Hybrid Hidden Markov Model and Generalized Linear Model for Auto Insurance Premiums

Lucas Berry

A Thesis
for The Department of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Arts (Mathematics) at
Concordia University
Montreal, Quebec, Canada

December 2016

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Lucas Berry**

Entitled: **Hybrid Hidden Markov Model and Generalized Linear Model for Auto Insurance Premiums**

and submitted in partial fulfillment of the requirements for the degree of

Master of Arts (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Examiner

Dr. L. Popovic

_____ Examiner

Dr. Y. Yang

_____ Thesis Supervisor

Dr. J. Garrido

_____ Thesis Supervisor

Dr. M. Mailhot

Approved by _____

Chair of Department or Graduate Program Director

_____

Dean of Faculty

Date _____

Abstract

Hybrid Hidden Markov Model and Generalized Linear Model for Auto Insurance Premiums

Lucas Berry

We describe a new approach to estimate the pure premium for automobile insurance. Using the theory of hidden Markov models (HMM) we derive a Poisson-gamma HMM and a hybrid between HMMs and generalized linear models (HMM-GLM). The hidden state is meant to represent a driver's skill thus capturing an unseen variable. The Poisson-gamma HMM and HMM-GLM have two emissions, severity and claim count, making it easier to compare to current actuarial models. The proposed models help deal with the overdispersion problem in claim counts and introduces dependence between the severity and claim count. We derive maximum likelihood estimates for the parameters of the proposed models and then using simulations with the Expectation Maximization algorithm we compare the three methods: GLMs, HMMs and HMM-GLMs. We show that in some instances the HMM-GLM outperforms the standard GLM, while the Poisson-gamma HMM under-performs the other models. Thus in certain situations it may be worth the added complexity of a HMM-GLM.

Acknowledgements

I would like to thank my supervisors, Dr. Jose Garrido and Dr. Melina Mailhot. Their patience and comments have been instrumental to the formation of my thesis. Also their guidance and encouragement was extremely helpful during my master's degree.

My regards go out to all my professors during my master's and undergraduate degree. Unbeknownst to them, their courses have greatly impacted my research, many of my ideas came from their courses.

Thank you to Dr. Lea Popovic and Dr. Yi Yang for reviewing this thesis.

My deepest gratitude goes to Clara Lacroce for providing me with someone to discuss my ideas with and for reviewing a large portion of my thesis.

I would also like to thank my family and friends. During this process they have encouraged me and provided needed distractions at time.

This work would not have been possible without the financial support of the Mathematics and Statistics Department of Concordia University and the Institut des sciences mathématiques.

Contents

List of Figures

List of Tables

Introduction

Auto insurance makes up a large portion of the Property and Casualty realm. Accurate predictions of future automobile claims is crucial for a company's survival. The most common techniques rely on generalized linear models (GLMs). Actuaries typically forecast two quantities, the number of claims or claim count and the average cost of a claim or claim severity. Claim counts are usually modeled with a Poisson distribution and claim severity with a gamma distribution. Commonly included in both models is a covariate that attempts to capture a driver's skill. This is something typically unseen by the actuary and thus is hard to determine. In this thesis we propose a Poisson-gamma hidden Markov model and a combination of an HMM and a generalized linear model (HMM-GLM) to capture this effect.

Chapter 1 gives a quick overview of classical linear regression, the exponential dispersion family and generalized linear models. Included is the process for deriving the maximum likelihood estimates and more advanced topics.

Chapter 2 provides an introduction to hidden Markov models and the expectation maximization (EM) algorithm, which is commonly used to estimate the parameters of an HMM. For HMMs the EM algorithm is also known as the Baum-Welch algorithm.

Chapter 3 is a description of the proposed Poisson-gamma HMM and the estimators for the parameters according to the expectation maximization algorithm. The estimators were derived assuming an actuary has one or multiple observation sequences. Also contained is the mathematical methodology for making future forecasts.

Chapter 4 introduces the HMM-GLM and goes over the derivation of the estimators of the model parameters. Again the model was considered for one or multiple observation sequences.

Chapter 5 provides three simulation studies each with different underlying assumptions.

A comparison of the models is also included. Before the results is an overview of implementation issues, the scaling problem for observation sequences with too long time horizons and the issue with zero claim counts.

Chapter 1

Linear Models

1.1 Introduction

Generalized Linear Models (GLMs) are an essential part of a statistician's tool box, one might say the hammer because it is used with such frequency. This tool has thus made its way into insurance and today is one of the most widely used methods to forecast losses. Thus there exists many research papers extending previous concepts to tackle different problems. GLMs themselves are an extension of the linear regression model. This chapter reviews the main GLM concepts with the help of De Jong et al. [2008].

1.2 Classical Linear Regression

Linear regression is taught in most first year Statistics courses and is the foundation for GLMs. Thus this topic will be covered first. Linear regression attempts to explain a relationship between a response variable, y, and a linear combination of explanatory variables, x_i's. As many fields use linear regression there exist different names for the response y, such as dependent variable, outcome, output, or target. The x_i's suffer from the same problem where covariates, independent variables, inputs, risk factors, features, or predictors. Note this can make it cumbersome when speaking to experts from different fields. To fit a model one needs a matrix $\mathbf{X}$, $n \times (p+1)$, containing data for the recorded explanatory variables and a vector $\mathbf{y}$, $n \times 1$, containing the recorded response variables. Here n refers to the number of

observations and p refers to the number of explanatory variables, the $+1$ is for the intercept term. The intercept term denotes the predicted value of the response given zeros for the explanatory variables. Note that future response values are often unobservable and therefore useful to model. In other cases the predicted value is compared to the current y value and then leveraged to gain a competitive edge. The classical linear regression model is written as

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon, \qquad \epsilon \sim N(0, \sigma^2). \tag{1.1}$$

Here the β_i's are the coefficients to the explanatory variables and ϵ is an error term. Note one can transform the explanatory variables however one likes, such as x_i^b, where $b \in \mathbb{R}$. Also one can include interaction terms, $x_i^b x_j^c$, where $b, c \in \mathbb{R}$ and $i \neq j$.

After choosing the appropriate explanatory variables one must derive estimates of the β's. Let $\boldsymbol{\beta}$ be a $(p+1) \times 1$ vector containing the coefficients. Thus having:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

In classical linear regression to estimate $\boldsymbol{\beta}$ one must minimize the least squared error,

$$S = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{1.2}$$

This optimization problem can be solved using calculus:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

With this solution and new data points one can then move on to forecasting. To arrive at a

point estimate solution one must take the expectation of (1.1),

$$\mathrm{E}(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p.$$

When making important decisions, practitioners often take this value under consideration. It is unlikely for future y values to be precisely equal to their point estimates and thus it helps to derive a confidence interval for y. Thankfully given the model assumptions we know the distribution of $\mathbf{y}$ given $\mathbf{X}$,

$$\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I). \tag{1.3}$$

When applying classic linear regression there are four assumptions that are implicit:

1. E(ϵ)=0.
2. Homoskedasticity. The variance of ϵ is constant and does not vary with different values of the explanatory variables.
3. Normality. ϵ is normally distributed.
4. Uncorrelation. Each observation is independent of every other observation or at least uncorrelated.

Each one of these assumptions can be tested, for the appropriate tests refer to Chapter 4 of De Jong et al. [2008].

The simplicity of classical linear regression has aided its popularity. Linear regression models are used across many different fields to make informed decisions. These models are one of the building blocks for GLMs. Please note there is an endless supply of literature on linear regression. Therefore there exists many more tests and different ways to improve one's model, if interested consult De Jong et al. [2008].

1.3 Exponential Family

The exponential dispersion family of distributions is another key building block to GLMs. Any density or probability function that can be written as

$$f(y) = c(y, \tau)\exp\left\{\frac{y\zeta - a(\zeta)}{\tau}\right\}, \tag{1.4}$$

where ζ and τ are parameters, denotes a member in the exponential dispersion family for the variable y. Note that ζ is referred to as the canonical parameter and τ as the dispersion parameter. Functions $a(\zeta)$ and $c(y, \tau)$ determine the exact density/probability function of y. Distributions that can be written in this form have the following nice properties.

From (1.4) one can show that $\frac{\partial a(\zeta)}{\partial \zeta} = \mathrm{E}(y)$. Let $\dot{a}(\zeta)$ be the partial derivative of a with respect to ζ. Then

$$\frac{\partial f(y)}{\partial \zeta} = f(y)\left[\frac{y - \dot{a}(\zeta)}{\tau}\right].$$

Integrating both sides over y gives

$$0 = \frac{\mathrm{E}(y) - \dot{a}(\zeta)}{\tau}. \tag{1.5}$$

The left side is zero because

$$\int \frac{\partial f(y)}{\partial \zeta} = \frac{\partial}{\partial \zeta}\int f(y)dy = \frac{\partial}{\partial \zeta}1 = 0.$$

Thus for the right side of (1.5) to equal zero the numerator must equal zero which implies that $\dot{a}(\zeta) = \mathrm{E}(y)$. Also from (1.4) one can show that $\frac{\partial^2 a(\zeta)}{\partial \zeta^2} = \frac{\mathrm{V}(y)}{\tau}$, where $\mathrm{V}(y)$ is the variance of y. Let $\ddot{a}(\zeta)$ be the second partial derivative of a with respect to ζ. Then

$$\frac{\partial^2 f(y)}{\partial \zeta^2} = f(y)\left[\frac{y - \dot{a}(\zeta)}{\tau}\right]^2 - f(y)\frac{\ddot{a}(\zeta)}{\tau}.$$

Integrating both sides over y yields

$$0 = \frac{\mathrm{E}[(y - \dot{a}(\zeta))^2]}{\tau^2} - \frac{\ddot{a}(\zeta)}{\tau}.$$

The left side is zero by the same logic as before, switching the order of integration and differentiation. Using our previous result, $\dot{a}(\zeta) = \mathrm{E}(y)$,

$$\frac{\mathrm{E}[(y - \mathrm{E}(y))^2]}{\tau^2} - \frac{\ddot{a}(\zeta)}{\tau} = \frac{\mathrm{V}(y)}{\tau^2} - \frac{\ddot{a}(\zeta)}{\tau}.$$

Thus for this difference to equal zero $\ddot{a}(\zeta)$ must equal $\frac{\mathrm{V}(y)}{\tau}$, where $\ddot{a}(\zeta)$ is known as the variance function. Note that the calculations above assume that the derivative and integral are interchangeable on the left-hand side of the equation.

These properties are very useful but a distribution must be member of the exponential family first. Using (1.4) one can derive a new form for the exponential family,

$$\ln[f(y)] = \ln[c(y, \tau)] + \frac{y\zeta - a(\zeta)}{\tau}. \tag{1.6}$$

Next rewriting distributions in this new form shows members of the exponential family. The following two examples show in detail why the Poisson and gamma distributions are part of the exponential family. Let $y \sim \mathrm{Poi}(\lambda)$ and f represent the probability mass function then

$$f(y) = \frac{\lambda^y e^{-\lambda}}{y!} \implies \ln[f(y)] = y\ln(\lambda) - \lambda - \ln(y!) = -\ln(y!) + \frac{y\zeta - a(\zeta)}{\tau},$$

if we set $\tau = 1$, $\zeta = \ln(\lambda)$, $a(\zeta) = e^{\zeta}$ and $c(y, \tau) = y!^{-1}$. Checking the mean and variance relations with $a(\zeta)$, $\dot{a}(\zeta) = e^{\zeta} = \lambda$ and $\tau\ddot{a}(\zeta) = e^{\zeta} = \lambda$.

Moving on to the gamma distribution, $y \sim G(k, \theta)$,

$$\begin{aligned} f(y) = \frac{y^{k-1}e^{\frac{-y}{\theta}}}{\Gamma(k)\theta^k} \implies \ln[f(y)] &= (k-1)\ln(y) - \frac{y}{\theta} - \ln\left[\Gamma(k)\right] - k\ln(\theta) \\ &= (k-1)\ln(y) - \ln\left[\Gamma(k)\right] + \frac{y\zeta - a(\zeta)}{\tau}, \end{aligned}$$

where $\zeta = \frac{-1}{k\theta}$, $\tau = \frac{1}{k}$, $a(\zeta) = -\ln\left(\frac{-\zeta}{k}\right)$ and $c(y, \tau) = \frac{y^{\frac{1}{\tau}-1}}{\Gamma(\frac{1}{\tau})}$. Checking the mean and variance

relation with $a(\zeta)$, $\dot{a}(\zeta) = \frac{-1}{\zeta} = k\theta$ and $\tau\ddot{a}(\zeta) = \tau\frac{1}{\zeta^2} = k\theta^2$. Thus the Poisson and gamma distributions are members of the exponential family. Table 1.1 provides a description of how to parameterize other distributions as members of the exponential dispersion family.

Distribution	ζ	$a(\zeta)$	τ
B(n,p)	$\ln\left(\frac{p}{1-p}\right)$	$n\ln(1+e^{\zeta})$	1
Poi(λ)	$\ln(\lambda)$	e^{ζ}	1
N(μ,σ^2)	μ	$\frac{\zeta^2}{2}$	σ^2
G(k,θ)	$\frac{-1}{k\theta}$	$-\ln\left(\frac{-\zeta}{k}\right)$	$\frac{1}{k}$
IG(μ,σ^2)	$\frac{-1}{2\mu^2}$	$-\sqrt{-2\zeta}$	σ^2
NB(μ,κ)	$\ln\left(\frac{\kappa\mu}{1+\kappa\mu}\right)$	$\frac{-1}{\kappa}\ln(1-\kappa e^{\zeta})$	1

Table 1.1: Exponential Dispersion Family of Distributions

1.4 Generalized Linear Models

As in linear regression, GLMs attempt to capture a relationship between a response variable, y, and explanatory variables, x_i's. GLMs differ from linear regression in two ways:

(i) The response is no longer normal but can be chosen from the exponential family of distributions.

(ii) A transformation of the mean of the response variable can be applied to get a linear combination of the explanatory variables, as in (1.1).

This allows GLMs the freedom to fit more different types of data sets than classical linear regression. One draw back is that the response might no longer be homoskedastic and therefore its variance will vary with the explanatory variables, heteroskedastic.

A GLM model is defined as a response distribution and mean response:

$$f(y) = c(y,\tau)\exp\left\{\frac{y\zeta - a(\zeta)}{\tau}\right\}, \qquad g(\mu) = \mathbf{x}\boldsymbol{\beta}, \tag{1.7}$$

where $g(\mu)$ is known as the link function, $\mathbf{x}$ is a $1\times(p+1)$ vector representing a data point and μ is the mean of the response. $f(y)$ guarantees that the distribution is in the exponential

family and g describes the relationship between the response and the explanatory variables. Given a data set, building a GLM involves the following steps:

(i) Pick a distribution for the response variable $f(y)$, choose $a(\zeta)$ in (1.7). The response variable is chosen based on the data.

(ii) Determine which link function, $g(\mu)$, to use. Note later in table 1.2 a list of common link functions is given.

(iii) Fit the model by estimating $\boldsymbol{\beta}$ and τ. This is commonly done using packages in SAS or R that implement maximum likelihood estimation or a variant.

(iv) After determining estimates for $\boldsymbol{\beta}$, generate predictions for y given observations and evaluate how well the model fits new data.

The choice of link function is not clear and therefore one might want to try different possibilities.

There exist special types of link function referred to as the canonical link functions. If $g(\mu) = \zeta$ then g is said to be the canonical link function. Using the canonical link function simplifies the estimation, but given the computing power of today one can choose different link functions. Table 1.2 lists some commonly used link functions and their canonical links.

Link Function	$g(\mu)$	Canonical link for
identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	normal
log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	Poisson
inverse	$\mathbf{X}\boldsymbol{\beta} = \mu^{-1}$	gamma, exponential
inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	inverse Gaussian
square root	$\mathbf{X}\boldsymbol{\beta} = \sqrt{\mu}$	
logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	binomial, Bernoulli, multinomial

Table 1.2: Commonly used Link Functions

1.5 Further Topics and Applications to Actuarial Science

When fitting linear models of any kind there are a plethora of problems one can run into, overfitting is one such issue. Overfitting occurs when the model does a good job of fitting past data, but does a poor job of forecasting new data points. This problem can be caused by too many explanatory variables being considered and thus the model is too complex. To resolve it some researchers have proposed regularization. Regularization penalizes large values of the β's. Some being sent to zero thereby eliminating their covariates and simplifying the model. There are two common types of regularization techniques, L_1 and L_2. Rewriting (1.2) as a sum one can pose the L_1 regularization optimization problem as,

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 + \nu ||\boldsymbol{\beta}||_1, \tag{1.8}$$

where

$$||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|,$$

and $\mathbf{x}_i$ represents a row vector of covariates fore one observation. Changing the L_1-norm to an L_2-norm in (1.8) is how L_2 regularization is applied. The effect of regularization is controlled by the ν parameter, larger values of ν produce smaller coefficients. L_2 regularization is more likely to keep all the β's while L_1 regularization will send them to zero faster as ν increases. Also one can apply linear combinations of the two as in Zou and Hastie [2005], which they call an elastic net. They go further and state that one can use any norm that one wants. Any form regularization helps prevent overfitting.

Another issue that can arise is in the instance of outliers. In (1.2) the errors are squared and thus outliers, points that lie far from the mean response, will cause the model to shift greatly. To mitigate this problem researchers have proposed more robust linear regression

models. Huber [1973] developed the following model

$$\min_{\beta} \sum_{i=1}^{n} g(y_i - \mathbf{x}_i\boldsymbol{\beta}), \tag{1.9}$$

where

$$g(y_i - \mathbf{x}_i\boldsymbol{\beta}) = \begin{cases} \frac{1}{2}(y_i - \mathbf{x}_i\boldsymbol{\beta})^2 & \text{for } |(y_i - \mathbf{x}_i\boldsymbol{\beta})| < b \\ b|(y_i - \mathbf{x}_i\boldsymbol{\beta})| - \frac{1}{2}b^2 & \text{for } |(y_i - \mathbf{x}_i\boldsymbol{\beta})| \geq b \end{cases}.$$

Note that b can be a predetermined value or depend on the size of the data set. Huber [1973] states that one does not need to choose g as above but is free to tailor it to the problem. Though it is usually convex. Optimizing (1.9) helps mitigate the effect of outliers, thus helping model the majority of the data and not deviate due to a small proportion of the data.

GLMs have become more popular in the actuarial field, as their data is typically non normal. Thus actuaries usually build models that have gamma or Poisson responses, for severity and counts respectively. Quijano and Garrido [2015] built off this line of thinking to develop a model with a Tweedie response. Their model allowed them to aggregate the claims and thus model the severity and counts together. This is typically done separately and can lead to problems. Another instance of GLMs being applied to actuarial data is in Kafková et al. [2014]. Their paper explores the efficacy of applying generalized linear models to automobile insurance. They analyze their data using different models and then report the performance of each. These are just two examples of how researchers are applying GLMs to actuarial problems.

Chapter 2

Hidden Markov Models

2.1 Introduction

Hidden Markov models (HMMs) have been applied to numerous problems with successful results: speech recognition, text processing, DNA analysis, mobile robot sensor processing, modeling hurricanes, etc. Many of these areas impact our daily lives and are shaping the world that we live in today. This has made the modeling technique very popular, and today researchers are adding their own unique twists to the model so as to best suit their problems. The model relies on the same properties of a Markov Process. This chapter introduces the basics of a Markov chain and an HMM with the help from Fink [2014].

