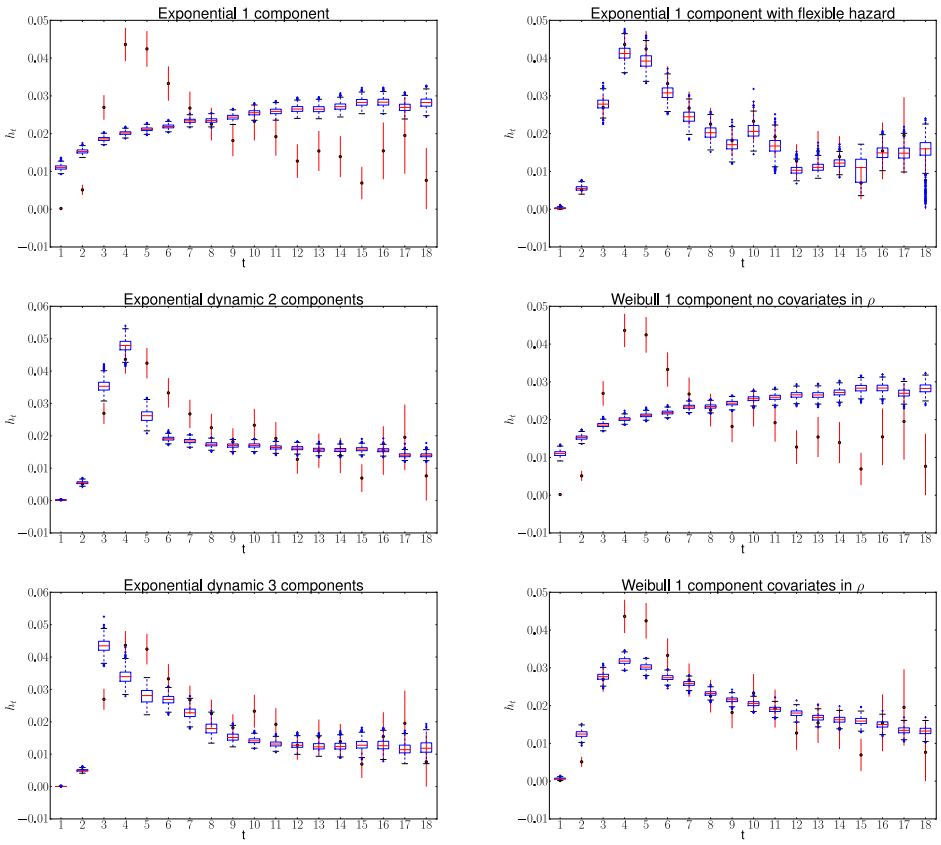


FIGURE 5.1. Hazard as a function of time for some models (box-plots) plotted against the empirical hazard (red vertical lines).

TABLE 1. Log Predictive Score (LPS) for the static and dynamic mixtures computed using 4-fold cross-validation. The best model for a given number of components are in bold typeface.

| **Static mixtures** | **Comp 1** | **Comp 2** | **Comp 3** |
|---|---|---|---|
| Exponential | -1784.83 | -1712.16 | -1687.54 |
| Weibull | -1785.17 | -1725.08 | -1683.60 |
| Weibull covariates in $\rho$ | -1696.96 | -1652.78 | -1648.73 |
| | | | |
| **Dynamic mixtures** | **Comp 1** | **Comp 2** | **Comp 3** |
| Exponential | -1784.83 | -1618.51 | -1570.20 |
| Weibull | -1785.17 | -1605.00 | -1561.97 |
| Weibull covariates in $\rho$ | -1696.96 | **-1585.07** | **-1553.43** |
| | | | |
| Exponential Flex Baseline | **-1686.41** | | |

shape parameter. The most striking result in Table 1 is the dramatically better out-of-sample predictive performance of the dynamic mixtures compared to their static counterparts. As an example, the three-components dynamic mixture of exponentials is 117.34 LPS units better than the three-components static mixture of exponentials. Table 1 also reports the LPS of the one-component exponential model with a free baseline hazard parameter estimated for each year. Using a flexible baseline hazard clearly improves the LPS, but also this model is clearly outperformed by the dynamic mixtures with $K > 1$. This suggests that these data are truly heterogeneous even after controlling for age and size effects and different baseline hazards.