2.2 Markov Process

Markov models have been used to model many different sequential time series data for example, biological sequences, temperature variations, speech utterances, financial data, etc... Markov models benefit greatly from the Markov Property, making them computationally efficient and tractable. Each observation depends on a selection of the previous ones, allowing one to store only the necessary observations as the others have no bearing on the future. Markov models rely on the basics of probability and thus some prior knowledge will be assumed, for a refresher please refer to Chapter 3 of Fink [2014].

Let us define a simple Markov chain in which the next observation only depends on the

previous one, also known as a first order model. Note that you can make each observation depend on as many previous observations as you deem relevant. Let $X_1, X_2, X_3, ...$ denote a sequence of random variables, $i, j \in \{1, .., L\}$ the possible states of a Markov chain. Thus

$$\Pr(X_t|X_{t-1}, ..., X_1) = \Pr(X_t|X_{t-1}), \qquad t \geq 2 \qquad \text{(Markov Property)}.$$

A Markov model is not fully defined without the initial state and transition probabilities. Let a_{ij} be the probability of starting in state i at time $t-1$ and going to state j at time t, i.e. a transition probability,

$$a_{ij} = \Pr(X_t = j|X_{t-1} = i), \qquad t \geq 2, \quad i, j \in \{1, ..., L\}.$$

Note that the transition probabilities are independent of time, t, thus making the model stationary. One can build a model in which these transition probabilities do depend on time. They are called non stationary models and tend to be less common as they add another dimension to the model thus making it less tractable. Lastly let π_i be the probability of starting the sequence in state i, i.e. a initial state probability,

$$\pi_i = \Pr(X_1 = i), \qquad i \in \{1, ..., L\}.$$

Thus the Markov chain (MC) model is summarized by the vector $\boldsymbol{\pi}$ containing all the initial state probabilities, the matrix $\mathbf{A}$ containing all the transition probabilities and the set S containing all possible states, denoted MC($\boldsymbol{\pi}$,$\mathbf{A}$,S).

2.3 Definitions

HMMs build off the Markov chain model and add an additional dimension of variability and unknown. The basic HMM is a two stage stochastic process. Let S_t denote the state at time t, where $S_t \in S$ and $t = 1, 2, ..., T$. This is the first stage which involves latent or hidden transitions from state S_{t-1} to S_t. Set S contains all possible states given the problem. This

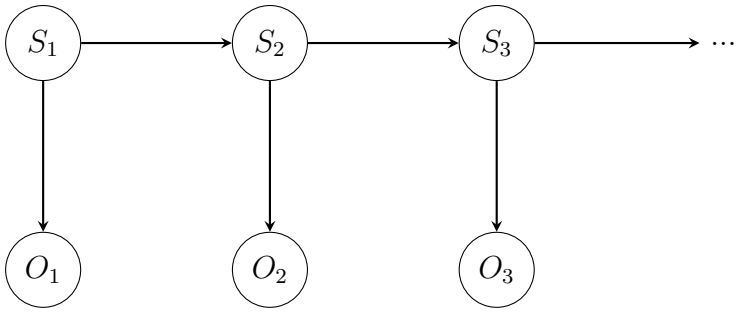

Figure 2.1: Graphical Representation of an HMM

model takes advantage of the Markov property,

$$\Pr(S_t|S_1, ..., S_{t-1}) = \Pr(S_t|S_{t-1}), \qquad t \geq 2.$$

The probability distribution of the next state only depends on the previous state, if known, and all prior states are therefore superfluous information. This property makes the model computationally efficient because less information needs to be stored. Note that these stages are called hidden as that are not observed.

The second stage captures the observations, what the researcher or practitioner actually observes. Let O_t denote the observation at time t, these can be discrete or continuous. The probability distribution for O_t is dependent on state S_t and none of the previous states,

$$\Pr(O_t|O_1, ..., O_{t-1}, S_1, ..., S_t) = \Pr(O_t|S_t).$$

In the literature this property is referred to as the output independence assumption.

HMMs are thus a way to model time series data and capture a hidden variable or uncertainty that changes with time. These models have a nice graphical representation depicted in Figure 2.1. An HMM is not complete without defining the necessary parameters, like a Markov chain. Given the model described above one would need:

- a matrix $\mathbf{A}$ containing the transition probabilities given L states $\mathbf{A}$ will be $L \times L$:

$$\mathbf{A} = \{a_{ij}|a_{ij} = \Pr(S_t = j|S_{t-1} = i)\}; \qquad i, j \in \{1, 2, ...L\}.$$

- a vector $\boldsymbol{\pi}$ of the initial probabilities which will be $L \times 1$:

$$\boldsymbol{\pi} = \{\pi_i | \pi_i = \Pr(S_1 = i)\}; \qquad i \in \{1, ..., L\}.$$

- and conditional/output probability distributions for specific states:

$$\mathbf{B} = \{b_i(o_t) | b_i(o_t) = \Pr(O_t = o_t | S_t = i)\}; \qquad t \in \{1, ..., T\} \quad i \in \{1, ..., L\}.$$

Note that o_t represents the value of the emissions at time t and the transition probabilities, a_{ij}, do not depend on t. The emissions can be discrete or continuous depending on the problem. After defining the model three problems remain: model evaluation, decoding, and parameter estimation.

2.4 Evaluation

The most widely used measure to evaluate a HMM's efficacy is the total production probability. Let ϕ denote the set of model parameters, $\phi = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}\}$, then the total output probability is defined as $\Pr(\mathbf{O}|\phi)$, where $\mathbf{O}$ denotes a sequence of observations up to T, the terminal time. Another criterion sometime considered is the probability of the observation sequence when traveling along the optimal state sequence. This is referred to as the optimal output probability. The total output probability considers all possible paths and takes the sum whereas the optimal path probability considers just the most likely path given the data. Currently, there exists no algorithm to find the optimal HMM model given the observations and a criterion. Therefore one has to consider a finite number of models and choose a criterion to evaluate said models.

2.4.1 Total Output Probability

If more than one HMM ϕ_i's are under consideration one would want to choose the model, ϕ_j, that best represents the sequence of data. One considers a space of models, $\phi_i \in \Omega_i$, and chooses the model that maximizes the posterior in that space:

$$\Pr(\phi_j|\mathbf{O}) = \max_i \frac{\Pr(\mathbf{O}|\phi_i)\Pr(\phi_i)}{\Pr(\mathbf{O})}. \tag{2.1}$$

In practice when calculating this expression, the denominator can be dropped as it does not depend on ϕ_i,

$$\phi_j = \underset{\phi_i}{\operatorname{argmax}} \Pr(\phi_i|\mathbf{O}) = \underset{\phi_i}{\operatorname{argmax}} \frac{\Pr(\mathbf{O}|\phi_i)\Pr(\phi_i)}{\Pr(\mathbf{O})} = \underset{\phi_i}{\operatorname{argmax}} \Pr(\mathbf{O}|\phi_i)\Pr(\phi_i). \tag{2.2}$$

Also for this probability to be attainable the prior probabilities $\Pr(\phi_i)$ need to be specified. This is often not the case and the priors tend to be difficult or impossible to calculate. Thus when choosing a model the $\Pr(\phi_i)$ term is often dropped. Next let us see how this quantity is calculated.

In order for us to compute the total output probability we will fist consider the probability of an observation sequence and a state sequence given ϕ, $\Pr(\mathbf{O}, \mathbf{S}|\phi)$. Both the observation and state sequence must be the same length, T. Then if we marginalize over the possible state sequences $\mathbf{S}$'s we get,

$$\sum_{\mathbf{S}} \Pr(\mathbf{O}, \mathbf{S}|\phi) = \Pr(\mathbf{O}|\phi).$$

Therefore we can use this to tabulate the total output probability. Using properties of conditional probabilities one can rewrite this sum as,

$$\sum_{\mathbf{S}} \Pr(\mathbf{O}, \mathbf{S}|\phi) = \sum_{\mathbf{S}} \Pr(\mathbf{O}|\mathbf{S}, \phi)\Pr(\mathbf{S}|\phi).$$

Then taking advantage of independences in the structure of HMMs and using the definitions from Section 2.3 we get

$$\Pr(\mathbf{O}|\mathbf{S}, \phi) = \prod_{t=1}^{T} b_{s_t}(O_t),$$
$$\Pr(\mathbf{S}|\phi) = \pi_{s_1} \prod_{t=2}^{T} a_{s_{t-1}s_t}.$$

Thus after estimating the parameters of the model we can then find the total output probability. Before writing out the full expression let us set $a_{0s_1} = \pi_{s_1}$ and $S_0 = 0$, this will simplify

the notation. Given these new definitions we can now write the total output probability as

$$\Pr(\mathbf{O}|\phi_i) = \sum_{\mathbf{S}} \Pr(\mathbf{O}, \mathbf{S}|\phi) = \sum_{\mathbf{S}} \prod_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(O_t).$$

Note that this brute force approach is typically avoided as the algorithm is computationally complex, $O(TL^T)$ where L is the number of states. There exists faster computational methods.

A more commonly used method is the forward algorithm, which exploits the model's structure. It allows us to compute the total output probability in a recursive fashion. Let $\alpha_t(i) = \Pr(O_1, ..., O_t, S_t = i|\phi)$, which is known as the forward variable, then the forward algorithm is

1. Initialization:
   $\alpha_1(i) := \pi_i b_i(O_1), \qquad i \in \{1, ..., L\}.$
2. Recursion for $t = 2, ..., T$:
   $\alpha_t(j) := \sum_i \left(\alpha_{t-1}(i) a_{ij}\right) b_j(O_t), \qquad j \in \{1, ..., L\}.$
3. Termination:
   $\Pr(\mathbf{O}|\phi) = \sum_{i=1}^{L} \alpha_T(i).$

This nice recursion can be shown using properties of conditional probability. Beginning with the initialization,

$$\begin{aligned}
\alpha_1(i) &= \Pr(O_1, S_1 = i|\phi) && \text{(by definition)} \\
&= \Pr(O_1|S_1 = i, \phi) \Pr(S_1 = i|\phi) && \text{(properties of conditional probability)} \\
&= b_i(O_1)\pi_i && \text{(parameters).}
\end{aligned}$$

Next we need to show the recursion. Let us proceed by induction; first base case

$$\begin{aligned}
\alpha_2(j) &= \Pr(O_1, O_2, S_2 = j|\phi) \\
&= \sum_i \Pr(O_1, O_2, S_1 = i, S_2 = j|\phi) \\
&= \sum_i \Pr(O_2|O_1, S_1 = i, S_2 = j, \phi) \Pr(S_2 = j|S_1 = i, O_1, \phi) \Pr(O_1, S_1 = i|\phi) \\
&= \sum_i \Pr(O_2|S_2 = j, \phi) \Pr(S_2 = j|S_1 = i, \phi) \Pr(O_1, S_1 = i|\phi) \\
&= \sum_i (\alpha_1(i)a_{ij})\, b_j(O_2).
\end{aligned}$$

Assuming that $\alpha_t(j) := \sum_i (\alpha_{t-1}(i)a_{ij})\, b_j(O_t)$ we will show this implies the $t+1$ case:

$$\begin{aligned}
\alpha_{t+1}(j) &= \Pr(O_1, ..., O_{t+1}, S_{t+1} = j|\phi) \\
&= \sum_i \Pr(O_1, ..., O_{t+1}, S_t = i, S_{t+1} = j|\phi) \\
&= \sum_i \Pr(O_{t+1}|O_1, ..., O_t, S_t = i, S_{t+1} = j, \phi) \Pr(S_{t+1} = j|S_t = i, O_1, ..., O_t, \phi) \\
&\qquad \Pr(S_t = i, O_1, ..., O_t|\phi) \\
&= \sum_i \Pr(O_{t+1}|S_{t+1} = j, \phi) \Pr(S_{t+1} = j|S_t = i, \phi) \Pr(S_t = i, O_1, ..., O_t|\phi) \\
&= \sum_i (\alpha_t(i)a_{ij})\, b_j(O_{t+1}).
\end{aligned}$$

Then to find the total output probability, termination of the algorithm, one needs to sum over all states at time T. Like the name suggests there is a backward algorithm as well. Also one can combine the two and use the forward-backward algorithm. It is left up to the practitioner which one to use, they both provide the same result.

2.4.2 Optimal Output Probability

Total output probability, which captures how well the model works on average, might not select the model that performs the best for a certain case, sequence of states. In this instance one could use the optimal output probability as the criterion, $\max_{\mathbf{S}} \Pr(\mathbf{O}, \mathbf{S}|\phi)$. This value can

also be calculated in a recursive fashion. Let $\delta_t(i) := \max_{S_1,...,S_{t-1}} \Pr(O_1, ..., O_t, S_1, ..., S_{t-1}, S_t = i|\phi)$, then the algorithm can be written as follows:

1. Initialization:
   $\delta_1(i) := \pi_i b_i(O_1), \qquad i \in \{1, ..., L\}.$
2. Recursion for $t = 2, ..., T$:
   $\delta_t(j) := \max_i \left(\delta_{t-1}(i) a_{ij}\right) b_j(O_t), \qquad j \in \{1, ..., L\}.$
3. Termination:
   $\max_{\mathbf{S}} \Pr(\mathbf{O}, \mathbf{S}|\phi) = \max_i \delta_T(i).$

This recursion can be shown using the same properties as before. Starting with the initialization,

$$\begin{aligned}
\delta_1(i) &= \Pr(O_1, S_1 = i|\phi) && \text{(by definition the max is dropped)} \\
&= \Pr(O_1|S_1 = i, \phi) \Pr(S_1 = i|\phi) && \text{(properties of conditional probability)} \\
&= b_i(O_1)\pi_i && \text{(parameters).}
\end{aligned}$$

Like before we will use induction, first the base case:

$$\begin{aligned}
\delta_2(j) &= \max_{S_1} \Pr(O_1, O_2, S_1, S_2 = j|\phi) \\
&= \max_{S_1} \Pr(O_2|O_1, S_1, S_2 = j, \phi) \Pr(S_2 = j|S_1, O_1, \phi) \Pr(O_1, S_1|\phi) \\
&= \max_{S_1} \Pr(O_2|S_2 = j, \phi) \Pr(S_2 = j|S_1, \phi) \Pr(O_1, S_1|\phi) \\
&= \max_i \left(\delta_1(i) a_{ij}\right) b_j(O_2).
\end{aligned}$$

Next we need to show that case t implies case $t+1$ by assuming that $\delta_t(j) := \max_i (\delta_{t-1}(i)a_{ij})$ $b_j(O_t)$,

$$\begin{aligned}
\delta_{t+1}(j) &= \max_{S_1,...,S_t} \Pr(O_1, ..., O_{t+1}, S_1, ..., S_t, S_{t+1} = j|\phi) \\
&= \max_{S_1,...,S_t} \Pr(O_{t+1}|O_1, ..., O_t, S_1, ..., S_t, S_{t+1} = j, \phi) \Pr(S_{t+1} = j|O_1, ..., O_t, S_1, ..., S_t, \phi) \\
&\qquad \Pr(O_1, ..., O_t, S_1, ..., S_t|\phi) \\
&= \max_{S_1,...,S_t} \Pr(O_{t+1}|S_{t+1} = j, \phi) \Pr(S_{t+1} = j|s_t, \phi) \Pr(O_1, ..., O_t, S_1, ..., S_t|\phi) \\
&= \max_i (\delta_t(i)a_{ij})\, b_j(O_t).
\end{aligned}$$

Then to terminate the recursion we need to find the state at time T that maximizes δ. This algorithm, like the one listed before can be carried out in a backward fashion or a combination of forward and backward. In practice to carry out the computation one usually transforms the probabilities by taking the log. This changes the product into a sum without affecting the max.

2.5 Decoding

Decoding is the process of finding the state sequence that best fits the observations. Often these states have no real life interpretation and thus decoding is not necessary. In certain cases the states have meaningful interpretations, in these instances decoding becomes an interesting problem. One can use the brute force method described at the beginning of Section 2.4.1 to find the optimal state sequence ($\mathbf{S}^* = \underset{\mathbf{S}}{\operatorname{argmax}} \Pr(\mathbf{S}|\mathbf{O}, \phi)$). Like before this method is computationally expensive.

In applications practitioners tend to use the Viterbi algorithm as it is more computationally efficient. Using the algorithm defined before to calculate the optimal output probability one can find the optimal path. It is performed in a backward fashion. Let $\psi_1(i) := 0$ and $\psi_{t+1}(j) := \underset{i}{\operatorname{argmax}} \{\delta_t(i)a_{ij}\}$, then picking up at the end of the algorithm:

3. Termination:

   $\max_{\mathbf{s}} \Pr(\mathbf{O}, \mathbf{s}|\phi) = \max_i \delta_T(i)$

$s_T^* := \underset{j}{\operatorname{argmax}}\, \delta_T(j)$.

4. Back-Tracking of the Optimal Path:

   for $t = T-1, ..., 1$,

   $s_t^* := \psi_{t+1}(s_{t+1}^*)$.

Unfortunately given the backward direction of the algorithm one cannot find the optimal path until one has computed the optimal output probability. This can become a problem for interactive systems, in such cases there exists algorithms to provide partial feedback while the computation is ongoing. These methods suffer from the problem of finding sub optimal solutions.

2.6 Parameter Estimation

All the algorithms described require that the parameters of the model be estimated. In general one should choose an HMM architecture that best resembles the statistical properties of the data and then estimate the required parameters. One is free to change the model architecture, the number of states and the emission probabilities. For instance, given a problem with continuous non zero emissions one might want to stay away from the common choice of normal distributions. After doing so there are multiple algorithms to optimize the parameters. We will focus on one, the expectation-maximization (EM) algorithm also known as the Baum-Welch algorithm for HMMs. This is the most commonly used method and focuses on maximizing the total output probability. Other algorithms exist such as the Viterbi and Segmental k-Means algorithms which focus on maximizing the optimal output probability.

Before describing the EM algorithm in further detail there are a few more quantities that are important to define. Let $\beta_t(j) = \Pr(O_{t+1}, ..., O_T | S_t = j, \phi)$, which is known as the backward variable. When implementing the forward-backward algorithm one uses the backward variable in tandem with $\alpha_t(i)$. Note it is calculated in a recursive fashion. $\beta_T(j) =$

1 for all $j \in \{1, ..., L\}$, then for $1 \leq t \leq T-1$

$$\begin{aligned}
\beta_t(j) &= \Pr(O_{t+1}, ..., O_T | S_t = j, \phi), \qquad j \in \{1, ..., L\}, \\
&= \sum_i \Pr(O_{t+1}, ..., O_T, S_{t+1} = i | S_t = j, \phi) \\
&= \sum_i \Pr(O_{t+1} | O_{t+2}, ..., O_T, S_{t+1} = i, S_t = j, \phi) \Pr(O_{t+2}, ..., O_T | S_{t+1} = i, S_t = j, \phi) \\
&\qquad \Pr(S_{t+1} = i | S_t = j, \phi) \\
&= \sum_i \Pr(O_{t+1} | S_{t+1} = i, \phi) \Pr(O_{t+2}, ..., O_T | S_{t+1} = i, \phi) \Pr(S_{t+1} = i | S_t = j, \phi) \\
&= \sum_i b_i(O_{t+1}) \beta_{t+1}(i) a_{ji}, \qquad j \in \{1, ..., L\}.
\end{aligned}$$

One can use the backward variable to compute the total output probability thus giving us the backward algorithm,

$$\Pr(\mathbf{O} | \phi) = \sum_i^L \pi_i b_i(O_1) \beta_1(i).$$

Using this, one can now imagine how to program the forward-backward algorithm.