Another interesting observation from Table 1 is that the LPS for the Weibull model improves considerably when allowing for covariates in both model parameters. This is true for models with multiple components as well. Covariates in the shape parameter of the Weibull is rare or perhaps even non-existent in practical work, but this is clearly an extension that should be considered.

In all models, the LPS improves for each added component but the rate of improvement decreases. It is worthwhile to mention that variable selection implies that adding components does not necessarily give a more complex model. See the Lidar example in Li et al. (2011) for a clear demonstration of how variable selection in mixture-of-experts models can be a very effective guard against overfitting.

To illustrate some of the interpretations of our models, Tables 2-4 presents parameter estimates for some selected one- and two-component models. Data have been standardized to have zero mean and unit variance for all covariates, hence all parameter estimates are on the same scale. The posterior mean and standard deviation are computed conditional on the covariate belonging to the model.

Starting with the results for the one component exponential model in Table 2, we see that the most significant variables are cash, age, earnings, and leverage, all with a posterior inclusion probability of unity. The variable selection effectively removes size, GDPG, and to

TABLE 2. Estimation results for exponential model with one component. IF: min = 0.55, median = 1, max = 1.70.

| **Component 1** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Intercept | Earnings | Leverage | Cash | Size | Age | GDPG | Repo |
| **Post Mean** | -4.397 | -0.258 | 0.25 | -1.174 | -0.02 | 0.399 | 0.057 | 0.1 |
| **Post Std** | 0.045 | 0.019 | 0.018 | 0.11 | 0.023 | 0.033 | 0.033 | 0.029 |
| **Post Incl Prob** | - | 1 | 1 | 1 | 0.012 | 1 | 0.051 | 0.837 |
| **Mean Acc Prob** | 0.404 | | | | | | | |

TABLE 3. Estimation results for Weibull model with one component and covariates in both parameters. IF: min = 0.92, median = 7.01, max = 14.5.

| **Parameter $\lambda$** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Intercept | Earnings | Leverage | Cash | Size | Age | GDPG | Repo |
| **Post Mean** | -4.985 | -0.265 | 0.213 | 0.81 | -0.04 | 1.971 | -0.02 | 0.057 |
| **Post Std** | 0.244 | 0.022 | 0.02 | 0.073 | 0.026 | 0.118 | 0.026 | 0.028 |
| **Post Incl Prob** | - | 1 | 1 | 1 | 0.029 | 1 | 0.014 | 0.063 |
| **Mean Acc Prob** | 0.712 | | | | | | | |
| **Parameter $\rho$** | | | | | | | | |
| | Intercept | Earnings | Leverage | Cash | Size | Age | GDPG | Repo |
| **Post Mean** | 0.231 | 0.021 | -0.013 | -1.07 | -0.008 | -0.782 | -0.003 | 0.015 |
| **Post Std** | 0.093 | 0.015 | 0.014 | 0.071 | 0.008 | 0.04 | 0.009 | 0.009 |
| **Post Incl Prob** | - | 0.007 | 0.01 | 1 | 0.002 | 1 | 0.004 | 0.016 |
| **Mean Acc Prob** | 0.788 | | | | | | | |

some extent Repo. In this model, a positive sign corresponds to increased hazard probability as a variable increases, and vice versa.

Moving to the dynamic mixture of two exponential components in Table 4 it is evident that the most significant covariates in the mixing function are age and cash. There is also a posterior inclusion probability of 1 for GDPG and Repo, but the magnitude of their effects are smaller. This means that the separation of the data into the two different classes is mostly determined by age and cash. Our parametrization is such that when age increases it is more likely to belong to the first component and the same holds for cash. To illustrate the interpretation of the mixture models, let us consider a newly founded firm. Since a newly founded firm is by definition of low age, such a firm tends to belongs to the second component

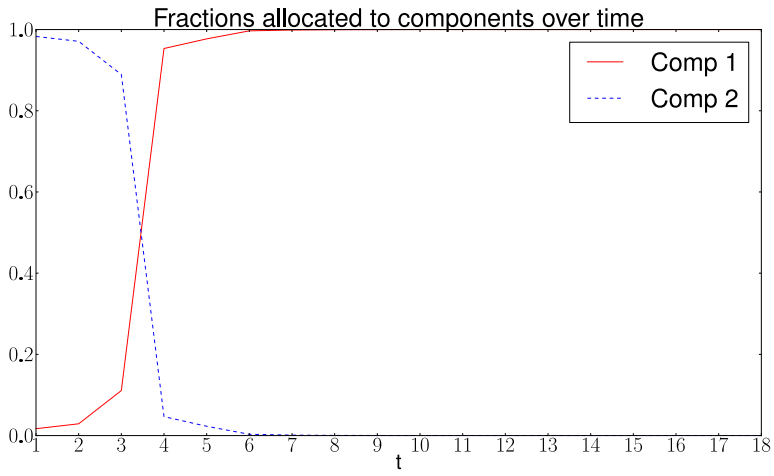


FIGURE 5.2. Fraction allocated to respective component over time for the dynamic exponential mixture.

with a large probability, everything else equal. Since age has a large positive coefficient in the second component, this young firm will initially experience a rapidly increasing hazard as it grows older. If the firm manages to survive the early years, it will eventually move over to the first mixture component where age is no longer a significant determinant of the hazard. The firm has managed to survive the first risky years and can now grow older without accelerating risk on account of its age. Figure 5.2 shows the posterior allocation of firms over time: firms that have survived for a long time are classified to component 1 in their later time periods, while firms in early time periods are classified to the second component.

Cash has a similar interpretation as age: with a large probability, a firm with low cash belongs to the second component where the coefficient on cash is strongly negative. This means that a low cash firm can drastically reduce the bankruptcy probability by increasing its holdings of cash. As the firm continues to improve its liquidity, it will eventually reach a point where it switches over to the first component. In this component, cash remains a positive factor for decreasing bankruptcy risk, but its effect is much smaller. Note however that this interpretation is only valid if holding of cash is exogenous.

To further explore the difference between the static and dynamic mixtures we plot the overall predictive hazard $h_t(x_t)$ in Figure 5.3 over the first four years for a firm that is born in the beginning of the sample period, i.e. 1991. Each subgraph shows the predictive hazard

TABLE 4. Estimation results for a dynamic exponential model with two components. Covariates in the mixing function are exponentially moving averages. Parameters in the mixing function corresponds to $P(s_t = 2|z_t)$. IF: min = 0.84, median = 1.23, max = 110.84.

| **Component 1** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Intercept | Earnings | Leverage | Cash | Size | Age | GDPG | Repo |
| **Post Mean** | -4.311 | -0.251 | 0.339 | -0.559 | 0.033 | 0.096 | -0.024 | 0.039 |
| **Post Std** | 0.058 | 0.026 | 0.023 | 0.108 | 0.035 | 0.086 | 0.05 | 0.067 |
| **Post Incl Prob** | - | 1 | 1 | 1 | 0.016 | 0.055 | 0.016 | 0.026 |
| **Mean Acc Prob** | 0.751 | | | | | | | |
| **Component 2** | | | | | | | | |
| **Post Mean** | -2.522 | -0.367 | 0.042 | -2.873 | -0.004 | 4.544 | -0.066 | 0.044 |
| **Post Std** | 0.238 | 0.036 | 0.053 | 0.501 | 0.052 | 0.237 | 0.045 | 0.043 |
| **Post Incl Prob** | - | 1 | 0.027 | 1 | 0.013 | 1 | 0.04 | 0.019 |
| **Mean Acc Prob** | 0.782 | | | | | | | |
| **Mixing** | | | | | | | | |
| **Post Mean** | -4.777 | -0.113 | 0.031 | -1.698 | 0.039 | -8.296 | 0.788 | 0.735 |
| **Post Std** | 0.496 | 0.088 | 0.094 | 0.39 | 0.089 | 0.815 | 0.184 | 0.196 |
| **Post Incl Prob** | - | 0.092 | 0.067 | 1 | 0.066 | 1 | 1 | 1 |
| **Mean Acc Prob** | 0.835 | | | | | | | |