Another important value to define is the probability of being in state i at time t given an observation sequence,

$$\begin{aligned}
\gamma_t(i) &= \Pr(S_t = i | \mathbf{O}, \phi), \qquad i \in \{1, ..., L\}, \quad t \geq 0, \\
&= \frac{\Pr(O_1, ..., O_T, S_t = i | \phi)}{\Pr(\mathbf{O} | \phi)} \\
&= \frac{\Pr(O_1, ..., O_t, S_t = i | \phi) \Pr(O_{t+1}, ..., O_T | S_t = i, \phi)}{\Pr(\mathbf{O} | \phi)} \\
&= \frac{\alpha_t(i) \beta_t(i)}{\Pr(\mathbf{O} | \phi)}.
\end{aligned}$$

We now have a way to compute the probability of being in state i at time t, next we will

find the probability of going from state i to j at time t:

$$\begin{aligned}
\gamma_t(i,j) &= \Pr(S_t = i, S_{t+1} = j|\mathbf{O}, \phi), \qquad i,j \in \{1,...,L\}, \quad t \geq 0, \\
&= \frac{\Pr(S_t = i, S_{t+1} = j, \mathbf{O}|\phi)}{\Pr(\mathbf{O}|\phi)} \\
&= \frac{1}{\Pr(\mathbf{O}|\phi)} \Pr(O_1,...,O_t, S_t = i|\phi) \Pr(S_{t+1} = j|S_t = i, \phi) \\
&\quad \Pr(O_{t+1}|S_{t+1} = j, \phi) \Pr(O_{t+2},...,O_T|S_{t+1} = j, \phi) \\
&= \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\Pr(\mathbf{O}|\phi)}.
\end{aligned}$$

Note that $\gamma_t(i) = \sum_j \gamma_t(i,j)$, a relationship that can be helpful in the numerical calculations. These three quantities are used in the derivation of the EM algorithm. It is also worth pointing out that these quantities might be different depending on the architecture of your model. We can now proceed to the EM algorithm.

First one must clearly define the parameters, $\phi = \{\mathbf{A}, \boldsymbol{\pi}, \mathbf{B}\}$ and initialize values for these parameters. Note that this step is important as EM does not guarantee convergence to the global maximum. It can get stuck at a local solutions if these exist. So initializing the parameters close to their optimal values would be ideal. This is why EM is usually run multiple times with different initial parameters. After initialization the EM algorithm has two steps, the E step and the M step. In order to proceed let us define the likelihood assuming complete data, including the hidden Markov chain states:

$$\begin{aligned}
L_T^c(\phi) &= \Pr(O_1 = o_1,...,O_T = o_T, S_1 = s_1,...,S_T = s_T) \\
&= \Pr(O_T = o_T|O_1 = o_1,...,O_{T-1} = o_{T-1}, S_1 = s_1,...,S_T = s_T)... \\
&\quad \Pr(O_1 = o_1|S_1 = s_1,...,S_T = s_T) \Pr(S_T = s_T|S_1 = s_1,...,S_{T-1} = s_{T-1})... \Pr(S_1 = s_1) \\
&= \pi_{s_1} b_{s_1}(o_1) \prod_{i=2}^{T} a_{s_{i-1}s_i} b_{s_i}(o_i).
\end{aligned}$$

Summing over $s_1, ..., s_T$ we obtain the likelihood for the incomplete data,

$$\begin{aligned} L_T(\phi) &= \Pr(O_1 = o_1, ..., O_T = o_T) \\ &= \sum_{s_1} .... \sum_{s_T} \pi_{s_1} b_{s_1}(o_1) \prod_{i=2}^{T} a_{s_{i-1}s_i} b_{s_i}(o_i). \end{aligned}$$

Given our situation of incomplete data the natural thing to do is to use the EM algorithm, to maximize the expected likelihood over the complete data. After defining the likelihoods we can now continue to describe the E and M steps. Given a set of observed but incomplete data, $\mathbf{O} = \{o_1, ..., o_T\}$:

1. Initialize the parameters in ϕ^0.

2. Compute
$$Q(\phi; \phi^k) = \mathrm{E}_{\phi^k} \left(\ln(L_T^c(\phi)|\mathbf{O}) \right).$$

3. Find ϕ^{k+1} that maximizes $Q(\phi; \phi^k)$.

4. Repeat Steps 2 and 3 in an alternating way until
$$\ln(L_T(\phi^{k+1})) - \ln(L_T(\phi^k))$$
is sufficiently small.

Wu [1983] showed that EM converges to at least a local maximum. The actual calculation of Steps 3 and 4 vary depending on the assumed emission distributions, this decision is left up to the practitioner.

2.6.1 Example

To demonstrate the EM algorithm we will consider a simple case, a discrete emission distribution only taking observed values (i.e. $b_{s_j}(o_k) = P(O_t = o_t | S_j = s_j)$ where o_t are only observed values). For notation and simplicity let us assume that $o_t \in \{1, ..., K\}$, also let L denote the number of states. First initialize our parameters ϕ^0, this should be done respecting

the fact that $\sum_{i=1}^{L} \pi_i = 1$, $\sum_{j=1}^{L} a_{ij} = 1$ and $\sum_{k=1}^{K} b_j(k) = 1$. Second we must calculate $\mathrm{E}_{\phi^0}\left(\ln(L_T^c(\phi)|\mathbf{O})\right)$, using the result from before $L_T^c(\phi) = \pi_{s_1} b_{s_1}(o_1) \prod_{i=2}^{T} a_{s_{i-1}s_i} b_{s_i}(o_i)$, we get

$$\begin{aligned}
\mathrm{E}_{\phi^0}\left(\ln(L_T^c(\phi)|\mathbf{O})\right) &= \mathrm{E}_{\phi^0}\left(\ln\left(\pi_{s_1} b_{s_1}(o_1) \prod_{i=2}^{T} a_{s_{i-1}s_i} b_{s_i}(o_i)\right)\right) \\
&= \mathrm{E}_{\phi^0}(\ln(\pi_{s_1})) + \sum_{i=1}^{T} \mathrm{E}_{\phi^0}(\ln(b_{s_i}(o_i))) + \sum_{i=2}^{T} \mathrm{E}_{\phi^0}(\ln(a_{s_{i-1}s_i})) \\
&= \sum_{j=1}^{L} \Pr(S_1 = j|\mathbf{O}, \phi) \ln(\pi_j) + \sum_{j=1}^{L} \sum_{i=1}^{T} \ln(b_{s_i=j}(o_i)) \Pr(S_i = j|\mathbf{O}, \phi) \\
&\qquad + \sum_{k=1}^{L} \sum_{j=1}^{L} \sum_{i=2}^{T} \ln(a_{s_{i-1}=k, s_i=j}) \Pr(S_{i-1} = k, S_i = j|\mathbf{O}, \phi) \\
&= \sum_{j=1}^{L} \gamma_1(j) \ln(\pi_j) + \sum_{j=1}^{L} \sum_{i=1}^{T} \ln(b_{s_i=j}(o_i)) \gamma_i(j) \\
&\qquad + \sum_{k=1}^{L} \sum_{j=1}^{L} \sum_{i=2}^{T} \ln(a_{s_{i-1}=k, s_i=j}) \gamma_{i-1}(k, j).
\end{aligned}$$

With our initial set of parameters this value can be calculated thus completing the E-step.

Moving to the M-step we need to maximize the above with respect to ϕ, under a few constraints, as probability mass functions need to sum to one. Thus the problem can be written as,

$$\max Q(\phi; \phi^0)$$

with repect to $\boldsymbol{\pi}$, $\mathbf{A}$, $\mathbf{B}$ and subject to

$$g_1(\boldsymbol{\pi}) = \sum_{j=1}^{L} \pi_j = 1, \tag{2.3}$$

$$g_2(a_{k1}, ..., a_{kL}) = \sum_{j=1}^{L} a_{kj} = 1, \qquad \text{for all } k \in \{1, ..., L\} \tag{2.4}$$

$$g_3(\mathbf{B}) = \sum_{k=1}^{K} b_j(k) = 1, \qquad \text{for all } j \in \{1, ..., L\}. \tag{2.5}$$

Then we can introduce a new function F using the theory of Lagrange multipliers,

$$F(\phi^0, \boldsymbol{\kappa}) = Q(\phi; \phi^0) + \kappa_1 \left(1 - \sum_{j=1}^{L} \pi_j\right) + \kappa_2 \left(1 - \sum_{j=1}^{L} a_{kj}\right) + \kappa_3 \left(1 - \sum_{k=1}^{K} b_j(k)\right)$$

where $\boldsymbol{\kappa} = \{\kappa_1, \kappa_2, \kappa_3\}$ are the Lagrange multipliers. Now first consider the initial state probabilities. Taking the derivative of F with respect to π_i we get

$$\frac{\partial F}{\partial \pi_i} = \frac{\gamma_1(i)}{\pi_i} - \kappa_1, \qquad i \in \{1, ..., L\}.$$

Setting equal to zero and solving for π_i gives

$$\pi_i = \frac{\gamma_1(i)}{\kappa_1}.$$

Using our Constraint (2.3) and the fact that the derivative would be the same for all π's we get

$$1 = \sum_{j=1}^{L} \frac{\gamma_1(j)}{\kappa_1} \implies \kappa_1 = \sum_{j=1}^{L} \gamma_1(i) \implies \kappa_1 = 1.$$

This is because $\gamma_1(j)$ represents the probability of being in state j at time 1 and summing over all j makes the left side equal to 1. Thus the optimal initial stat probabilities are given by:

$$\pi_i = \gamma_1(i), \qquad i \in \{1, ..., L\}.$$

Next we need to optimize our transition probabilities. Taking the derivative with respect to a_{kj} gives

$$\frac{\partial F}{\partial a_{kj}} = \frac{\sum_{i=2}^{T} \gamma_{i-1}(k, j)}{a_{kj}} - \kappa_2, \qquad k, j \in \{1, ..., L\}.$$

Setting equal to zero and solving for a_{kj},

$$a_{kj} = \frac{\sum_{i=2}^{T} \gamma_{i-1}(k,j)}{\kappa_2}.$$

Then plugging the above solution for a_{kj} into the Constraint 2.4 we obtain

$$1 = \sum_{j=1}^{L} \frac{\sum_{i=2}^{T} \gamma_{i-1}(k,j)}{\kappa_2} \implies \kappa_2 = \sum_{j=1}^{L} \sum_{i=2}^{T} \gamma_{i-1}(k,j)$$
$$\kappa_2 = \sum_{i=2}^{T} \sum_{j=1}^{L} \gamma_{i-1}(k,j) = \sum_{i=2}^{T} \gamma_{i-1}(k).$$

Thus the optimal transition probabilities are given by

$$a_{kj} = \frac{\sum_{i=2}^{T} \gamma_{i-1}(k,j)}{\sum_{i=2}^{T} \gamma_{i-1}(k)}, \qquad k, j \in \{1, ..., L\}.$$

Last, to derive an estimate for our emission probabilities, take the derivative with respect to $b_j(k)$:

$$\frac{\partial F}{\partial b_j(k)} = \frac{\sum_{i:o_i=k} \gamma_i(j)}{b_j(k)} - \kappa_3.$$

Setting equal to zero and solving for $b_j(k)$ gives

$$b_j(k) = \frac{\sum_{i:o_i=k} \gamma_i(j)}{\kappa_3}.$$

Then substituting into Constraint (2.5) and solving for κ_3 yields

$$1 = \sum_{k=1}^{K} \frac{\sum_{i:o_i=k} \gamma_i(j)}{\kappa_3} \implies \kappa_3 = \sum_{k=1}^{K} \sum_{i:o_i=k} \gamma_i(j)$$

and summing over all possible k allows us to simplify to

$$\kappa_3 = \sum_{i=1}^{T} \gamma_i(j).$$

Thus

$$b_j(k) = \frac{\sum_{i:o_i=k} \gamma_i(j)}{\sum_{i=1}^{T} \gamma_i(j)}, \qquad j,k \in \{1,...,L\}.$$

We know that we are moving in the direction of a critical point. Thus to ensure that it is at least a local maximum we can check the Hessian. This is the matrix of all the second derivatives. Using our function F the Hessian becomes

$$H = \begin{bmatrix} \frac{\partial^2 F}{\partial \pi_1^2} & 0 & \dots & 0 \\ 0 & \frac{\partial^2 F}{\partial \pi_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \frac{\partial^2 F}{\partial b_N(K)^2} \end{bmatrix}.$$

Note the diagonal captures the second derivative with respect to each parameter $(\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$. Here the matrix is diagonal as the sums break apart nicely, the second derivative with respect to two different parameters becomes zero. Also notice that we can ignore the Lagrange multipliers as we they were created as dummy variables. Now inputing the derivatives we get

$$H = \begin{bmatrix} \frac{-\gamma_1(1)}{\pi_1^2} & 0 & \dots & 0 \\ 0 & \frac{-\gamma_1(2)}{\pi_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \frac{-\sum_{i:o_i=k} \gamma_i(j)}{b_N(K)^2} \end{bmatrix}.$$

Notice how the elements on the diagonal are negative because the numerator and denominator of each fraction are probabilities, which by definition are greater than zero. This implies that H is negative definite and therefore the point we are marching towards is indeed a local maximum. This provides the estimates for Steps 2 and 3 of the EM algorithm for the simplest case HMM and ensures that is moving towards a maximum.

2.7 Further Topics

Another popular variation of HMMs is to change the emissions to a normal distribution or a sum of normals, for more information refer to Fink [2014]. One can also derive estimates for an HMM with multiple observation sequences of varying length; we discuss this idea in Chapter 3. Another extension is that the dependency arrows, as in Figure 2.1, can be drawn in more general patterns. Longer time dependencies can be created to fit different problems. Similarly, the number of observations for each latent state can be changed. The model architecture is flexible and can be made to fit different settings, leading to the popularity of HMMs. For other algorithms or more examples of HMMs please refer to Fink [2014].

Chapter 3

An HMM for Modeling Claims

3.1 Introduction

Automobile insurance plays a pivotal role in the property and casualty realm. The problem of forecasting claim counts and claim severity within auto insurance is vital for a business and companies are constantly looking for new state of the art methods to gain a competitive edge. Their very livelihood depends on accurate models. Thus far actuaries have used loss models based on GLMs to predict claims. Alternatively, here we derive a model using HMMs to forecast auto insurance claims. Note that without loss of generality, this type of model could be used for many other kinds of insurance.

3.2 HMMs for Auto Insurance

Unlike common loss models HMMs have a time dependency; in particular the model considered here has a one time period dependency. This dependency can be changed by either increasing the number of past states that influence the current state or by changing the state that influences the current state, that is have state t depend on state $t - i$ where $i > 1$. In the case of auto insurance the latent states could be viewed as representing the driving ability of the policy holder in the past year, for example s_1 = good driving year, s_2 = average driving year, and s_3 = bad driving year. Note that the the time period does not need to be a year. This builds a nice intuitive reasoning as the insurer does not see

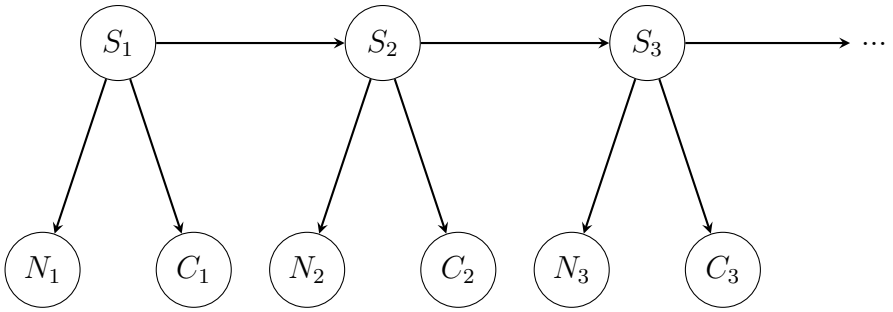

Figure 3.1: Graphical Representation of a Poisson-Gamma HMM

directly how well a policy holder drove in the past year, unless they have a remote control sensor in the car, the only information that the insurer sees is the claims. Also this approach views driving skills as a dynamic ability, a diver can change from a good to a bad driver over different time periods. This allows for hypothetical situations like: given a bad driving year a policyholder might become more cautious/better driver, or given a few good years one might become overconfident and take more risks to become a bad driver. Before testing if these situations are possible and should be accommodated for we must build an HMM for insurance.

First we need to determine the type of distributions for our emissions. As is common in auto insurance, we propose the Poisson distribution to model claim counts and the gamma distribution to model the claim severity. Thus the model will have two outputs, count and severity, and these distributions we chose so we can compare any results to other loss models. Above is a graphical depiction of the proposed model as in Figure 2.1, where N_i is the claim count and C_i refers to the average severity incurred during the ith time period. The model described in Figure 3.1 has the advantage of estimating the parameters of the gamma and Poisson distribution together through a similar latent variable making them dependent on a hidden variable. Commonly these two quantities are estimated separately and then multiplied together, which some find controversial as there could be a dependence at play.

The idea of applying HMMs to model non life insurance is not new, see for instance Paroli et al. [2000]. In their research they use Poisson hidden Markov models (PHMMs), where the emissions are Poisson distributed, to model the daily frequencies of injury in the work

place in Italy. They derive maximum likelihood estimates for the λ parameters that govern the Poisson distributions. They claim that this helps deal with the overdispersion problem in count data, the fact that the Poisson distribution cannot handle too much variability. Overdispersion occurs when there is greater variability in the data than would be expected given the statistical model, which usually manifests in car insurance with too many zero observations. Switching λ based on latent states provides a more accurate model for the overdispersion by distributing the variability across multiple Poisson distributions. Lu and Zeng [2012] also used PHMMs to model hurricanes to assess the risk of an insurer. They proposed using a non homogeneous PHMM, the parameters changed with time, to better predict the seasonal variations.

The idea of an HMM with gamma emissions has also been researched, see Zhang et al. [2012] and Mohammadiha et al. [2013]. The first paper builds a model to represent ozone levels, while Mohammadiha et al. [2013] propose a similar model for speech signals. In both Zhang et al. [2012] and Mohammadiha et al. [2013], maximum likelihood estimates are derived for the parameters (k, θ) that summarize the gamma distribution. Thus far no one has written about the above Poisson-gamma HMM in any field nor has anyone written about a gamma HMM for insurance.

3.3 MLE Estimates

The proposed model differs from the one described in Chapter 2 as it emits two observations, which describes a compound Poisson-gamma with a latent state changing their parameters. Also the field of application will be different, unlike Paroli et al. [2000] we are not just curious about number of claims but also the severity. Let L be the number of states, there will be L^2 transition probabilities, L initial probabilities, and $3L$ parameters for the Poisson and gamma distributions to estimate.