$h_t(x_t)$ as a function of the covariate cash for a given year. The analysis in Figure 5.3 is conditioned on fixed paths for the other covariates. We have chosen to set the covariate paths for Repo, GDPG and age as the realized values at each time point but with a one year lag for repo and GDPG; when predicting bankruptcy at period $t$, macro variables from $t-1$ are used. For the financial ratios and the size variable, the average covariate value in the sample for each respective year is used as conditioning paths. The covariate paths together with their moving averages are presented in Figure 5.4. This example clearly illustrates the main difference in these models; the dynamically evolving proportions in the dynamic mixture (left panel) gives a much more flexible hazard than the static mixture (right panel) where the mixture weights have a much more restrictive form, thus not allowing the same flexibility.

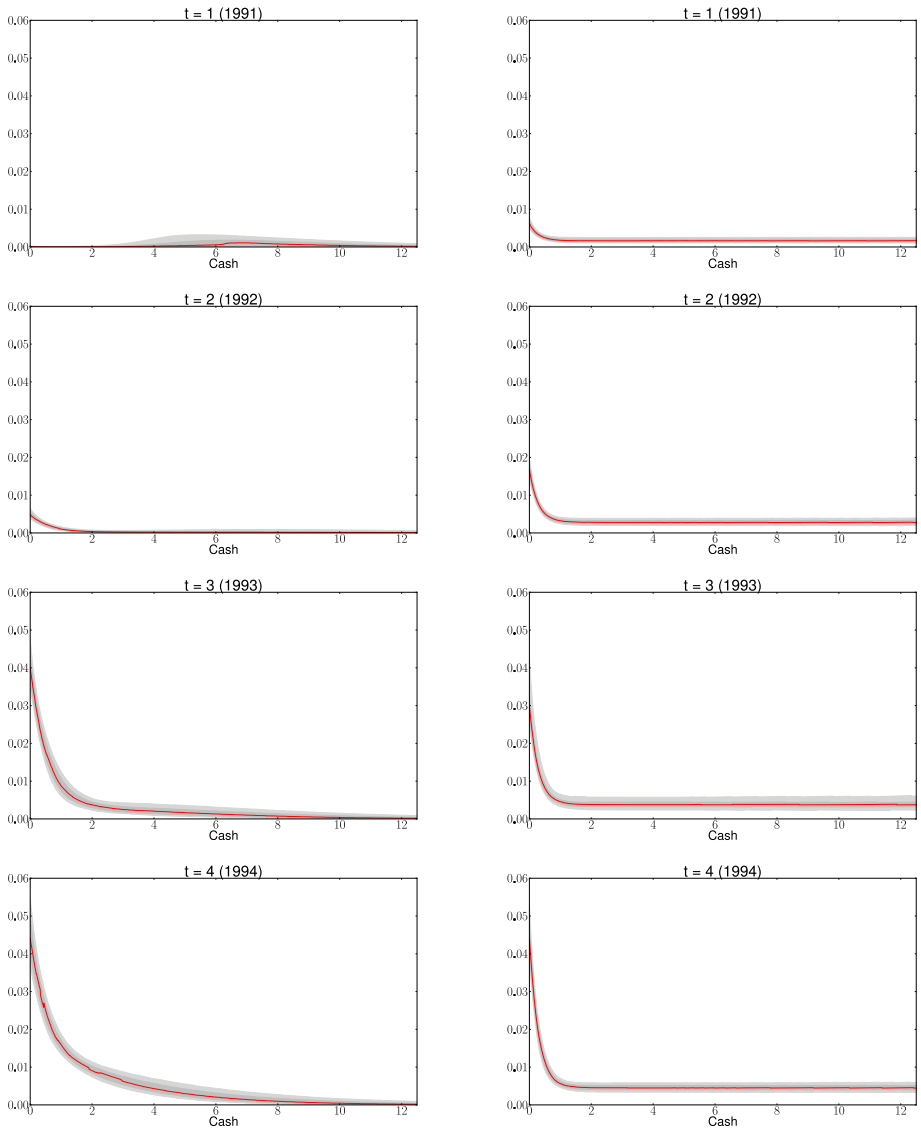


FIGURE 5.3. Posterior distribution of the hazard probability of the representative firm as a function of cash for a dynamic (left panel) and a static (right panel) exponential mixture with two components for $t = 1, \ldots, 4$. The dark shaded area corresponds to 68% Highest Posterior Density (HPD) regions and lighter shaded area are the 95% HPD regions. The red solid line is the posterior mode.