3.3.1 Single Observation Sequence

First we will deal with the case of modeling all policy holders in a portfolio as individuals. We will need estimates for our parameters $\phi = \{\mathbf{A}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \mathbf{k}, \boldsymbol{\theta}\}$, where $\boldsymbol{\lambda}$ is a vector of

the parameters for the Poisson distributions and $\mathbf{k}$ and $\boldsymbol{\theta}$ are vectors of the parameters for the gamma distributions. Also let O_i denote the random vector of observations, i.e. $O_i = (N_i, C_i)$, and $\mathbf{O}$ be all the observed values in the sequence. First we write the Q function:

$$\begin{aligned}
Q(\phi;\phi^0) &= \mathrm{E}_{\phi^0}\left(\ln(L_T^c(\phi)|\mathbf{O})\right) \\
&= \mathrm{E}_{\phi^0}\left(\ln\left(\pi_{s_1} b_{s_1}(o_1)\prod_{i=2}^{T} a_{s_{i-1}s_i} b_{s_i}(o_i)\right)|\mathbf{O}\right) \\
&= \mathrm{E}_{\phi^0}\left(\ln(\pi_{s_1})|\mathbf{O}\right) + \sum_{i=2}^{T}\mathrm{E}_{\phi^0}\left(\ln(a_{s_{i-1}s_i})|\mathbf{O}\right) + \sum_{i=1}^{T}\mathrm{E}_{\phi^0}\left(\ln(b_{s_i}(o_i))|\mathbf{O}\right) \\
&= \sum_{j=1}^{L}\ln(\pi_{s_1=j})\Pr(S_1=j|\mathbf{O}) + \sum_{k=1}^{L}\sum_{j=1}^{L}\sum_{i=2}^{T}\ln(a_{s_{i-1}=k,s_i=j})\Pr(S_i=k,S_{i-1}=j|\mathbf{O}) \\
&\quad + \sum_{j=1}^{L}\sum_{i=1}^{T}\ln(b_{s_i=j}(o_i))\Pr(S_i=l|\mathbf{O}) \\
&= \sum_{j=1}^{L}\gamma_1(j)\ln(\pi_j) + \sum_{k=1}^{L}\sum_{j=1}^{L}\sum_{i=2}^{T}\ln(a_{s_{i-1}=k,s_i=j})\gamma_{i-1}(k,j) + \sum_{j=1}^{L}\sum_{i=1}^{T}\ln(b_{s_i=j}(o_i))\gamma_i(j).
\end{aligned}$$

Again we started at the first iteration to simplify the notation. The Q function will be the same at each iteration making the results applicable at each update. Thanks to the nice structure of the model the Q function breaks up very nicely, making the estimation of the parameters easier. Next, split $b_{s_i=j}(o_i)$ by following model structure:

$$\begin{aligned}
\sum_{j=1}^{L}\sum_{i=1}^{T}\ln(b_{s_i=j}(o_i))\gamma_i(j) &= \sum_{j=1}^{L}\sum_{i=1}^{T}\ln(\Pr(N_i=n_i,C_i=c_i|S_i=j))\gamma_i(j) \\
&= \sum_{j=1}^{L}\sum_{i=1}^{T}\ln(\Pr(N_i=n_i|S_i=j)\Pr(C_i=c_i|S_i=j))\gamma_i(j) \\
&= \sum_{j=1}^{L}\sum_{i=1}^{T}\left[\ln\Big(\Pr(N_i=n_i|S_i=j)\Big) + \ln\Big(\Pr(C_i=c_i|S_i=j)\Big)\right]\gamma_i(j).
\end{aligned}$$

Here n_i is the observed number of claims and c_i is the observed average claim. The Q function breaks apart again nicely because the random variables C_i and N_i are conditionally

independent given our latent state S_i. Thus the Q function becomes,

$$Q(\phi;\phi^0) = \sum_{j=1}^{L}\gamma_1(j)\ln(\pi_j) + \sum_{k=1}^{L}\sum_{j=1}^{L}\sum_{i=2}^{T}\ln(a_{s_{i-1}=k,s_i=j})\gamma_{i-1}(k,j)$$
$$+ \sum_{j=1}^{L}\sum_{i=1}^{T}\left[\ln\Big(\Pr(N_i=n_i|S_i=j)\Big) + \ln\Big(\Pr(C_i=c_i|S_i=j)\Big)\right]\gamma_i(j). \qquad (3.1)$$

Therefore with initialized values of the parameters this Q function can be calculated, leading to the second step of the EM algorithm. Next we must derive optimal values for our parameters so we can iterate.

Fortunately since the expectation of the log likelihood separated nicely the estimates for the initial state probabilities are the same as before:

$$\pi_i = \gamma_1(i), \qquad i \in \{1,...,L\}.$$

This is true of the transition probabilities as well:

$$a_{kj} = \frac{\sum_{i=2}^{T}\gamma_i(k,j)}{\sum_{i=2}^{T}\gamma_i(k)}, \qquad k,j \in \{1,...,L\}.$$

Unlike before here we need to derive estimates for the Poisson and gamma distribution parameters, $(\boldsymbol{\lambda}, \mathbf{k}, \boldsymbol{\theta})$. Starting with the Poisson parameters, take the partial derivative of Q with respect to λ_j, which refers to the λ value at the hidden state j:

$$\frac{\partial Q}{\partial \lambda_j} = \sum_{i=1}^{T}\left(\frac{n_i}{\lambda_j} - 1\right)\gamma_i(j), \qquad j \in \{1,...,L\}.$$

This is done by replacing $\Pr(N_i = n_i|S_i = j)$ by the probability mass function of the Poisson distribution, and noticing that the rest of the sum does not depend on λ_j. Setting it equal

to zero and solving gives

$$\sum_{i=1}^{T}\left(\frac{n_i}{\lambda_j}-1\right)\gamma_i(j)=0 \implies \sum_{i=1}^{T}(n_i-\lambda_j)\,\gamma_i(j)=0,$$
$$\lambda_j\sum_{i=1}^{T}\gamma_i(j)=\sum_{i=1}^{T}n_i\gamma_i(j) \implies \lambda_j=\frac{\sum_{i=1}^{T}n_i\gamma_i(j)}{\sum_{i=1}^{T}\gamma_i(j)}, \qquad j\in\{1,...,L\}.$$

This is consistent with the result from Paroli et al. [2000]. All that is left is the estimation of the parameters for the gamma distribution. Since it is continuous and therefore has probability 0 of taking a particular value we use the probability density function, as is commonly done when deriving the maximum likelihood estimates for continuous distributions. We use the following parameterization of the PDF

$$f(c_i)=\frac{c_i^{k_j-1}e^{\frac{-c_i}{\theta_j}}}{\Gamma(k_j)\theta_j^{k_j}}, \qquad j\in\{1,...,L\}, \quad k_j,\theta_j>0,$$

where c_i represents the average claim severity of the ith observation, $i\in\{1,...,T\}$, while k_j and θ_j are the parameters of the gamma distribution for hidden state j. Thus we can proceed to derive estimates for k_j and θ_j, beginning with θ_j:

$$\frac{\partial Q}{\partial\theta_j}=\sum_{i=1}^{T}\left(\frac{c_i}{\theta_j^2}-\frac{k_j}{\theta_j}\right)\gamma_i(j), \qquad j\in\{1,...,L\}.$$

This time replace $\Pr(C_i=c_i|S_i=j)$ by the density function of the gamma and disregard the rest of the sum as it does not depend on θ_j. Setting equal to zero and solving gives

$$\sum_{i=1}^{T}\left(\frac{c_i}{\theta_j^2}-\frac{k_j}{\theta_j}\right)\gamma_i(j)=0 \implies \sum_{i=1}^{T}(c_i\gamma_i(j)-k_j\theta_j\gamma_i(j))=0,$$
$$\sum_{i=1}^{T}k_j\theta_j\gamma_i(j)=\sum_{i=1}^{T}c_i\gamma_i(j) \implies \theta_j=\frac{\sum_{i=1}^{T}c_i\gamma_i(j)}{\sum_{i=1}^{T}k_j\gamma_i(j)}, \qquad j\in\{1,...,L\}.$$

This allows us to write θ_j in terms of k_j, so now we must derive an estimate for k_j . Simpli-

fying notation gives

$$\bar{c_T} = \frac{\sum_{i=1}^{T} c_i \gamma_i(j)}{\sum_{i=1}^{T} \gamma_i(j)} \implies \theta_j = \frac{\bar{c_T}}{k_j}, \qquad j \in \{1, ..., L\}.$$

Substituting this estimate for θ_j into Equation (3.1) and taking the partial derivative with respect to k_j yields

$$\frac{\partial Q}{\partial k_j} = \sum_{i=1}^{T} \left(\ln(c_i) - \frac{c_i}{\bar{c_T}} - \psi_0(k_j) - \ln\left(\frac{\bar{c_T}}{k_j}\right) + 1 \right) \gamma_i(j),$$

where $\psi_0(k_j)$ is the digamma function. Setting equal to zero does not lead to an analytical solution, therefore a numerical technique is needed. Note that the estimates of k_j and θ_j are consistent with Zhang et al. [2012] and Mohammadiha et al. [2013]. Then, as before, we need to verify that we are moving towards a local maximum. Writing the Hessian out gives

$$H = \begin{bmatrix} \frac{-\gamma_1(1)}{\pi_1^2} & 0 & \dots & 0 \\ 0 & \frac{-\gamma_1(2)}{\pi_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \sum_{i=1}^{T} \left(-\psi_1(k_K) + \frac{1}{k_K} \right) \gamma_i(K) \end{bmatrix}.$$

The diagonal represents the second derivatives with respect to the parameters of the model, i.e. $\phi = \{\mathbf{A}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \mathbf{k}, \boldsymbol{\theta}\}$. By the same reasoning as before the second derivatives with respect to a_{ij} and π_i are negative, thus we must check the second derivatives with respect to our emission probability parameters. Since we replaced θ_j by a value dependent on k_j we can disregard the second derivatives with respect to θ_j and this explains why we have a diagonal matrix, the second derivatives with respect to two different parameters is zero as before. Then, consider the second derivative of Q with respect to λ_j,

$$\frac{\partial^2 Q}{\partial \lambda_j^2} = \sum_{i=1}^{T} \left(\frac{-n_i}{\lambda_j^2} \right) \gamma_i(j), \qquad j \in \{1, ..., L\}.$$

This value is negative because n_i is positive as it represents claim counts, then $\gamma_i(j)$ is a probability thus nonnegative and $\lambda_j > 0$ by definition of the Poisson Distribution. The next

part is more complicated and will require the use of the results of others. Calculating the second derivative of Q with respect to k_j we have

$$\frac{\partial^2 Q}{\partial k_j^2} = \sum_{i=1}^{T} \left(-\psi_1(k_j) + \frac{1}{k_j} \right) \gamma_i(j), \qquad j \in \{1, ..., L\},$$

where $\psi_1(k_j)$ is the trigamma function, i.e. $\psi_1(k_j) = \psi_0'(k_j)$. Since by definition $k_j > 0$ we will only be concerned with that space. Also $\gamma_i(j) > 0$ and $\psi_1(k_j) > 0$, and for this reason all that is left to show is that $\psi_1(k_j) > \frac{1}{k_j}$.

Proof: To prove this, split the space into two sections, $(0, 1]$ and $(1,\infty)$. Dealing with the first part, $(0, 1]$, we need to rewrite the trigamma function like in Sebah and Gourdon [2002]

$$\psi_1(k_j) = \sum_{p=0}^{\infty} \frac{1}{(p + k_j)^2}, \qquad k_j \neq 0, -1, -2, ...$$

Then if we look at the first term in this sum, $\frac{1}{k_j^2}$, this is clearly larger than $\frac{1}{k_j}$ for $k_j \in (0, 1)$. Similarly when $k_j = 1$ just take the first two terms in the series. Next for the second part if we let

$$\theta_1(k_j) = (\psi_1)^{-1} \left(\frac{1}{k_j} \right) - k_j, \qquad j \in \{1, ..., L\}.$$

Batir [2007] showed that $\theta_1(k_j) > 0$ for all $k_j > 0$ and

$$\frac{1}{k_j} = \psi_1(k_j + \theta_1(k_j)), \qquad j \in \{1, ..., L\}.$$

This implies that $\psi_1(k_j) > \frac{1}{k_j}$, as we showed that $\psi_1(k_j)$ starts off larger than $\frac{1}{k_j}$ in the considered region. ■

Thus $\frac{\partial^2 Q}{\partial k_j^2} < 0$ which makes all the diagonal elements of the Hessian negative. Therefore it is negative definite and we are moving towards a local maximum. One more thing to note is that by definition we need λ_j, θ_j and k_j to be greater than zero which is easily verified as they are defined by a sum and product of strictly positive values.

3.3.2 Multiple Observation Sequences

In the case of a new policy holder there is a lack of data and therefore it is impossible to estimate the necessary parameters for the described HMM model. Luckily one can use an unsupervised learning algorithm to split the policy holders into classes/groups, i.e. clustering procedures, and leverage similar policy holders to estimate parameters for the new policy holder. In credibility theory the problem of estimating premiums is often presented by assuming that the insurance company will have a portfolio already split into different risk classes that have similar characteristics. Therefore the assumption of an insurance portfolio already split by classes is not so far fetched and adding this to the model does not require much additional work. What is needed is the machinery to estimate parameters for multiple observation sequences.

Let $\mathbf{O}$ now represent a set containing multiple observation sequences,

$$\mathbf{O} = \{O^{(1)}, O^{(2)}, ..., O^{(M)}\},$$

where

$$O^{(m)} = \{o_1^{(m)}, ..., o_{T_m}^{(m)}\},$$

and $1 \leq m \leq M$. Note that different sequences are allowed to be of varying lengths, which we can take advantage of, as policy holders beginning at different times are bound to have dissimilar sequence lengths. Let T_m denote the length of the mth sequence. One usually does not know if the sequences are independent or not and if independence is assumed and then proven not to be the case a catastrophe can occur. In either circumstance let us redefine the total output probability for multiple sequences, thanks to properties of conditional

probability we can write it in m different ways:

$$\begin{aligned}
\Pr(\mathbf{O}|\phi) &= \Pr(O^{(1)}|\phi)\Pr(O^{(2)}|O^{(1)},\phi)...\Pr(O^{(M)}|O^{(M-1)},...,O^{(1)},\phi) \\
\Pr(\mathbf{O}|\phi) &= \Pr(O^{(2)}|\phi)\Pr(O^{(3)}|O^{(2)},\phi)...\Pr(O^{(1)}|O^{(M)},...,O^{(2)},\phi) \\
&\vdots \\
\Pr(\mathbf{O}|\phi) &= \Pr(O^{(M)}|\phi)\Pr(O^{(1)}|O^{(M)},\phi)...\Pr(O^{(M-1)}|O^{(M)},O^{(M-2)},...,O^{(1)},\phi).
\end{aligned}$$

Summing the above expressions we can rewrite the total out probability as,

$$\Pr(\mathbf{O}|\phi) = \sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi), \tag{3.2}$$

where

$$\begin{aligned}
w_1 &= \frac{1}{M}\Pr(O^{(2)}|O^{(1)},\phi)...\Pr(O^{(M)}|O^{(M-1)},...,O^{(1)},\phi) \\
w_2 &= \frac{1}{M}\Pr(O^{(3)}|O^{(2)},\phi)...\Pr(O^{(1)}|O^{(M)},...,O^{(2)},\phi) \\
&\vdots \\
w_K &= \frac{1}{M}\Pr(O^{(1)}|O^{(M)},\phi)...\Pr(O^{(M-1)}|O^{(M)},O^{(M-2)},...,O^{(1)},\phi).
\end{aligned}$$

These weights are conditionals probabilities and thus capture the dependence relationship between different observation sequences. Using the above relations Li et al. [2000] showed that the Q function can be rewritten as

$$Q(\phi;\phi^k) = \sum_{m=1}^{M} w_m Q_m(\phi;\phi^k), \tag{3.3}$$

where

$$Q_m(\phi;\phi^k) = \sum_{S} \Pr(O^{(m)},S|\phi)\ln(\Pr(O^{(m)},S|\phi^k)).$$

Li et al. [2000] first built the hardware needed to derive estimates given multiple observation sequences. As they switched from expectations to probabilities let us verify that this

coincides with our previous results for the case of $M = 1$. Note that $w_1 = 1$ for $M = 1$, thus making

$$\begin{aligned}
Q_m(\phi;\phi^k) &= \sum_S Pr(O,S|\phi)\ln(\Pr(O,S|\phi^k)) \\
&= \sum_S \Pr(O,S|\phi)\left(\ln\left(\pi_{s_1}b_{s_1}(o_1)\prod_{i=2}^{T}a_{s_{i-1}s_i}b_{s_i}(o_i)\right)\right) \\
&= \sum_S \Pr(S|O,\phi)\Pr(O|\phi)\left(\ln(\pi_{s_1}) + \sum_{i=1}^{T}\ln(b_{s_i}(o_i)) + \sum_{i=2}^{T}\ln(a_{s_{i-1}s_i})\right) \\
&= \Pr(O|\phi)\Bigg(\sum_S \Pr(S|O,\phi)\ln(\pi_{s_1}) + \sum_S \Pr(S|O,\phi)\sum_{i=1}^{T}\ln(b_{s_i}(o_i)) \\
&\quad + \sum_S \Pr(S|O,\phi)\sum_{i=2}^{T}\ln(a_{s_{i-1}s_i})\Bigg) \\
&= \Pr(O|\phi)\Bigg(\sum_{j=1}^{L}\gamma_1(j)\ln(\pi_{s_1=j}) + \sum_{j=1}^{L}\sum_{i=1}^{T}\gamma_i(j)\ln(b_{s_i=j}(o_i)) \\
&\quad + \sum_{k=1}^{L}\sum_{j=1}^{L}\sum_{i=2}^{T}\gamma_{i-1}(k,j)\ln(a_{s_{i-1}=k,s_i=j})\Bigg).
\end{aligned}$$

When optimizing this equation for ϕ the $\Pr(O|\phi)$ cancels out and thus we are left with the same estimates as before.