## 6. CONCLUSIONS

We propose flexible smooth mixture models for longitudinal data, with special emphasis on models for survival data in discrete time. We discuss how the longitudinal dimension opens up for two different types of mixture models, the static and dynamic mixture. In the static mixture, subjects have to remain in the same component in all time periods,

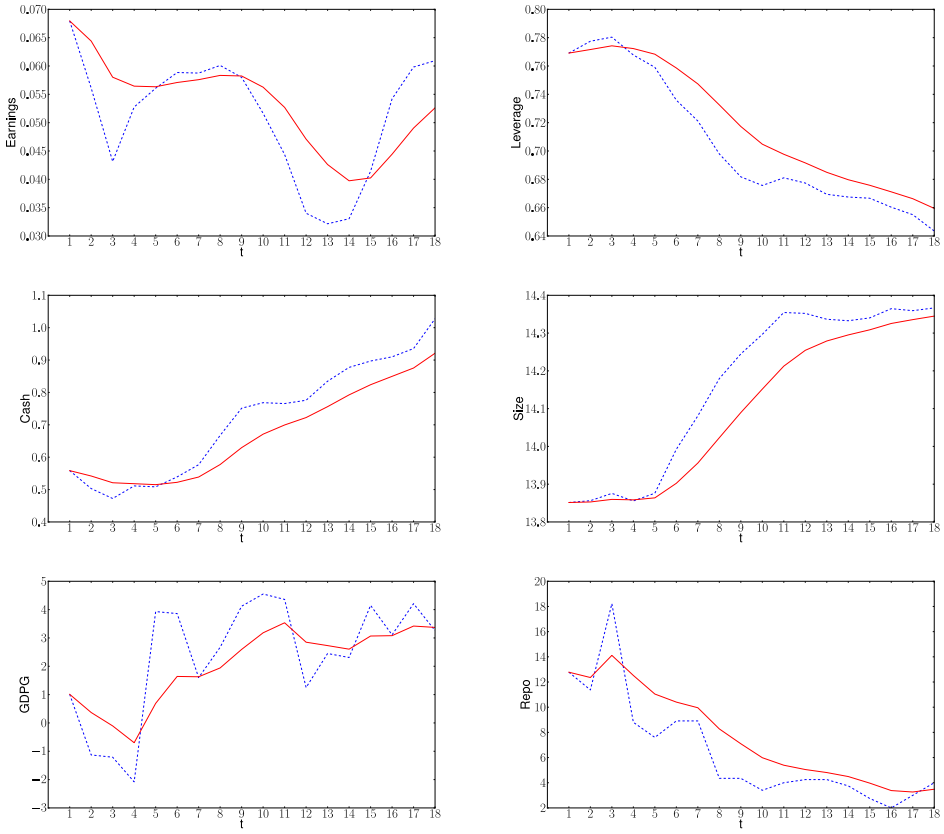


FIGURE 5.4. Covariate paths for a representative firm. The dashed blue line corresponds to realized covariates and the solid red line are the exponentially moving averages with $\alpha = 0.3$.

whereas in the dynamic mixture they can move between mixture components over time. We argue that the obvious Markov transition model would be prohibitively time-consuming for datasets with a large number of subjects, and we propose an alternative approach where the within-subject dynamics is determined by subject-specific time-varying covariates. We prove that the proposed longitudinal dynamic mixture model with sufficiently many components can approximate a large class of models.

We compare the static and dynamic mixtures in bankruptcy modeling for a large panel of Swedish firms over the time period 1991-2008. The main result is that the dynamic mixture formulation dramatically outperforms the static mixture, a result that holds both when exponential or Weibull mixture components are used. We also show that the MCMC algorithm with variable selection in Villani et al. (2012) can be straightforwardly extended

to the longitudinal case and we document a high MCMC efficiency in our application to firm bankruptcy.

It is also shown that the firm bankruptcy data are heterogeneous even after the standard firm specific variables in the literature are included in the model and when a flexible baseline hazard is used. This result suggests that there are different classes of firms and the effect of the covariates on the hazard probability is different in each class. Furthermore, it is also shown that model with multiple classes is able to generate a non-monotonic hazard function which agrees with the empirical hazard and also with models that uses a flexible baseline hazard with a separate parameter for each time period.

Although our way of modeling within-subject dynamics by mixture-of-experts with time-varying mixing covariates is computationally attractive in comparison to other standard approaches, data sets with several millions of observations remain a challenge. We are currently working on extensions of the MCMC algorithm presented here that may reduce computing times substantially for large data sets, see Quiroz et al. (2015). In terms of model extensions it would be interesting to explore the role of a continuous frailty in the components. The hierarchical structure of such a model requires two extra steps in the MCMC scheme; sampling the frailty and the parameters in its distribution. This is in principle straightforward, but will add to the computing time, which again requires innovations in the MCMC methodology.

## 7. Acknowledgments

The authors thank the associated editor and two anonymous referees for their comments which helped improving the paper. Matias Quiroz was partially supported by VINNOVA grant 2010-02635.

## References

Allison, P. (1982). Discrete-time methods for the analysis of event histories. *Sociological methodology*, 13(1):61–98.

Baum, L. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563.

Carling, K., Edin, P.-A., Harkman, A., and Holmlund, B. (1996). Unemployment duration, unemployment benefits, and labor market programs in sweden. *Journal of Public Economics*, 59(3):313–334.

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

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

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208.

Franzén, J. (2008). *Bayesian Cluster Analysis: Some Extensions to Non-standard Situations*. PhD thesis, Stockholm University, Department of Statistics.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of time series analysis*, 15(2):183–202.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer-Verlag.

Frühwirth-Schnatter, S. (2011). Panel data analysis: a survey on model-based clustering of time series. *Advances in Data Analysis and Classification*, 5(4):251–280.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7(1):57–68.

Geweke, J. (2007). Interpretation and inference in mixture models: Simple mcmc works. *Computational Statistics & Data Analysis*, 51(7):3529–3550.

Geweke, J. and Keane, M. (2007). Smoothly mixing regressions. *Journal of Econometrics*, 138(1):252–290.

Giordani, P., Jacobson, T., Von Schedvin, E., and Villani, M. (2013). Taking the twists into account: Predicting firm bankruptcy risk with splines of financial ratios. *Journal of Financial and Quantitative Analysis*, forthcoming.

Huynh, K. and Voia, M. (2009). Mixed proportional hazard models with finite mixture unobserved heterogeneity: An application to nascent firm survival. Manuscript.

Ibrahim, J., Chen, M., and Sinha, D. (2005). *Bayesian survival analysis*. Wiley Online Library.

Jacobs, R., Jordan, M., Nowlan, S., and Hinton, G. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Jacobson, T., Lindé, J., and Roszbach, K. (2011). Firm default and aggregate fluctuations. *Journal of European Economic Association*, forthcoming:to appear.

Jasra, A., Holmes, C., and Stephens, D. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20(1):50–67.

Jiang, W. and Tanner, M. (1999). Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *The Annals of Statistics*, 27(3):987–1011.

Jordan, M. and Jacobs, R. (1994). Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214.

Kass, R. (1993). Bayes factors in practice. *The Statistician*, 42(5):551–560.

Kim, C.-J. and Nelson, C. R. (2003). State-space models with regime switching: classical and gibbs-sampling approaches with applications. *MIT Press Books*, 1.

Kohn, R., Smith, M., and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, 11(4):313–322.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica: Journal of the Econometric Society*, 47(4):939–956.

Li, F., Villani, M., and Kohn, R. (2011). Modeling conditional densities using finite smooth mixtures. In Mengersen, K., Robert, C., and Titterington, M., editors, *Mixtures: estimation and applications*, pages 123–144. John Wiley & Sons.

McLachlan, G., McGiffin, D., et al. (1994). On the role of finite mixture models in survival analysis. *Statistical methods in medical research*, 3(3):211.

Miller, R., Gong, G., and Muñoz, A. (1981). *Survival analysis*. Wiley New York.

Mosler, K. (2003). Mixture models in econometric duration analysis. *Applied Stochastic Models in Business and Industry*, 19(2):91–104.