Next we must derive estimates for our parameters with $M > 1$. Like before we can set up a Lagrange multiplier problem,

$$F(\phi^k,\boldsymbol{\kappa}) = Q(\phi;\phi^k) + \kappa_1\left(1 - \sum_{j=1}^{L}\pi_j\right) + \kappa_2\left(1 - \sum_{j=1}^{L}a_{kj}\right), \tag{3.4}$$

where $Q(\phi;\phi^k)$ is the same as Equation (3.3), $\boldsymbol{\kappa} = \{\kappa_1,\kappa_2\}$ are the Lagrange Multipliers and the constraints are the same as before

$$g_1(\boldsymbol{\pi}) = \sum_{j=1}^{L}\pi_j = 1, \tag{3.5}$$

$$g_2(a_{k1},...,a_{kL}) = \sum_{j=1}^{L}a_{kj} = 1. \tag{3.6}$$

Given a similar set up as before and we need to optimize our parameters. Commencing with the initial state probabilities

$$\frac{\partial F}{\partial \pi_i} = \sum_m w_m \Pr(O^{(m)}|\phi) \left(\frac{\gamma_1^{(m)}(i)}{\pi_i} \right) - \kappa_1, \qquad i \in \{1, ..., L\},$$

where $\gamma_t^{(m)}(i)$ refers to the $\Pr(S_t = i|O^{(m)}, \phi)$. Note the calculation for this value is the same as before it is just with respect to the mth observation sequence. Then setting equal to zero and solving for π_i,

$$\sum_m w_m \Pr(O^{(m)}|\phi) \left(\frac{\gamma_1^{(m)}(i)}{\pi_i} \right) - \kappa_1 = 0 \implies \sum_m w_m \Pr(O^{(m)}|\phi) \left(\frac{\gamma_1^{(m)}(i)}{\pi_i} \right) = \kappa_1$$
$$\sum_m w_m \Pr(O^{(m)}|\phi) \left(\gamma_1^{(m)}(i) \right) = \kappa_1 \pi_i \implies \pi_i = \frac{1}{\kappa_1} \sum_m w_m \Pr(O^{(m)}|\phi) \left(\gamma_1^{(m)}(i) \right).$$

Using our constraint 3.5 we can then solve for κ_1,

$$\sum_{j=1}^{L} \frac{1}{\kappa_1} \sum_m w_m \Pr(O^{(m)}|\phi) \left(\gamma_1^{(m)}(j) \right) = 1 \implies \kappa_1 = \sum_{j=1}^{L} \sum_m w_m \Pr(O^{(m)}|\phi) \left(\gamma_1^{(m)}(j) \right),$$
$$\kappa_1 = \sum_m w_m \Pr(O^{(m)}|\phi) \left(\sum_{j=1}^{L} \gamma_1^{(m)}(j) \right) = \sum_m w_m \Pr(O^{(m)}|\phi).$$

Thus making the estimate

$$\pi_i = \frac{\sum_m w_m \Pr(O^{(m)}|\phi) \left(\gamma_1^{(m)}(i) \right)}{\sum_m w_m \Pr(O^{(m)}|\phi)}. \tag{3.7}$$

Proceeding to the transition probabilities, a_{kj}. Taking the derivative of F with respect to a_{kj} we get

$$\frac{\partial F}{\partial a_{kj}} = \sum_m w_m \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \frac{\gamma_{i-1}^{(m)}(k, j)}{a_{kj}} \right) - \kappa_2, \qquad k, j \in \{1, ..., L\}$$

where $\gamma_i^{(m)}(k, j) = \Pr(S_i = k, S_{i+1} = j|O^{(m)}, \phi)$. Again this is similar to the previous calculation for this value just with respect to the mth observation sequence. Then to find a

critical point we set the derivative equal to zero and solve for a_{kj},

$$\sum_m w_m \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \frac{\gamma_{i-1}^{(m)}(k,j)}{a_{kj}} \right) - \kappa_2 = 0 \implies \sum_m w_m \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \frac{\gamma_{i-1}^{(m)}(k,j)}{a_{kj}} \right) = \kappa_2,$$

$$\frac{1}{\kappa_2} \sum_m w_m \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k,j) \right) = a_{kj}.$$

Then applying this to Constraint (3.6)

$$\begin{aligned}
\sum_{j=1}^{L} \frac{1}{\kappa_2} \sum_m w_m \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k,j) \right) &= 1 \\
\kappa_2 &= \sum_{j=1}^{L} \sum_m w_m \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k,j) \right) \\
\kappa_2 &= \sum_m w_m \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \sum_{j=1}^{L} \gamma_{i-1}^{(m)}(k,j) \right) \\
\kappa_2 &= \sum_m w_m \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k) \right).
\end{aligned}$$

Substituting this in to our previous solution for a_{kj} gives

$$a_{kj} = \frac{\sum_m w_m \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k,j) \right)}{\sum_m w_m \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k) \right)}. \tag{3.8}$$

Both the result for a_{kj} and π_i are consistent with Li et al. [2000]. This makes sense as the sum for F splits nicely. Now to estimate our emission probability parameters, as we are still in the auto insurance realm, let us continue with the assumption of Poisson and gamma emissions. As in Section 3.3.1, we split $b_{s_i=j}(o_i)$ by claim count and severity in our Q_m

functions,

$$Q_m(\phi;\phi^k) = \Pr(O^{(m)}|\phi)\Bigg(\sum_{j=1}^{L}\gamma_1^{(m)}(j)\ln(\pi_{s_1=j}) + \sum_{j=1}^{L}\sum_{i=1}^{T_m}\gamma_i^{(m)}(j)\Bigg[\ln\Big(\Pr(N_i=n_i^{(m)}|S_i=j)\Big)+\ln\Big(\Pr(C_i=c_i^{(m)}|S_i=j)\Big)\Bigg] + \sum_{k=1}^{L}\sum_{j=1}^{L}\sum_{i=2}^{T_m}\gamma_{i-1}^{(m)}(k,j)\ln(a_{s_{i-1}=k,s_i=j})\Bigg), \tag{3.9}$$

where $n_i^{(m)}$ and $c_i^{(m)}$ are the number of claims and the claim severity at time i for the mth observation sequence respectively. We were able to disregard the other parameters when deriving estimates for π_i and a_{kj} because this sum breaks apart nicely as we took the derivative. Then to derive an estimate for λ_j, the Poisson parameter for the jth state, let us take the derivative of F with respect to λ_j,

$$\frac{\partial F}{\partial \lambda_j} = \sum_m w_m \Pr(O^{(m)}|\phi)\sum_{i=1}^{T_m}\left(\frac{n_i^{(m)}}{\lambda_j}-1\right)\gamma_i^{(m)}(j), \qquad j\in\{1,...,L\}.$$

Setting equal to zero and solving

$$\begin{aligned}
&\sum_m w_m \Pr(O^{(m)}|\phi)\sum_{i=1}^{T_m}\left(\frac{n_i^{(m)}}{\lambda_j}-1\right)\gamma_i^{(m)}(j)=0\\
&\frac{1}{\lambda_j}\sum_m w_m \Pr(O^{(m)}|\phi)\sum_{i=1}^{T_m}\Big(n_i^{(m)}-\lambda_j\Big)\gamma_i^{(m)}(j)=0\\
&\sum_m\sum_{i=1}^{T_m}\Big(w_m \Pr(O^{(m)}|\phi)n_i^{(m)} - w_m \Pr(O^{(m)}|\phi)\lambda_j\Big)\gamma_i^{(m)}(j)=0\\
&\lambda_j\sum_m\sum_{i=1}^{T_m} w_m \Pr(O^{(m)}|\phi)\gamma_i^{(m)}(j) = \sum_m\sum_{i=1}^{T_m} w_m \Pr(O^{(m)}|\phi)n_i^{(m)}\gamma_i^{(m)}(j)\\
&\lambda_j = \frac{\sum_m\sum_{i=1}^{T_m} w_m \Pr(O^{(m)}|\phi)n_i^{(m)}\gamma_i^{(m)}(j)}{\sum_m\sum_{i=1}^{T_m} w_m \Pr(O^{(m)}|\phi)\gamma_i^{(m)}(j)}.
\end{aligned} \tag{3.10}$$

Last but not least we must optimize our gamma parameters. Taking the derivative of F

with respect to θ_j,

$$\frac{\partial F}{\partial \theta_j} = \sum_m w_m \Pr(O^{(m)}|\phi) \sum_{i=1}^{T_m} \left(\frac{c_i^{(m)}}{\theta_j^2} - \frac{k_j}{\theta_j} \right) \gamma_i^{(m)}(j), \qquad j \in \{1, ..., L\}.$$

Following the theme

$$\begin{aligned}
&\sum_m w_m \Pr(O^{(m)}|\phi) \sum_{i=1}^{T_m} \left(\frac{c_i^{(m)}}{\theta_j^2} - \frac{k_j}{\theta_j} \right) \gamma_i^{(m)}(j) = 0 \\
&\sum_m w_m \Pr(O^{(m)}|\phi) \sum_{i=1}^{T_m} \left(c_i^{(m)} - k_j \theta_j \right) \gamma_i^{(m)}(j) = 0 \\
&\sum_m \sum_{i=1}^{T_m} \left(c_i^{(m)} w_m \Pr(O^{(m)}|\phi) - k_j \theta_j w_m \Pr(O^{(m)}|\phi) \right) \gamma_i^{(m)}(j) = 0 \\
&\theta_j \sum_m \sum_{i=1}^{T_m} k_j w_m \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j) = \sum_m \sum_{i=1}^{T_m} c_i^{(m)} w_m \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j) \\
&\theta_j = \frac{\sum_m \sum_{i=1}^{T_m} c_i^{(m)} w_m \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j)}{\sum_m \sum_{i=1}^{T_m} k_j w_m \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j)}, \qquad j \in \{1, ..., L\}.
\end{aligned} \tag{3.11}$$

To simplify notation let

$$\bar{c}^{(m)} = \frac{\sum_m \sum_{i=1}^{T_m} c_i^{(m)} w_m \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j)}{\sum_m \sum_{i=1}^{T_m} w_m \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j)}$$

then

$$\theta_j = \frac{\bar{c}^{(m)}}{k_j}, \qquad j \in \{1, ..., L\}.$$

Substituting this into Equation (3.4) and taking the derivative with respect to k_j yields

$$\frac{\partial F}{\partial k_j} = \sum_m w_m \Pr(O^{(m)}|\phi) \sum_{i=1}^{T_m} \gamma_i^{(m)}(j) \left(\ln(c_i^{(m)}) - \frac{c_i^{(m)}}{\bar{c}^{(m)}} - \psi_0(k_j) - \ln\left(\frac{\bar{c}^{(m)}}{k_j} \right) + 1 \right), \tag{3.12}$$

for $j \in \{1, ..., L\}$. When setting equal to zero this is again impossible to solve analytically and thus a numerical procedure is needed. Next we must check if we are moving towards a

local maximum. Taking the second derivatives gives

$$\frac{\partial^2 F}{\partial \pi_i^2} = \sum_m w_m \Pr(O^{(m)}|\phi) \left(\frac{-\gamma_1^{(m)}(i)}{\pi_i^2} \right), \tag{3.13}$$

$$\frac{\partial^2 F}{\partial a_{kj}^2} = \sum_m w_m \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \frac{-\gamma_{i-1}^{(m)}(k,j)}{a_{kj}^2} \right), \tag{3.14}$$

$$\frac{\partial^2 F}{\partial \lambda_j^2} = \sum_m w_m \Pr(O^{(m)}|\phi) \sum_{i=1}^{T_m} \frac{-n_i^{(m)}}{\lambda_1^2} \gamma_i^{(m)}(j), \tag{3.15}$$

$$\frac{\partial^2 F}{\partial k_j^2} = \sum_m w_m \Pr(O^{(m)}|\phi) \sum_{i=1}^{T_m} \gamma_i^{(m)}(j) \left(-\psi_1(k_j) + \frac{1}{k_j} \right). \tag{3.16}$$

The Hessian therefore becomes a diagonal matrix, as the second derivatives with respect to two different parameters are zero, with the previous second derivatives as it's diagonals. Note that $\psi_1(k_j)$ and $\psi_0(k_j)$ represent the same functions as before, trigamma and digamma respectively. To ensure that the critical point is a local maximum the Hessian must be negative definite. Equation (3.15) is negative because $w_m > 0$, as it is a probability, and $\Pr(O^{(m)}|\phi) > 0$. The rest of values in Equation (3.15) are greater than zero by the same logic as before. This reasoning holds true for Equation (3.14) and (3.13) as well. Then for Equation (3.16), this value is negative as by the same logic as the proof from Section 3.3.1. The value is just being multiplied by $\sum_m w_m \Pr(O^{(m)}|\phi)$ which is positive as it represents a probability. Therefore the algorithm is moving in the direction of a local maximum.

Considering multiple observation sequences has the advantage of modeling a dependence relation between policy holders, w_m. In the past it was often thought that the observations were independent but in light of what has transpired in the past couple years that seems not to be the case. Companies and government regulators have been transitioning to predictive models that capture dependencies to help safeguard from economic catastrophes. With this being said let us point out two special cases of dependence. First, assuming independent observation sequences allows for nice simplifications of the model parameters. The total output probability from (3.2) can be rewritten as

$$\Pr(\mathbf{O}|\phi) = \prod_{m=1}^{M} \Pr(O^{(m)}|\phi),$$

and the weights as

$$w_m = \frac{\Pr(\mathbf{O}|\phi)}{M \Pr(O^{(m)}|\phi)}, \qquad m \in \{1, ..., M\}. \tag{3.17}$$

Substituting (3.17) into (3.7) and (3.8) gives

$$\pi_i = \frac{\sum_m \frac{\Pr(\mathbf{O}|\phi)}{M \Pr(O^{(m)}|\phi)} \Pr(O^{(m)}|\phi) \gamma_1^{(m)}(i)}{\sum_m \frac{\Pr(\mathbf{O}|\phi)}{M \Pr(O^{(m)}|\phi)} \Pr(O^{(m)}|\phi)} = \frac{1}{M} \sum_m \gamma_1^{(m)}(i), \qquad i \in \{1, ..., L\},$$

$$a_{kj} = \frac{\sum_m \frac{\Pr(\mathbf{O}|\phi)}{M \Pr(O^{(m)}|\phi)} \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k, j) \right)}{\sum_m \frac{\Pr(\mathbf{O}|\phi)}{M \Pr(O^{(m)}|\phi)} \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k) \right)} = \frac{\sum_m \sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k, j)}{\sum_m \sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k)}, \qquad k, j \in \{1, ..., L\}.$$

These results coincide with Li et al. [2000]. Then replacing again (3.17) into (3.10) and (3.11) yields

$$\lambda_j = \frac{\sum_m \sum_{i=1}^{T_m} \frac{\Pr(\mathbf{O}|\phi)}{M \Pr(O^{(m)}|\phi)} \Pr(O^{(m)}|\phi) n_i^{(m)} \gamma_i^{(m)}(j)}{\sum_m \sum_{i=1}^{T_m} \frac{\Pr(\mathbf{O}|\phi)}{M \Pr(O^{(m)}|\phi)} \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j)} = \frac{\sum_m \sum_{i=1}^{T_m} n_i^{(m)} \gamma_i^{(m)}(j)}{\sum_m \sum_{i=1}^{T_m} \gamma_i^{(m)}(j)}, \qquad j \in \{1, ..., L\},$$

$$\theta_j = \frac{\sum_m \sum_{i=1}^{T_m} c_i^{(m)} \frac{\Pr(\mathbf{O}|\phi)}{M \Pr(O^{(m)}|\phi)} \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j)}{\sum_m \sum_{i=1}^{T_m} k_j \frac{\Pr(\mathbf{O}|\phi)}{M \Pr(O^{(m)}|\phi)} \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j)} = \frac{\sum_m \sum_{i=1}^{T_m} c_i^{(m)} \gamma_i^{(m)}(j)}{\sum_m \sum_{i=1}^{T_m} k_j \gamma_i^{(m)}(j)}, \qquad j \in \{1, ..., L\}.$$

Lastly substituting (3.17) into (3.12),

$$\begin{aligned}
\frac{\partial F}{\partial k_j} &= \sum_m w_m \Pr(O^{(m)}|\phi) \sum_{i=1}^{T_m} \gamma_i^{(m)}(j) \left(\ln(c_i^{(m)}) - \frac{c_i^{(m)}}{\bar{c}^{(m)}} - \psi_0(k_j) - \ln\left(\frac{\bar{c}^{(m)}}{k_j} \right) + 1 \right) \\
&= \sum_m \frac{\Pr(\mathbf{O}|\phi)}{M \Pr(O^{(m)}|\phi)} \Pr(O^{(m)}|\phi) \sum_{i=1}^{T_m} \gamma_i^{(m)}(j) \left(\ln(c_i^{(m)}) - \frac{c_i^{(m)}}{\bar{c}^{(m)}} - \psi_0(k_j) - \ln\left(\frac{\bar{c}^{(m)}}{k_j} \right) + 1 \right) \\
&= \frac{\Pr(\mathbf{O}|\phi)}{M} \sum_m \sum_{i=1}^{T_m} \gamma_i^{(m)}(j) \left(\ln(c_i^{(m)}) - \frac{c_i^{(m)}}{\bar{c}^{(m)}} - \psi_0(k_j) - \ln\left(\frac{\bar{c}^{(m)}}{k_j} \right) + 1 \right),
\end{aligned}$$

for $j \in \{1, ..., L\}$, where

$$\bar{c}^{(m)} = \frac{\sum_m \sum_{i=1}^{T_m} c_i^{(m)} \gamma_i^{(m)}(j)}{\sum_m \sum_{i=1}^{T_m} \gamma_i^{(m)}(j)}.$$

Note that a numerical procedure is still required to solve for k_j when determining for a

critical point. The second special case is assuming that there is a uniform dependence across all observation sequences. This allows us to rewrite the weights as

$$w_m = a, \qquad m \in \{1, ..., M\}, \tag{3.18}$$

where a is a constant that does not depend on m. Using this (3.13) and (3.14) become

$$\pi_i = \frac{\sum_m a \Pr(O^{(m)}|\phi) \left(\gamma_1^{(m)}(i)\right)}{\sum_m a \Pr(O^{(m)}|\phi)} = \frac{\sum_m \Pr(O^{(m)}|\phi) \left(\gamma_1^{(m)}(i)\right)}{\sum_m \Pr(O^{(m)}|\phi)}, \qquad i \in \{1, ..., L\}$$
$$a_{kj} = \frac{\sum_m a \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k,j)\right)}{\sum_m a \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k)\right)} = \frac{\sum_m \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k,j)\right)}{\sum_m \Pr(O^{(m)}|\phi) \left(\sum_{i=2}^{T_m} \gamma_{i-1}^{(m)}(k)\right)},$$

for $k, j \in \{1, ..., L\}$. This again accords with the Li et al. [2000]. Proceeding to the other parameters, substituting (3.18) into (3.10) and (3.11) gives

$$\lambda_j = \frac{\sum_m \sum_{i=1}^{T_m} a \Pr(O^{(m)}|\phi) n_i^{(m)} \gamma_i^{(m)}(j)}{\sum_m \sum_{i=1}^{T_m} a \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j)} = \frac{\sum_m \sum_{i=1}^{T_m} \Pr(O^{(m)}|\phi) n_i^{(m)} \gamma_i^{(m)}(j)}{\sum_m \sum_{i=1}^{T_m} \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j)}, \qquad j \in \{1, ..., L\},$$
$$\theta_j = \frac{\sum_m \sum_{i=1}^{T_m} c_i^{(m)} a \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j)}{\sum_m \sum_{i=1}^{T_m} k_j a \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j)} = \frac{\sum_m \sum_{i=1}^{T_m} \Pr(O^{(m)}|\phi) c_i^{(m)} \gamma_i^{(m)}(j)}{\sum_m \sum_{i=1}^{T_m} \Pr(O^{(m)}|\phi) k_j \gamma_i^{(m)}(j)}, \qquad j \in \{1, ..., L\}.$$

Then using (3.18) in (3.12) yields

$$\frac{\partial F}{\partial k_j} = \sum_m a \Pr(O^{(m)}|\phi) \sum_{i=1}^{T_m} \gamma_i^{(m)}(j) \left(\ln(c_i^{(m)}) - \frac{c_i^{(m)}}{\bar{c}^{(m)}} - \psi_0(k_j) - \ln\left(\frac{\bar{c}^{(m)}}{k_j}\right) + 1 \right),$$

for $j \in \{1, ..., L\}$, where

$$\bar{c}^{(m)} = \frac{\sum_m \sum_{i=1}^{T_m} \Pr(O^{(m)}|\phi) c_i^{(m)} \gamma_i^{(m)}(j)}{\sum_m \sum_{i=1}^{T_m} \Pr(O^{(m)}|\phi) \gamma_i^{(m)}(j)}.$$

As before a numerical procedure is needed to find the critical point. Note that the actuary can set these dependencies how wanted, one is not forced to choose one of the cases described.

The multiple observation twist makes it easy to use the HMM model in credibility. One can use the whole portfolio and derive an estimate for the next time period claims and then

derive the estimate for risk class or individual policy holder. Using these two quantities one can take a weighted sum, as is done in credibility, and use that as a pure premium. Thus the application of the proposed HMM model would be an easy transition for most actuaries as they would be able to easily incorporate some methods that they already apply.

3.4 Prediction

After computing the model parameters starts the part of forecasting future values. First begin with the expected number of claims for the next time period. We can use the mechanisms we defined earlier to find the probabilities of our terminal states:

$$\gamma_T(j) = \Pr(S_T = j|\mathbf{O}, \phi), \qquad j \in \{1, ..., L\}.$$

Then we can use this probability and propagate forward with our estimated transition probabilities to the next time period. Thus allowing us to find the expected number of claims, using the tower property:

$$\begin{aligned}
\mathrm{E}(N_{T+1}|\phi) &= \mathrm{E}(\mathrm{E}(N_{T+1}|S_{T+1}, \phi)) \\
&= \mathrm{E}(\lambda_j) \\
&= \sum_j \lambda_j \Pr(S_{T+1} = j) \\
&= \sum_i \sum_j \lambda_j \gamma_T(i) a_{ij}.
\end{aligned}$$

The expected claim severities are calculated in the same manner except for λ_j, which is the mean of the Poisson, here is replaced by the mean of the gamma $k_j\theta_j$,

$$\mathrm{E}(C_{T+1}|\phi) = \sum_i \sum_j k_j \theta_j \gamma_T(i) a_{ij}.$$

These values can be used as estimates of pure premiums for the next time period and then a company can apply the appropriate premium loading.

After short term planning comes the question of long term forecasting. The further down

the road you look the more computationally expensive the calculation becomes. Let us define the probability of the first step after T,

$$\Pr(S_{T+1} = j) = \sum_i \gamma_T(i) a_{ij}.$$

This result agrees with the previous derivation of the $T+1$ estimates. Then the probability of the next step becomes,

$$\Pr(S_{T+2} = j) = \sum_i \Pr(s_{T+1} = i) a_{ij}.$$

Thus if we let $r = 3, 4, ...$ then

$$\Pr(S_{T+r} = j) = \sum_i \Pr(S_{T+r-1} = i) a_{ij}.$$

Making it possible to estimate the probability of being in state j for any future time $T+r$. This makes the expected values of future total severity and claim counts calculable:

$$\mathrm{E}(N_{T+r}|\phi) = \sum_j \lambda_j \Pr(S_{T+r} = j),$$
$$\mathrm{E}(C_{T+r}|\phi) = \sum_j k_j \theta_j \Pr(S_{T+r} = j).$$

The algorithm becomes more expensive when we increase the number of states but thanks to the Markov property if we know the probability distribution of the previous hidden state we can iterate forward easily. Thus one would store these probabilities while progressing through the HMM.

Together with the pure premium actuaries often like to estimate intervals around these point estimates. This is because it is often extremely unlikely that the next observation will be the same as the estimated one. Using the distributions for all the states we can build intervals of any percentage for any future time period. The following result can be viewed as creating risk measures for HMMs. Start with the interval for the claim count of the time period $T+r$, where r is as defined above. We want to find the point, call it b, at which we

capture $1-\alpha$ of the possible values of claim counts given our model:

$$b = \inf\left\{b \in \mathbb{N} : \Pr(N_{T+r} \leq b) > 1-\alpha\right\}.$$

The distribution for N_{T+r} becomes a weighted sum of the different Poissons which we can write down and thus solve for b. Let $f_{N_{T+r}}(n_{T+r})$ denote the probability mass function and $F_{N_{T+r}}(n_{T+r})$ be the cumulative distribution function (CDF) for N_{T+r}. One can write $f_{N_{T+r}}(n_{T+r})$ as

$$f_{N_{T+r}}(n_{T+r}) = \sum_j \frac{\lambda_j^{n_{T+r}} e^{\lambda_j}}{n_{T+r}!} \Pr(S_{T+r} = j).$$

Using this one can then write the CDF:

$$F_{N_{T+r}}(n_{T+r}) = \sum_{i=0}^{n_{T+r}} \sum_j \frac{\lambda_j^i e^{\lambda_j}}{i!} \Pr(S_{T+r} = j). \tag{3.19}$$

Thus b becomes the lowest value for which (3.19) is at least $1-\alpha$. Then proceeding in the same manner for severities, let d be

$$d = \inf\left\{d \in \mathbb{R} : \Pr(C_{T+r} \leq d) > 1-\alpha\right\}.$$

Also let $f_{C_{T+r}}(c_{T+r})$ denote the probability mass function:

$$f_{C_{T+r}}(c_{T+r}) = \sum_j \frac{c_{T+r}^{k_j-1} e^{\frac{-c_{T+r}}{\theta_j}}}{\Gamma(k_j)\theta_j^{k_j}} \Pr(S_{T+r} = j),$$

and then the CDF, $F_{C_{T+r}}(c_{T+r})$, can be written as

$$F_{C_{T+r}}(c_{T+r}) = \int_0^{c_{T+r}} \sum_j \frac{x^{k_j-1} e^{\frac{-x}{\theta_j}}}{\Gamma(k_j)\theta_j^{k_j}} \Pr(S_{T+r} = j) dx. \tag{3.20}$$

Thus d is the lowest such value for which (3.20) is at least $1-\alpha$. Unfortunately both the sum and integral for the Poisson and gamma CDF respectively do not simplify nicely. The

procedure described above provides a method for determining the value at risk, VaR, using an HMM. Note that in practice one can choose the risk measure that one deems relevant, CVaR, EVaR or etc. These risk measures are useful for regulatory institutions to prevent economic catastrophe.

3.5 Conclusion

The above results describe how to create an HMM for auto insurance and then proceeds to derive values, using this model, that actuaries often consider when making decisions. The above model, Poisson-gamma HMM, has never been proposed in actuarial science nor in another field and can be used to estimate any data that exhibits a time series Poisson-gamma distribution. The model does not need to be relegated to actuarial science. The model was conceived as trying to capture the hidden ability of a policy holder's capacity to drive. Even though some current actuarial models, based on GLMs, have a bonus-malus variable that tries to capture the hidden driving ability of a policyholder they do not provide a model that intuitively and dynamically captures this effect. Instead they are forced to create an explanatory variable as a proxy to capture this effect thus forcibly adding it to their model. The HMM model provides a more realistic interpretation of what the actuary sees and thus better describes the real life situation without adding to the model complexity.

Chapter 4

HMM-GLM Hybrid

4.1 Introduction

Thus far the ideas of HMMs and GLMs have been introduced separately. This chapter combines the two and introduces a HMM-GLM hybrid, which has been proposed before in Fan [2015]. Unlike the proposal in Fan [2015] the model described here has two emissions, one Poisson and the other gamma. Fan [2015] derived a HMM-GLM with one GLM emission. This proposal is similar to that in Figure 3.1, except that after determining the hidden state there is a set of covariates used to estimate the emissions. The HMM-GLM relies on estimation techniques used for both HMMs and GLMs.

4.2 Definitions

The first step is to simplify the notation. Let $\mathbf{W}$ and $\mathbf{U}$ denote the matrices of coefficients for C_t and N_t, respectively, in Figure 3.1. Assuming that the same number of covariates are used to estimate C_t and N_t makes

$$\mathbf{W} = \begin{bmatrix} w_{10} & w_{11} & \dots & w_{1p} \\ w_{20} & w_{21} & \dots & w_{2p} \\ \vdots & \vdots & \dots & \vdots \\ w_{L0} & x_{L1} & \dots & x_{Lp} \end{bmatrix}, \qquad \mathbf{U} = \begin{bmatrix} u_{10} & u_{11} & \dots & u_{1p} \\ u_{20} & u_{21} & \dots & u_{2p} \\ \vdots & \vdots & \dots & \vdots \\ u_{L0} & u_{L1} & \dots & u_{Lp} \end{bmatrix}.$$

Note that both these matrices are $L \times (p+1)$. This can be adapted to dissimilar matrices but similar ones allow for simpler notation. Also let $R_t = (R_{t,1}, ..., R_{t,L})$ denote a $1 \times L$ vector with $R_{t,i} = 1$, if $S_t = i$, and zero otherwise. Please note that all other definitions stand as before.

4.3 Parameter Estimation

Combining GLMs with HMMs adds parameters to estimate and also another hyper-parameter known as the link function. The derivation of the coefficients changes when choosing different link functions. There are two cases often considered, the canonical link function and other link functions. Also one can adapt HMMs to multiple observation sequences as before.

4.3.1 Single Observation Sequence

Before estimating the parameters the likelihood function, given complete data, must be rewritten to represent the new model. Let g_1 and g_2 be the probability density functions of the Poisson and gamma respectively. Then,

$$L_T^c(\phi) = \pi_{s_1} b_{s_1}(o_1) \prod_{i=2}^{T} a_{s_{i-1}s_i} b_{s_i}(o_i),$$

$$\ln(L_T^c(\phi)) = \ln(\pi_{s_1}) + \sum_{i=1}^{T} \ln(b_{s_i}(o_i)) + \sum_{i=2}^{T} \ln(a_{s_{i-1}s_i}).$$

Then taking the expectation the Q function becomes,

$$\begin{aligned} Q(\phi; \phi^k) &= \mathrm{E}(\ln(L_T^c(\phi))) \\ &= \sum_{j=1}^{L} \ln(\pi_{s_1=j})\gamma_1(j) + \sum_{j=1}^{L}\sum_{i=1}^{T} \ln(b_{s_i=j}(o_i))\gamma_i(j) \\ &\quad + \sum_{j=1}^{L}\sum_{k=1}^{L}\sum_{i=2}^{T} \ln(a_{s_{i-1}s_i})\gamma_{i-1}(k,j). \end{aligned} \tag{4.1}$$

Notice how in (4.1) the maximization of Q with respect to π_i and a_{ij} do not depend on the part of the sum with $\ln(b_{s_i=j}(o_i))$. Therefore the estimates from the previous chapters are still valid, however the forward and backward variable will need to be altered slightly.

Consider first the case of the canonical link function, one must then estimate $\mathbf{W}$, $\mathbf{U}$, τ_1 and τ_2. Note that one must estimate different τ_1 and τ_2 corresponding to the gamma and Poisson GLM respectively. Disregarding the nonessential parts of the sum, the optimization problem for $\mathbf{W}$ can be rewritten as

$$\begin{aligned}
\max_{\mathbf{W}} Q(\phi;\phi^k) &= \max_{\mathbf{W}} \sum_{j=1}^{L}\sum_{i=1}^{T} \ln(b_{s_i=j}(o_i))\gamma_i(j) \\
&= \max_{\mathbf{W}} \sum_{j=1}^{L}\sum_{i=1}^{T} \gamma_i(j)\left(\ln(c(y_i,\tau_1)) + \left\{\frac{y_i\zeta_i - p(\zeta_i)}{\tau_1}\right\}\right) \\
&= \max_{\mathbf{W}} \sum_{j=1}^{L}\sum_{i=1}^{T} \gamma_i(j)\left(\left\{\frac{y_i\zeta_i - p(\zeta_i)}{\tau_1}\right\}\right). \qquad (4.2)
\end{aligned}$$

Note that here $b_{s_i=j}(o_i)$ splits into a sum of two parts, the other not depending on $\mathbf{W}$ and in this case $\zeta_i = \mathbf{x}_i\mathbf{W}'R_i'$, where $\mathbf{x}_i$ is the ith row of $\mathbf{X}$ corresponding to the covariates at time step i. This expression is very similar to the MLE for GLMs except for the $\gamma_i(j)$ term. Plugging the canonical inverse link function for the gamma and applying the chain rule to (4.2) yields

$$\mathbf{X}'\mathbf{\Gamma}_k\mathbf{y} = \mathbf{X}'\mathbf{\Gamma}_k\boldsymbol{\mu}_k, \qquad k \in \{1,..,L\}, \qquad (4.3)$$

where $\mu_{t,k} = g_1^{-1}(\mathbf{x}_t\mathbf{w}_k') = \frac{1}{\mathbf{x}_t\mathbf{w}_k}$, $\boldsymbol{\mu}_k = (\mu_{1,k},...,\mu_{T,k})$, and $\mathbf{\Gamma}_k = diag(\gamma_1(k),...,\gamma_T(k))$. Note that (4.3) is used to estimate the coefficients for hidden state k and the subscript to the link function was added to distinguish the links for the gamma and Poisson. Proceeding along the same steps one can show that

$$\mathbf{X}'\mathbf{\Gamma}_k\mathbf{y} = \mathbf{X}'\mathbf{\Gamma}_k\mathbf{m}_k, \qquad k \in \{1,..,L\}, \qquad (4.4)$$

where $m_{t,k} = g_2^{-1}(\mathbf{x}_t\mathbf{u}_k') = \exp(\mathbf{x}_t\mathbf{u}_k)$ and $\mathbf{m}_k = (m_{1,k},...,m_{T,k})$. One can then apply Fisher's scoring, Newton-Raphson, or other numerical procedures to solve for $\mathbf{u}_k$ and $\mathbf{w}_k$. In the case

of the gamma distribution one might not want to choose the canonical link as it can map the mean outside its support.

The procedure for non-canonical links is identical. This produces the following likelihood equations of

$$\mathbf{X}'\mathbf{\Gamma}_k\mathbf{Z}_k\mathbf{G}_k\mathbf{y} = \mathbf{X}'\mathbf{\Gamma}_k\mathbf{Z}_k\mathbf{G}_k\boldsymbol{\mu}_k, \qquad k \in \{1, .., L\},$$

where $\mathbf{Z}_k = diag(z_{1,k}, ..., z_{T,k})$, $z_{t,k} = \left[v(\mu_{t,k})g^2_{1,\mu}(\mu_{t,k})\right]^{-1}$, $v(\mu_{t,k}) = \mathrm{V}(y_t)/\tau_1^2$ is the variance function, $\mathbf{G}_k = diag(g_{1,\mu}(\mu_{1,k}), ..., g_{1,\mu}(\mu_{T,k}))$ and $g_{1,\mu} = \partial g/\partial\mu$.

The result for the Poisson coefficients are of the same form,

$$\mathbf{X}'\mathbf{\Gamma}_k\mathbf{Z}_k\mathbf{G}_k\mathbf{y} = \mathbf{X}'\mathbf{\Gamma}_k\mathbf{Z}_k\mathbf{G}_k\boldsymbol{m}_k, \qquad k \in \{1, .., L\},$$

where $g_{1,\mu}$ is replaced by $g_{2,\mu}$ for the other link function and τ_1 replaced by τ_2. Arriving at these two estimates one still needs to choose a numerical procedure to evaluate the coefficients and estimate the τ_i's.

Using a moment estimator for τ_i, and its relation with y_t, McCullagh and Nelder [1989] showed that

$$\hat{\tau} = \frac{1}{n-p}\sum_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}.$$

This expression can be adapted to the HMM-GLM setting,

$$\tau_1 = \frac{1}{T_s - (p+1)}\sum_{l=1}^{L}\sum_{j=1}^{T}\frac{(y_j - \hat{\mu}_{j,l})^2}{v(\hat{\mu}_{j,l})}\gamma_j(l), \tag{4.5}$$

and

$$\tau_2 = \frac{1}{T - (p+1)}\sum_{l=1}^{L}\sum_{j=1}^{T}\frac{(y_j - \hat{m}_{j,l})^2}{v(\hat{m}_{j,l})}\gamma_j(l), \tag{4.6}$$

where T_s is the number of observations of the gamma. Note that $T_s \leq T$. These derivations take advantage of the HMM-GLM model, that is given a state S_t the severity and count, C_t

and N_t, are independent of all other observations and of the hidden states. For more details please refer to Fan [2015].

Given our new model the forward and backward variables need to be redefined. In place of $b_i(O_t)$ put $f(N_t|S_t)h(C_t|S_t)$, where f and h are the conditional distributions given a hidden state. Note that these values are dependent on the coefficients. Thus the new iterative rules become

$$\alpha_t(j) = \sum_{i=1}^{L} \left(\alpha_{t-1}(i)a_{ij}\right) f(N_t|S_t = j)h(C_t|S_t = j), \qquad j \in \{1, ..., L\},$$

and

$$\beta_t(j) = \sum_{i=1}^{L} a_{ji} f(N_{t+1}|S_{t+1} = i)h(C_{t+1}|S_{t+1} = i)\beta_{t+1}(i), \qquad j \in \{1, ..., L\}.$$

In addition to these alterations one needs to change the form of $\gamma_t(i, j)$. Using the same reasoning as before,

$$\begin{aligned}\gamma_t(i, j) &= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\Pr(\mathbf{O}|\phi)} \\ &= \frac{\alpha_t(i)a_{ij}f(N_{t+1}|S_{t+1} = i)h(C_{t+1}|S_{t+1} = i)\beta_{t+1}(j)}{\Pr(\mathbf{O}|\phi)}.\end{aligned}$$

All previous derivations involving these values are the same since the Q function broke apart nicely.

This summarizes some of the results of Fan [2015]. If interested this PhD thesis also goes on to approximate the Kullback-Leibler divergence of a HMM-GLM. They use it iteratively to choose a model and illustrate this in a simulation study at the end.

4.3.2 Multiple Observation Sequences

As has been demonstrated one can generalize the HMM model to multiple observation sequences. In Fan [2015] he does not give an interpretation to the model and thus does not consider multiple observation sequences. An actuary would want to build a multiple sequence model for new policy holders and young police holders whose Markov chain are not

long enough, or nonexistent, to estimate parameters accurately. Firstly the case of a general HMM-GLM will be shown and then an adaption to the considered insurance specific model.

Recalling the Q function from multiple sequences (3.9),

$$\begin{aligned} Q(\phi;\phi^k) &= \sum_{m=1}^{M} w_m Q_m(\phi;\phi^k) \\ &= \sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi)\Bigg(\sum_{j=1}^{L}\gamma_1^{(m)}(j)\ln(\pi_{s_1=j}) + \sum_{j=1}^{L}\sum_{i=1}^{T_m}\gamma_i^{(m)}(j)\ln(b_{s_i=j}(o_i)) \\ &\quad + \sum_{k=1}^{L}\sum_{j=1}^{L}\sum_{i=2}^{T_m}\gamma_{i-1}^{(m)}(k,j)\ln(a_{s_{i-1}=k,s_i=j})\Bigg). \end{aligned}$$

The derivations for the transition probabilities and initial state probabilities do not depend on the emissions and thus the formulas from before are still valid. Therefore the MLE for the coefficients and an estimate for the dispersion parameters complete the model. Like before the forward and backward variables need to be redefined. Considering the part of Q that depends on $\mathbf{W}$,

$$\max_{\mathbf{W}} Q(\phi;\phi^k) = \max_{\mathbf{W}} \sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi) \sum_{j=1}^{L}\sum_{i=1}^{T_m}\gamma_i^{(m)}(j)\left(\exp\left\{\frac{y_{i,m}\zeta_{i,m} - p(\zeta_{i,m})}{\tau_1}\right\}\right).$$

This expression is very similar to the one for a single sequence from the previous section, except for the extra sum out front. Assuming that one is using the canonical link function and using the chain rule like before one can show that the estimators are of this form

$$\sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi)\mathbf{X}_m'\mathbf{\Gamma}_{k,m}\mathbf{y}_m = \sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi)\mathbf{X}_m'\mathbf{\Gamma}_{k,m}\boldsymbol{\mu}_{k,m}, \tag{4.7}$$

where the subscript m denotes the observation sequence and the variables are the same as (4.3) except that the appropriate ones need to be denoted with a subscript m.

A similar process can be followed to derive the estimates for the non-canonical case. Thus

making the likelihood equations,

$$\sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{Z}_{k,m}\mathbf{G}_{k,m}\mathbf{y}_m = \sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{Z}_{k,m}\mathbf{G}_{k,m}\boldsymbol{\mu}_{k,m}. \quad (4.8)$$

Again $\mathbf{Z}$ and $\mathbf{G}$ need to be subscripted by m to signify the different sequences. Using the above equations one can calculate the parameters needed for a HMM-GLM model.