Muthén, B. and Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics*, 30(1):27–58.

Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *The Annals of statistics*, 38(3):1733–1766.

Nott, D. and Leonte, D. (2004). Sampling schemes for bayesian variable selection in generalized linear models. *Journal of Computational and Graphical Statistics*, 13(2):362–382.

Ntzoufras, I., Dellaportas, P., and Forster, J. (2003). Bayesian variable and link determination for generalised linear models. *Journal of statistical planning and inference*, 111(1):165–180.

Qi, Y. and Minka, T. (2002). Hessian-based markov chain monte-carlo algorithms. Manuscript.

Quiroz, M., Villani, M., and Kohn, R. (2015). Speeding up mcmc by efficient data subsampling. *arXiv preprint arXiv:1404.4178v2*.

Richardson, S. and Green, P. (2002). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model*. *The Journal of Business*, 74(1):101–124.

Singer, J. and Willett, J. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational and Behavioral Statistics*, 18(2):155–195.

Vaupel, J., Manton, K., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Villani, M., Kohn, R., and Giordani, P. (2009). Regression density estimation using smooth adaptive gaussian mixtures. *Journal of Econometrics*, 153(2):155–173.

Villani, M., Kohn, R., and Nott, D. (2012). Generalized smooth finite mixtures. *Journal of Econometrics*, 171(2):121–133.

Xue, X. and Brookmeyer, R. (1997). Regression analysis of discrete time survival data under heterogeneity. *Statistics in medicine*, 16(17):1983–1993.

## Appendix A. On the flexibility of the dynamic mixture

**Definition 1.** A model for the time sequence $y_{1:n} = (y_1, \ldots, y_n)$ is called a *longitudinal p-lag GLM* if its joint distribution is of the form

$$p(y_{1:n}|x_{1:n}, \theta) \quad = \quad \prod_{j=1}^{n} p(y_j|y_{j-p:j-1}, x_j, \theta)$$

where

$$p(y_j|y_{j-p:j-1}, x_j, \theta) \quad = \quad \exp\left[a(h_j)y_j + b(h_j) + c(y_j)\right]$$

for known analytic functions $a(\cdot), b(\cdot)$ and $c(\cdot)$ with non-zero derivatives on $\mathbb{R}$. Furthermore, $h_j = h_j(x_j, y_{j-p:j-1}) = \alpha + \beta^T x_j + \eta^T y_{j-p:j-1}$ and $\theta = (\alpha,\beta,\eta)$. The conditional mean is $E[y_j|y_{j-p:j-1}, x_j, \theta] = \Psi^{-1}(h_j)$ for some smooth invertible link function $\Psi(\cdot)$. We use the shorthand notation $y_{1:n}|x_{1:n} \sim LGLM(a, b, c, h_{1:n}, \Psi, \theta, p)$.

The proposed dynamic mixture approximates the class of target distributions

$$y_{1:n}|x_{1:n} \quad \sim \quad LGLM(a, b, c, \tilde{h}_{1:n}, \Psi, \theta, p)$$

where $\tilde{h}_j$ is more flexible than $h_j = \alpha + \beta^T x_j + \eta^T y_{j-p:j-1}$. $\tilde{h}_j$ is essentially any non-linear function with continuous second derivatives. From a technical point of view, this choice of $\tilde{h}_j$ gives that the conditional mean in the target class belongs to a Sobolev space transformed by the inverse link function, see Jiang and Tanner (1999, p. 992) for details. We denote this

class by $SLGLM(a, b, c, h_{1:n}, \Psi, \theta, p)$ where $S$ stands for smooth and the tilde notation on $h$ is suppressed. We now introduce the approximating class.

**Definition 2.** Let $g_K$ be an approximation of $f \in SLGLM(a, b, c, h_{1:n}, \Psi, \theta, p)$. The approximator $g_K$ is a dynamic mixture of $K$ $LGLM(a, b, c, h_{1:n}, \Psi, \theta_l, p)$ experts with joint distribution

$$g_K(y_{1:n}|x_{1:n}) = \prod_{j=1}^{n} \left( \sum_{l=1}^{K} w_l(z_j, \gamma_l) p(y_j|y_{j-p:j-1}, x_j, \theta_l) \right)$$

where $z_j = (x_j, y_{j-p:j-1})^T$ and

$$w_l(z_j, \gamma_l) = \frac{\exp\left(z_j^T \gamma_l\right)}{\sum_{m=1}^{K} \exp\left(z_j^T \gamma_m\right)} \text{ with } \gamma_1 = 0$$

for identification.