The w_m captures the dependencies between the observation sequences and the prudent actuary might want to try a finite number of options and choose the best, cross-validation. Two cases will be shown as they are often considered. Independence is often assumed, justifiably or not, which in turn provides for nice simplifications. The HMM-GLM model is no different. Recall from (3.17) that assuming independence implies

$$w_m = \frac{\Pr(\mathbf{O}|\phi)}{M \Pr(O^{(m)}|\phi)}, \qquad m \in \{1, ..., M\}.$$

Thus making the likelihood equations in (4.7) and (4.8)

$$\sum_{m=1}^{M} \mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{y}_m = \sum_{m=1}^{M} \mathbf{X}'_m\mathbf{\Gamma}_{k,m}\boldsymbol{\mu}_{k,m},$$
$$\sum_{m=1}^{M} \mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{Z}_{k,m}\mathbf{G}_{k,m}\mathbf{y}_m = \sum_{m=1}^{M} \mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{Z}_{k,m}\mathbf{G}_{k,m}\boldsymbol{\mu}_{k,m}.$$

Next assuming that the sequences are uniformly dependent implies that

$$w_m = a, \qquad m \in \{1, ..., M\},$$

where a is constant. Inputting this into (4.7) and (4.8) yields,

$$\sum_{m=1}^{M} \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{y}_m = \sum_{m=1}^{M} \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\boldsymbol{\mu}_{k,m},$$
$$\sum_{m=1}^{M} \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{Z}_{k,m}\mathbf{G}_{k,m}\mathbf{y}_m = \sum_{m=1}^{M} \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{Z}_{k,m}\mathbf{G}_{k,m}\boldsymbol{\mu}_{k,m}.$$

Both cases provide for nice simplifications but in practice these assumptions probably do

not hold. In applications, one is not restricted to either case and can model any dependence that seems reasonable.

Following the theme, the model will now be adapted to Poisson and gamma emissions. Keeping with the same notation as before the likelihood equations become

$$\sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{y}_m = \sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\boldsymbol{\mu}_{k,m}, \qquad k \in \{1,..,L\},$$
$$\sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{y}_m = \sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{m}_{k,m}, \qquad k \in \{1,..,L\},$$

when using the canonical link function. When using the non-canonical link function these change to

$$\sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{Z}_{k,m}\mathbf{G}_{k,m}\mathbf{y}_m = \sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{Z}_{k,m}\mathbf{G}_{k,m}\boldsymbol{\mu}_{k,m},$$
$$\sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{Z}_{k,m}\mathbf{G}_{k,m}\mathbf{y}_m = \sum_{m=1}^{M} w_m \Pr(O^{(m)}|\phi)\mathbf{X}'_m\mathbf{\Gamma}_{k,m}\mathbf{Z}_{k,m}\mathbf{G}_{k,m}\mathbf{m}_{k,m},$$

for $k \in \{1,..,L\}$. Note one must be careful to change the link function appropriately changing $\mathbf{Z}_{k,m}$ and $\mathbf{G}_{k,m}$ in the two above lines. The solutions can also be adapted for the two special cases of uniform dependence and independence. Also the estimate for the dispersion parameter changes,

$$\hat{\tau} = \frac{1}{\widetilde{T} - (p+1)} \sum_{m=1}^{M} \sum_{l=1}^{L} \sum_{j=1}^{T_m} \frac{\left(y_j^{(m)} - \hat{\mu}_{j,l}^{(m)}\right)^2}{v\left(\hat{\mu}_{j,l}^{(m)}\right)} \gamma_i^{(m)}(l),$$

where

$$\widetilde{T} = \sum_{m=1}^{M} T_m.$$

Note the values for $\widetilde{T}$ should be different for the Poisson and gamma emissions as before.

4.4 Conclusion

The considered model outputs different sets of coefficients depending on the hidden state. Given that one does not see the latent state to make predictions one has to marginalize over the hidden state and the covariates. It also becomes harder to make long term forecasts as one can probably not collect accurate covariates further in the future, as policy holders are bound to move, or values for other covariates change as well. Assuming one can gather covariates all the results from Section 3.4 are still valid. The work described extends the previous work of Fan [2015] and derives a model specific to insurance. There exists other methods beside MLE to derive estimates that can be used.

Chapter 5

Numerical Implementation

5.1 Introduction

Predictive modeling can be tricky at the implementation stage. The theory can be very nice but if the model cannot be used on a computer then this is a problem as then applications of this model cannot be realized. Many packages already exist, in R, Python, MatLab, etc., that will estimate the parameters to build an HMM. Given the model described in Figure 3.1 there exists no code to estimate the parameters. This chapter describes the implementation steps of the algorithms required to estimate the model parameters and simulate data to see their theoretical properties. These algorithms were derived with the assistance of R Development Core Team [2008], Shen [2008], Soetaert [2009] and Soetaert and Herman [2009].

5.2 Implementation Issues

Issues can arise when implementing the Baum-Welch algorithm to build an HMM model; one such issue is known as the scaling problem. This occurs when a data set has a large time horizon, say T. Recall the forward variable from Chapter 2, $\alpha_t(i) = \Pr(O_1, ..., O_t, s_t = i|\phi)$, which was recursively defined as,

1. Initialization:

$\alpha_1(i) := \pi_i b_i(O_1), \qquad i \in \{1, ..., L\}.$

2. Recursion for $t = 2, ..., T$:
   $\alpha_t(j) := \sum_i (\alpha_{t-1}(i) a_{ij})\, b_j(O_t), \qquad j \in \{1, ..., L\}.$

This value involves recursively multiplying together probabilities whose values are less than one. Thus the summation goes to zero exponentially fast as t grows. Thus given the current capabilities of computer processors this values goes to zero with a large enough data set and thus researchers, such as Rabiner [1989], have developed solutions. His solution involves scaling the parameters appropriately, such that they remain probabilities. The algorithm changes to:

1. Initialization:
   $\ddot{\alpha}_1(i) := \pi_i b_i(O_1)$
   $d_1 := \frac{1}{\sum_{i=1}^{L} \ddot{\alpha}_1(i)}$
   $\hat{\alpha}_1(i) := d_1 \ddot{\alpha}_1(i), \qquad i \in \{1, ..., L\}.$

2. Recursion for $t = 2, ..., T$:
   $\ddot{\alpha}_t(j) := \sum_i (\hat{\alpha}_{t-1}(i) a_{ij})\, b_j(O_t)$
   $d_t = \frac{1}{\sum_{i=1}^{L} \ddot{\alpha}_1(i)}$
   $\hat{\alpha}_t(i) := d_t \ddot{\alpha}_t(i), \qquad j \in \{1, ..., L\}.$

Thus d_t becomes the scaling by which we ensure that the $\alpha_t(i)$ does not go to zero. Note that d_t only depends on t and not i. This makes $\sum_{i=1}^{L} \hat{\alpha}_t(i)$ always equal to one and thus a new method is needed to find the total output probability. By induction, one can show that

$$\hat{\alpha}_t(i) = \left(\prod_{j=1}^{t} d_j \right) \alpha_t(i).$$

Using this modified forward algorithm one can then redefine the total output probability such that it does not go to zero,

$$\begin{aligned} 1 &= \sum_{i=1}^{N} \hat{\alpha}_T(i) = \sum_{i=1}^{N} \left(\prod_{j=1}^{T} d_j \right) \alpha_T(i) \\ &= \left(\prod_{j=1}^{T} d_j \right) \sum_{i=1}^{N} \alpha_T(i) = \left(\prod_{j=1}^{T} d_j \right) \Pr(\mathbf{O}|\phi). \end{aligned}$$

Manipulating the above result gives

$$D = \ln\left(\Pr(\mathbf{O}|\phi)\right) = -\sum_{j=1}^{T} \ln(d_j). \tag{5.1}$$

This is a more useful criterion for determining, given a large T, when to stop the EM algorithm. Notice how the left-hand side of (5.1) increases to zero as the sum on the left approaches zero as well. Thus we want the right-hand side to be as close as possible to zero. As was true with the forward variables, the backward variable also suffers from the same problem and scaling becomes appropriate with a large T. Consider the algorithm:

1. Initialization:
   $\ddot{\beta}_T(i) := 1$
   $\hat{\beta}_T(i) := d_T \ddot{\beta}_T(i), \qquad i \in \{1, ..., L\}.$

2. Recursion for $t = T-1, ..., 1$:
   $\ddot{\beta}_t(j) := \sum_i \left(\hat{\beta}_{t-1}(i) a_{ij} \right) b_i(O_{t+1})$
   $\hat{\beta}_t(i) := d_t \ddot{\beta}_t(i), \qquad j \in \{1, ..., L\}.$

Using this recursion one can show that $\hat{\beta}_t(i) = \prod_{i=t}^{T} d_i \beta_t(i)$. Given our new definitions of the forward and backward variables we can rewrite our $\gamma_t(j)$ and $\gamma_t(i,j)$ as:

$$\gamma_t(i) = \frac{\hat{\alpha}_t(i)\hat{\beta}_t(i)}{d_t}, \qquad\qquad \gamma_t(i,j) = \hat{\alpha}_t(i) a_{ij} b_j(o_{t+1}) \hat{\beta}_{t+1}(i).$$

These new definitions make it easier to implement the algorithm and help in the case of a large time horizon.

The proposed model has a subtlety that has yet to be discussed. In the case of zero claims there is no corresponding severity output. The gamma distribution is not defined at zero and thus one needs to be careful when implementing the EM algorithm. When estimating the forward and backward variables for example, one should be careful with the value for $b_i(O_t)$. Let f_i and g_i be the probability function of the Poisson and the density of the gamma for state i respectively. Then

$$b_i(O_t) = f_i(n_t)g_i(c_t),$$

if $n_t > 0$ and

$$b_i(O_t) = f_i(n_t),$$

otherwise. Note that there will be less severity observations to calculate the MLE estimates for the gamma severity distribution. Thus it should take longer for it to converge.

5.3 Simulations

In this section we will compare the three methods, HMMs, GLMs and HMM-GLM, to see which performs better under certain assumptions. All the parameters were estimated using the algorithms described in the previous sections. We will see how different values of T and L affect our estimates, first examining the case of one observation sequence and then expanding to multiple observation sequences.

5.3.1 One Observation Sequence

First we will assume that the data follow an HMM-GLM type model with two hidden states. The data was simulated according to the following scheme (simulation scheme 1):

$$(n_i|S_i = j) \sim Poisson(\lambda_{ij}), \qquad i \in \{1, \ldots, T\} \quad j \in \{1, \ldots, L\},$$
$$(c_i|S_i = j) \sim Gamma(\theta_{ij}, k), \qquad i \in \{1, \ldots, T\} \quad j \in \{1, \ldots, L\},$$

where

$$\ln(\lambda_{ij}) = x_{i1}u_{j1} + x_{i2}u_{j2} + x_{i3}u_{j3},$$
$$\ln(\theta_{ij}k) = x_{i1}w_{j1} + x_{i2}w_{j2} + x_{i3}w_{j3}.$$

Note that the same covariates were used for severity and claim counts in this simulation and no intercept term was included. The covariate values were drawn independently from a uniform (0,1) distribution. We used the following true parameter matrices to simulate our data:

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = \begin{bmatrix} .3 \\ .7 \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} .8 & .2 \\ .35 & .65 \end{bmatrix}$$
$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \end{bmatrix} = \begin{bmatrix} .5 & .25 & .75 \\ -.5 & 1.75 & 1.0 \end{bmatrix} \quad \mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} = \begin{bmatrix} .1 & .46 & .8 \\ -.6 & 1.2 & 2 \end{bmatrix},$$

and the shape parameter was set to $k = \frac{3}{7}$.

The first state represents a good driver while the second state represents a bad driver, this is shown by both sets of coefficients. State 2 produces higher average claims with higher severities. Also the initial probabilities $\boldsymbol{\pi}$ depict that it is more likely to start in the bad state as we are assuming that this is the policyholder's first year driving. The transition matrix conveys that a driver's skill is more likely to stay the same, though it is easier to transition from the bad to the good state than vice versa. For each value of T we randomly initialized our parameters 100 times, ran the EM algorithm for 2 hidden states and chose the set of parameters which produced the lowest total output probability. Below in Table 5.1 and 5.2 show the results for different values of T.

The EM algorithm performed well for $T = 1000$, 5000 for all parameters except for the initial probabilities as with one observation sequence there is only one initial value to derive estimates. This problem should be alleviated when we transition to multiple observation sequences. When $T = 500$ the algorithm had trouble converging to the coefficients for the severity as there were less observations, recall that when the count is 0 there is no severity

T	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{a}_{11}$	$\hat{a}_{12}$	$\hat{a}_{21}$	$\hat{a}_{22}$	$\hat{k}$	$\ln(\Pr(\mathbf{O}\|\phi))$
100	.01	.99	.81	.19	.18	.82	.51	-305.34
500	.05	.95	.76	.24	.23	.77	.48	-1666.85
1000	.03	.97	.75	.25	.36	.64	.46	-3367.56
5000	.01	.99	.73	.27	.39	.61	.42	-16833.34

Table 5.1: Estimated Probabilities for an HMM-GLM using Simulation Scheme 1

T	$\hat{u}_{11}$	$\hat{u}_{12}$	$\hat{u}_{13}$	$\hat{u}_{21}$	$\hat{u}_{22}$	$\hat{u}_{23}$	$\hat{w}_{11}$	$\hat{w}_{12}$	$\hat{w}_{13}$	$\hat{w}_{21}$	$\hat{w}_{22}$	$\hat{w}_{23}$
100	1.07	$-.23$	.34	-.81	1.87	1.10	.54	-1.98	1.65	-.15	.57	1.70
500	.52	.21	.73	-.45	1.62	1.04	.18	1.06	.08	-.61	1.43	1.68
1000	.58	.39	.53	-.63	1.75	1.11	.09	.47	.62	-.63	.96	2.19
5000	.57	.19	.68	-.5	1.72	1.00	.17	.39	.87	-.79	1.19	1.94

Table 5.2: Estimated Coefficients for an HMM-GLM using Simulation Scheme 1

observation. Therefore for the part of the model related to severity there will always be a number of observations less than or equal to T. When comparing the distances according to absolute value,

$$ad_T = |(w_{11} - \hat{w}_{11})| + |(w_{12} - \hat{w}_{12})| + ... + |(w_{23} - \hat{w}_{23})|,$$

one can clearly see that the estimates for $T = 500$ were worse, in fact $ad_{500} = 1.96$ which is considerably greater than $ad_{1000} = .65$ or $ad_{5000} = .48$. For $T = 100$ the algorithm had difficulties converging in general. Figure 5.1 shows one random initialization of the EM algorithm with $D = \ln(\Pr(\mathbf{O}|\phi))$ on the vertical axis and the number of iterations on the horizontal axis. The algorithm stopped when the relative logarithm of the total output probability had reached a level .001 or smaller. This occurred after 8 iterations.

Of course when running the algorithm for real data one will not know the number of hidden states. Therefore one will need a criterion to choose the number of hidden states. For this we use the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) defined as:

$$AIC = 2\ln(\Pr(\mathbf{O}|\phi)) - 2\iota, \qquad BIC = \ln(\Pr(\mathbf{O}|\phi)) - \frac{\ln(T)\iota}{2},$$

where ι represents the number of parameters to be estimated. One wants to pick the number of hidden states that maximizes either AIC or BIC. Tables 5.3 and 5.4 below show the AIC and BIC values for different number of hidden states, were the subscript denote the number of hidden states. Both statistics favor the model with two hidden states, which makes sense as the simulation included two hidden states.

T	AIC_2	AIC_3	AIC_4	AIC_5
100	-648.68	-664.04	-677.12	-685.70
500	-3371.70	-3391.54	-3405.26	-3420.54
1000	-6773.12	-6793.00	-6828.16	-6854.98

Table 5.3: AIC Statistic for HMM-GLM using Simulation Scheme 1

T	BIC_2	BIC_3	BIC_4	BIC_5
100	-349.09	-372.40	-379.18	-422.31
500	-1725.89	-1761.10	-1797.46	-1838.82
1000	-3433.18	-3472.57	-3524.50	-3577.18

Table 5.4: BIC Statistic for HMM-GLM using Simulation Scheme 1

Using the same simulation scheme and stopping criterion we will now fit a Poisson-gamma HMM to the data. The results are summarized in Table 5.5. As expected the model provided a worse fit, which can be seen by the column of the log likelihood. This is due to the fact that we simulated according to a HMM-GLM, therefore the Poisson-gamma HMM does not capture all the details of the data. Also, when estimating the shape parameter, the Poisson-gamma HMM estimates two different values when the data was simulated using one, though as T increases they both approach the true value of $\frac{3}{7}$.

Figure 5.2 illustrates initialization of the EM algorithm. On average it took more iterations to converge than the HMM-GLM, but the code executed a lot faster as the HMM has fewer parameters. Like before a comparison of the AIC and BIC statistics is given in Tables 5.6 and 5.7. Both statistics for each value of T favor the fitted model with 2 hidden states.

Next we fitted a standard GLM to the data using the R function glm.fit(). The data were fitted to two separate GLMs, Poisson and gamma, both with log link functions as

T	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{a}_{11}$	$\hat{a}_{12}$	$\hat{a}_{21}$	$\hat{a}_{22}$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{k}_1$	$\hat{k}_2$	$\ln(\Pr(\mathbf{O}\|\phi))$
100	.97	.03	.81	.19	.72	.28	1.67	6.30	5.50	8.70	.36	.50	-328.02
500	.01	.99	.41	.59	.18	.82	5.30	2.00	14.19	4.55	.49	.41	-1764.99
1000	.99	.01	.86	.14	.81	.19	2.18	6.64	5.17	19.22	.43	.39	-3538.99
5000	.01	.99	.85	.15	.76	.24	2.05	5.98	4.84	17.38	.43	.39	-17564.45

Table 5.5: Estimated Parameters for a Poisson-Gamma HMM using Simulation Scheme 1

T	AIC_2	AIC_3	AIC_4	AIC_5
100	−680.04	−688.70	-688.68	-707.64
500	−3553.98	−3560.12	-3565.80	-3586.64
1000	−7101.98	−7113.94	-7132.78	-7148.72

Table 5.6: AIC Statistic for Poisson-Gamma HMM using Simulation Scheme 1

in the HMM-GLM example. Table 5.8 gives the results for different values of T. The estimates always lie in between the two true values for both hidden states. This is because the simulation is switching between the two as the hidden Markov chain progresses.

Now that all model types have been fitted an analysis to determine how they perform in predicting future values is carried out. We are interested in forecasting total claims and therefore will consider the product of number of claims and average severity. To compare the models we will consider the root mean squared error (RMSE) of future values,

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)}{n}},$$

where $\hat{y}_t$ and y_t are the predicted and actual values, respectively. Using the estimates after 5000 observations, as these appear to be the most accurate, and simulating another 1000 observations with respect to the final hidden state produced the following RMSEs described in Table 5.9. The HMM performed the worst while the difference between the HMM-GLM and GLM seems rather negligible. This is not surprising as the further into the future you forecast the advantages of the HMM-GLM over the GLM should get averaged out with the two performing relatively the same. If instead we simulated the next value in the HMM-GLM a 1000 times the HMM-GLM should outperform the GLM. Doing so we arrive at the results in Table 5.10. This time the HMM-GLM outperformed the GLM. Thanks to the estimates

T	BIC_2	BIC_3	BIC_4	BIC_5
100	-355.65	-371.70	-386.02	-412.44
500	-1802.28	-1824.31	-1850.33	-1888.15
1000	-3580.44	-3608.50	-3644.91	-3684.78

Table 5.7: BIC Statistic for Poisson-Gamma HMM using Simulation Scheme 1

T	$\hat{u}_1$	$\hat{u}_2$	$\hat{u}_3$	$\hat{w}_1$	$\hat{w}_2$	$\hat{w}_3$
100	$-.17$	1.16	.89	-.05	-.22	1.82
500	$-.56$	1.06	.94	-.45	1.32	1.22
1000	$-.09$	1.11	.88	-.37	.73	1.56
5000	.03	.98	.84	-.33	.83	1.41

Table 5.8: Estimated Coefficients for a GLM using Simulation Scheme 1

for the probability of the hidden state at time T, $\gamma_T(i)$, the HMM-GLM was able to exploit its hidden Markov chain. If the values of $\gamma_T(i)$ were more extreme like in the example with 1000 observations the advantage could be greater, as is shown in Table 5.11. Let the relative gain be defined as,

$$rg_T = \frac{RMSE_{GLM} - RMSE_{HMM-GLM}}{RMSE_{GLM}}.$$

Thus $rg_{1000} = .10$ and $rg_{5000} = .02$, rg_{1000} being larger indicates more of an impact. If $\gamma_T(i)$ placed all the weight on the correct hidden state this advantage would be at its max. Of course these estimates need to be accurate for the advantage to be realized. This short term benefit bodes well for a whole portfolio of observation sequences as there will be more sequences to forecast.