We will prove that $g_K$ approximates any $f \in SLGLM(a, b, c, h_{1:n}, \Psi, \theta, p)$ arbitrarily close in the Kullback-Leibler distance as the number of components increase. We need the following lemma.

**Lemma 3.** *Let $f = f(y_{1:n})$ and $g = g(y_{1:n})$ be two joint distributions. The Kullback-Leibler (KL) distance between $f$ and $g$ can be expressed as*

$$KL(f, g) = KL(f_1, g_1) + E_1 KL(f_{2|1}, g_{2|1}) + \cdots + E_{1:n-1} KL(f_{n|1:n-1}, g_{n|1:n-1})$$

*where*

$$KL(f_{j|1:j-1}, g_{j|1:j-1}) = \int_{\mathbb{R}} f_{j|1:j-1} \log \left( \frac{f_{j|1:j-1}}{g_{j|1:j-1}} \right) dy_j,$$

*with $h_{n|1:n-1} = h(y_n|y_{1:n-1})$ and $E_{1:j}$ denotes the expectation with respect to $h(y_{1:j})$.*

*Proof.* Theorem 2.5.3 in Cover and Thomas (2012) proves the lemma for $p = 2$. Repeated application gives the general result for $n$ variables. □

**Theorem 4.** *Let $f$ be the joint distribution of the target model*

$$y_{1:n}|x_{1:n} \sim SLGLM(a, b, c, h_{1:n}, \Psi, \theta, p).$$

*Let $g_K$ be the joint distribution of the approximating dynamic mixture with $K$ components as defined in Definition 2, with the parameters estimated by maximum likelihood. Then*

$$KL(f, g_K) \quad = \quad \frac{c}{K^{4/s}}$$

*for any $f$ in the SLGLM class, where $c$ is a constant independent of $K$.*

*Proof.* From Lemma 3 and the $p$-lag structure it follows that

$$KL(f, g_k) \quad = \quad KL(f_1, g_1^K) + E_1 KL(f_{2|1}, g_{2|1}^K) + \cdots + E_{n-p:n-1} KL(f_{n|n-p:n-1}, g_{n|n-p:n-1}^K).$$

Now, for any $j$, $f_{j|j-p:j-1}$ is a (non-longitudinal) GLM with a Sobolev smooth mean function and therefore belongs to the target class in Jiang and Tanner (1999). Furthermore $g_{j|j-p:j-1}^K$ is a (non-longitudinal) approximator for $f_{j|j-p:j-1}$ of the form

$$g_{j|j-p:j-1}^K \quad = \quad \sum_{l=1}^{K} w_l(z_j, \gamma_l) p(y_j|z_j, \theta_l)$$

with $z_j = (x_j, y_{j-p:j-1})^T \in s \times 1$. Hence $g_{j|j-p:j-1}^K$ has the form as in Equation (2.4) in Jiang and Tanner (1999, p. 992). By Theorem 2 in Jiang and Tanner (1999) it follows that

$$KL(f, g_K) \quad = \quad \frac{c_1}{K^{4/s}} + \frac{E_1[c_2]}{K^{4/s}} + \cdots + \frac{E_{n-p:n-1}[c_n]}{K^{4/s}} = \frac{c}{K^{4/s}}$$

where $s$ is the number of covariates (including the lags of $y$), $c_j$ is a constant independent of $K$ and $c = c_1 + E_1[c_2] + ... + E_{n-p:n-1}[c_n]$. Under the assumption that $E_{j-p:j-1}[c_j] < \infty$ for $j = 2, \ldots, n$ the proof is completed. □

*Remark.* When $y$ is continuous one can prove an alternative version of Theorem (4) using the approximation results in Norets (2010) instead of Jiang and Tanner (1999). Norets (2010)

result is derived under more general conditions and holds for a general class of (continuous) target densities.