Model	RMSE
GLM	31.12
HMM	32.36
HMM-GLM	31.10

Table 5.9: RMSE for Different Models with Simulation Scheme 1

Given the fact that the current models typically rely on GLMs, a simulation of a GLM is needed for comparison. Note that this is equivalent to an HMM-GLM with just one hidden

Model	RMSE
GLM	74.16
HMM-GLM	72.71

Table 5.10: RMSE-Next Value, for Different Models with Simulation Scheme 1

Model	RMSE
GLM	12.55
HMM-GLM	11.26

Table 5.11: RMSE-Next Value, for Different Models with Simulation Scheme 1 and the Estimates of $T = 1000$

state. The data was simulated according to the following scheme (simulation scheme 2):

$$(n_i) \sim Poisson(\lambda_i), \qquad i \in \{1, \ldots, T\},$$
$$(c_i) \sim Gamma(\theta_i, k), \qquad i \in \{1, \ldots, T\},$$

where

$$\ln(\lambda_i) = x_{i1}u_1 + x_{i2}u_2 + x_{i3}u_3,$$
$$\ln(\theta_i k) = x_{i1}w_1 + x_{i2}w_2 + x_{i3}w_3.$$

The true parameters for the coefficients and the shape parameter are as follows,

$$\mathbf{U} = \begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix} = \begin{bmatrix} .7 & .68 & 1.3 \end{bmatrix} \qquad \mathbf{W} = \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} = \begin{bmatrix} .45 & 2.2 & 1.69 \end{bmatrix},$$

where the shape parameter $k = 7/9$. The covariates were again drawn from a uniform (0,1) distribution, independently. As opposed to the previous simulation the drivers no longer switch from good to bad and vice versa. Therefore the true values represent an all encompassing state.

For this scheme it is unclear what the correct number of hidden states is, so let us first compare the AIC and BIC statistics to determine an appropriate value. Tables 5.12 and 5.13 show the different AIC and BIC statistics for a varying T for a HMM-GLM. For all

values of T the statistics seem to agree on two hidden states. This is not surprising as the simulation had one hidden state thus making two the closest option. Therefore we will now fit an HMM-GLM with two hidden states for varying values of T, as before.

T	AIC_2	AIC_3	AIC_4	AIC_5
100	-995.54	-996.18	-1004.54	-1002.14
500	-5075.52	-5101.18	-5124.38	-5136.56
1000	-10180.14	-10207.08	-10232.68	-10261.02

Table 5.12: AIC Statistic for a HMM-GLM using Simulation Scheme 2

T	BIC_2	BIC_3	BIC_4	BIC_5
100	-522.52	-538.47	-560.89	-580.53
500	-2577.80	-2615.92	-2657.02	-2696.83
1000	-5136.69	-5179.61	-5226.76	-5280.20

Table 5.13: BIC Statistic for a HMM-GLM using Simulation Scheme 2

Using the same stopping criterion as before Tables 5.14 and 5.15 contain the values for the fitted parameters. In each case one state seems to be closer to the true parameters and the Markov chain favors that state. For example when $T = 5000$ the absolute distances for state one and two for the Poisson GLM are $ad_{5000} = 2.36$ and $ad_{5000} = .03$, respectively. The gamma portion favors the same state as well, the distance for state one and two are $ad_{5000} = .90$ and $ad_{5000} = .20$, respectively. Then looking at the probability of transitioning to state two is much higher from each state, .87 and .99. This same phenomenon can be found for each value of T. Also the total distance for each the Poisson and gamma coefficients is decreasing as T increases, suggesting that they are converging to the correct values. In addition the shape parameter seems to be converging nicely. After 1000 observations the estimate is within .1 of the true value. The initial state probabilities are hard to interpret as there are no initial values to compare them to.

We would like to compare the three models therefore a Poisson-gamma HMM will be fitted to the data next. Tables 5.16 and 5.17 give the AIC and BIC statistics for the model. Comparing both indicates that an Poisson-gamma HMM with two hidden states should be

T	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{a}_{11}$	$\hat{a}_{12}$	$\hat{a}_{21}$	$\hat{a}_{22}$	$\hat{k}$	$\ln(\Pr(\mathbf{O}\|\phi))$
100	.01	.99	.20	.80	.26	.74	1.17	-478.77
500	.01	.99	.80	.20	.71	.29	.91	-2518.76
1000	.03	.97	.16	.84	.21	.79	.76	-5071.07
5000	.01	.99	.12	.87	.01	.99	.78	-25403.20

Table 5.14: Estimated Probabilities for a HMM-GLM using Simulation Scheme 2

T	$\hat{u}_{11}$	$\hat{u}_{12}$	$\hat{u}_{13}$	$\hat{u}_{21}$	$\hat{u}_{22}$	$\hat{u}_{23}$	$\hat{w}_{11}$	$\hat{w}_{12}$	$\hat{w}_{13}$	$\hat{w}_{21}$	$\hat{w}_{22}$	$\hat{w}_{23}$
100	$-.10$	.84	2.04	.98	.38	1.22	.41	-.54	-1.02	.08	2.41	2.10
500	.81	.60	1.24	.41	.79	1.33	.41	2.40	1.94	.66	1.14	.90
1000	.53	$-.55$	1.42	.74	.76	1.25	.66	1.39	1.25	.52	2.27	1.68
5000	.18	$-.93$	1.54	.72	.67	1.3	.58	1.55	1.57	.54	2.10	1.68

Table 5.15: Estimated Coefficients for a HMM-GLM using Simulation Scheme 2

fitted. The models estimates are provided in Table 5.18. In the cases when $T = 500$, 1000, or 5000 the HMM seems to slightly favor the state which more correctly estimates the shape parameter, giving more weight to transitioning in the more correct state. When $T = 100$ the results were inconclusive in that regard. Also for all values of T the model did an adequate job for $\hat{k}$. As with the HMM-GLM it is difficult to compare the probabilities as the simulation did not include any.

T	AIC_2	AIC_3	AIC_4	AIC_5
100	-1110.16	-1119.48	-1132.54	-1147.84
500	-5466.42	-5533.08	-5555.62	-5579.60
1000	-11024.08	-11029.74	-11052.40	-11080.84

Table 5.16: AIC Statistic for a Poisson-Gamma HMM using Simulation Scheme 2

Proceeding to a generalized linear model, note that Poisson and gamma GLM were fitted to the simulation. The results are contained in Table 5.19. As expected the model fits the true parameters very well as it was the model used to simulate the data.

Next we want to compare the predictive power of each model given the simulation scheme. Using the estimates for $T = 5000$ and simulating another 1000 values produced Table 5.20. The HMM this time performs a lot worse than the other two models. This is probably due to

T	BIC_2	BIC_3	BIC_4	BIC_5
100	-579.83	-600.12	-624.89	-653.38
500	-2800.25	-2831.87	-2872.64	-2918.35
1000	-5558.66	-5590.94	-5636.62	-5690.11

Table 5.17: BIC statistic for a Poisson-Gamma HMM using Simulation Scheme 2

T	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{a}_{11}$	$\hat{a}_{12}$	$\hat{a}_{21}$	$\hat{a}_{22}$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{k}_1$	$\hat{k}_2$	$\ln(\Pr(\mathbf{O}\|\phi))$
100	.99	.01	.51	.49	.30	.70	6.85	2.36	38.44	5.41	.57	.80	-536.08
500	.01	.99	.25	.75	.29	.71	6.81	2.96	43.02	8.70	.66	.73	-2741.21
1000	.01	.99	.40	.60	.44	.56	6.49	2.59	28.02	7.53	.68	.73	-5493.04
5000	.99	.01	.47	.53	.36	.64	6.51	2.75	32.12	8.38	.64	.71	-27634.70

Table 5.18: Estimated parameters for a Poisson-Gamma HMM using Simulation Scheme 2

the fact that this time it is a cruder approximation to the data. It does not include covariates and the data was not simulated according to a hidden Markov chain. The GLM and HMM-GLM provide relatively the same performance, with the GLM performing slightly better. This is suggesting that given a GLM scheme for the simulation the models have relatively equal predictive capabilities. Note that given the simulation one cannot simulate the next observation, given the ending hidden state, as the simulation did not rely on a hidden Markov chain.

T	$\hat{u}_1$	$\hat{u}_2$	$\hat{u}_3$	$\hat{w}_1$	$\hat{w}_2$	$\hat{w}_3$
100	.76	.49	1.40	.04	2.22	1.95
500	.72	.64	1.26	.42	2.26	1.80
1000	.70	.72	1.28	.52	2.16	1.62
5000	.72	.67	1.30	.54	2.10	1.68

Table 5.19: Estimated Coefficients for a GLM using Simulation Scheme 2

5.3.2 Multiple Observation Sequences

Now consider simulation Scheme 1 but with multiple observation sequences. Recall that the number of sequences is represented by M. This more realistically depicts an insurance portfolio with many policyholders. In the simulation described below each insuree had

Model	RMSE
GLM	77.30634
HMM-GLM	77.30656
HMM	93.03

Table 5.20: RMSE for Different Models using Simulation Scheme 2

equivalent time horizons, but note that this assumption can be relaxed and all algorithms still hold. Fitting the proposed HMM-GLM produces Tables 5.21 and 5.22. The initial state probabilities $\boldsymbol{\pi}$ converge nicely at $M \geq 500$. Thus given a new driver one can more accurately predict their initial state than before. Note that R does not need to invert large matrices to find the estimates for the HMM-GLM fit as before with a single sequence. When using multiple sequences the estimates rely on the sums of matrices thus reducing the size of the matrices needed.

M	T	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{a}_{11}$	$\hat{a}_{12}$	$\hat{a}_{21}$	$\hat{a}_{22}$	$\hat{k}$	$\ln(\Pr(\mathbf{O}\|\phi))$
100	5	.69	.31	.75	.25	.45	.55	.39	-1693.21
100	10	.21	.79	.72	.18	.47	.53	.44	-3398.38
500	5	.31	.69	.63	.37	.42	.58	.39	-8514.41
500	10	.35	.65	.75	.25	.48	.52	.44	-16867.95
1000	5	.34	.66	.82	.18	.38	.62	.47	-17077.13
1000	10	.32	.68	.78	.22	.37	.63	.44	-33967.57
5000	5	.31	.69	.77	.23	.36	.64	.42	-86179.40
5000	10	.30	.70	.79	.21	.36	.64	.43	-169874.60
10000	5	.31	.69	.81	.19	.35	.65	.42	-172455.80
10000	10	.29	.71	.79	.21	.37	.63	.43	-340206.50

Table 5.21: Estimated Probabilities for an HMM-GLM using Simulation Scheme 1 with Multiple Observation Sequences

Following the same theme as before, a Poisson-gamma HMM will now be fitted to the data. The results are contained in Table 5.23. The estimates seem to converge and do a better job recovering the initial state probabilities than before, but the model still does a poorer job to fit the data than a HMM-GLM. This can be seen by comparing the last columns of Tables 5.23 and 5.21 for similar M's and T's. The Poisson-gamma HMM cannot capture all the variability in the data as it was simulated according to an HMM-GLM. It

M	T	$\hat{u}_{11}$	$\hat{u}_{12}$	$\hat{u}_{13}$	$\hat{u}_{21}$	$\hat{u}_{22}$	$\hat{u}_{23}$	$\hat{w}_{11}$	$\hat{w}_{12}$	$\hat{w}_{13}$	$\hat{w}_{21}$	$\hat{w}_{22}$	$\hat{w}_{23}$
100	5	.22	.44	1.06	-.77	2.12	.67	.48	.06	1.10	-1.40	1.88	1.69
100	10	.46	.12	.89	-.45	1.77	.91	-.07	.19	1.58	-.25	1.48	1.13
500	5	.44	.11	.86	-.47	1.75	.91	.35	-.07	.98	-.64	1.08	.98
500	10	.47	.32	.75	-.52	1.81	.96	-.06	.54	.84	-.37	1.15	.84
1000	5	.37	.31	.82	-.52	1.79	.99	.03	.30	1.15	-.54	1.22	2.01
1000	10	.48	.28	.74	-.47	1.78	.93	.07	.52	.73	-.47	1.12	1.97
5000	5	.53	.22	.75	-.53	1.74	1.02	.14	.37	.78	-.59	1.25	1.99
5000	10	.47	.25	.76	-.50	1.74	1.00	.04	.49	.82	-.54	1.17	2.00
10000	5	.51	.26	.74	-.52	1.75	1.02	.13	.36	.82	-.62	1.19	1.94
10000	10	.48	.25	.76	-.49	1.74	1.00	.07	.47	.82	-.58	1.21	1.99

Table 5.22: Estimated Coefficients for an HMM-GLM using Simulation Scheme 1 with Multiple Observation Sequences

does seem to converge to certain values which appear difficult to determine, though the shape parameters converge nicely. $\hat{k}_1$ and $\hat{k}_2$ approach the true value of $k = \frac{3}{7}$. As was true with the HMM-GLM, the Poisson-gamma HMM estimates also require matrices of smaller size. Thus the EM algorithm does not need to invert a large matrix but many smaller ones.

M	T	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{a}_{11}$	$\hat{a}_{12}$	$\hat{a}_{21}$	$\hat{a}_{22}$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{k}_1$	$\hat{k}_2$	$\ln(\Pr(\mathbf{O}\|\phi))$
100	5	.24	.76	.26	.74	.29	.71	5.41	1.91	18.21	4.83	.34	.43	-1798.45
100	10	.29	.71	.34	.66	.25	.75	5.03	1.89	14.06	4.46	.43	.44	-3589.54
500	5	.21	.79	.21	.79	.15	.85	6.39	2.15	20.42	5.60	.40	.41	-8980.15
500	10	.22	.78	.20	.80	.12	.88	6.77	2.19	19.24	5.39	.42	.42	-17809.48
1000	5	.26	.74	.21	.79	.16	.84	6.23	2.09	21.25	5.67	.40	.40	-18050.16
1000	10	.24	.76	.22	.78	.13	.87	6.52	2.15	19.56	5.37	.41	.42	-35777.48
5000	5	.25	.75	.22	.78	.15	.85	6.45	2.12	21.14	5.39	.40	.42	-91248.68
5000	10	.26	.74	.24	.76	.13	.87	6.43	2.13	19.79	5.41	.41	.42	-179007.70
10000	5	.25	.75	.22	.78	.14	.86	6.56	2.14	21.61	5.46	.40	.42	-182745.60
10000	10	.26	.74	.23	.77	.13	.87	6.48	2.14	19.87	5.38	.41	.42	-358479.00

Table 5.23: Estimated Parameters for a Poisson-Gamma HMM using Simulation Scheme 1 with Multiple Observation Sequences

Lastly a GLM model will be fitted. Table 5.24 provide the results. As before the estimates lie in between the values for both hidden states. Using the three fitted models at values of $M = 10000$ and $t = 10$, Table 5.25 gives a comparison of the RMSE's based on simulating the claims in the next time period for each policyholder. The HMM-GLM outperforms the

M	T	$\hat{u}_1$	$\hat{u}_2$	$\hat{u}_3$	$\hat{w}_1$	$\hat{w}_2$	$\hat{w}_3$
100	5	$-.22$	1.22	.90	-.37	.98	1.43
100	10	$-.01$	.99	.91	-.17	.91	1.33
500	5	$-.13$	1.15	.89	-.35	.74	1.67
500	10	$-.04$	1.09	.87	-.30	.91	1.40
1000	5	$-.13$	1.14	.92	-.35	.87	1.64
1000	10	$-.03$	1.09	.86	-.29	.90	1.45
5000	5	$-.11$	1.13	.93	-.38	.97	1.57
5000	10	$-.05$	1.06	.90	-.33	.91	1.49
10000	5	$-.10$	1.14	.92	-.39	.93	1.60
10000	10	$-.04$	1.05	.91	-.34	.92	1.48

Table 5.24: Estimated Coefficients for a GLM using Simulation Scheme 1 with Multiple Sequences

other models. Using the values of $\gamma_T^{(m)}(i)$ the model can leverage these values to accurately predict the next emission. Another value of interest would be the estimates of the initial time period. Since the simulation included multiple observations the estimates for the initial state probabilities are more accurate. Simulating 1000 initial values and comparing the RMSE of the three models shows that HMM-GLM again outperforms the other two. These values are contained in Table 5.26. This time the gains are higher.

Model	RMSE
GLM	38.57
HMM-GLM	37.26
HMM	40.15

Table 5.25: RMSE-NEXT Value, for Different Models using Simulation Scheme 1 with Multiple Observation Sequences

Model	RMSE
GLM	62.14
HMM-GLM	59.10
HMM	65.75

Table 5.26: RMSE-Initial Value, for Different Models using Simulation Scheme 1 with Multiple Observation Sequences

5.4 Conclusion

Once accounting for the scaling issue all the simulations converged. Note that normally, depending on the time period considered, for insurance data one would not need to scale. For instance if a company was considering a model with yearly time periods it is unlikely that they would have data for more than 10 years for each policyholder, but if instead they considered seasonal data scaling might be needed. This need should be alleviated as computers become more efficient. The simulations for single observation sequences do a poorer job of representing reality, therefore considering a large set of multiple observation sequences with short time horizons more accurately depicts an insurance portfolio.

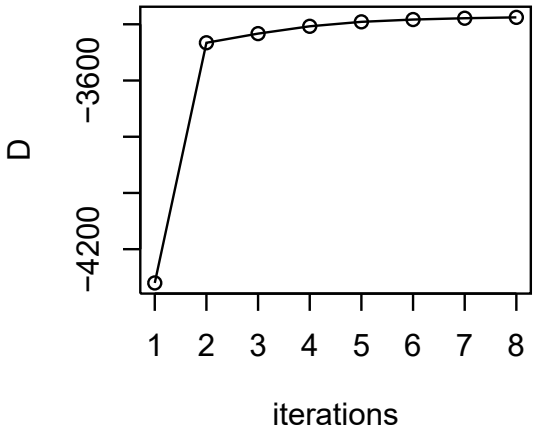

Figure 5.1: Convergence of HMM-GLM

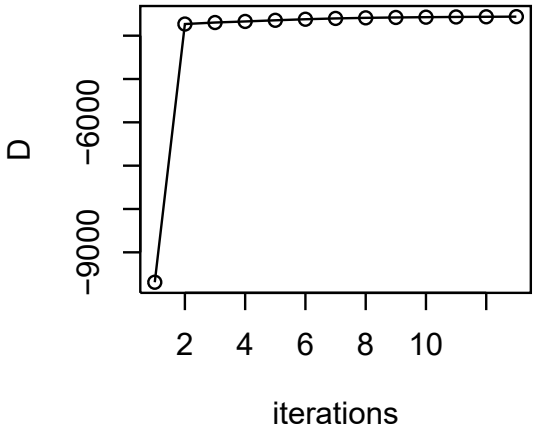

Figure 5.2: Convergence of Poisson-Gamma HMM

Conclusion

GLMs are often used to model claims and in most models there is a covariate created to capture a driver's skill based on their past claims. The HMM aspect of both proposed models views skill as something unseen which better depicts reality and considers that there are multiple distributions that the claims are generated from. In other words, good and bad drivers draw from different distributions. The Markov chain allows drivers to switch from bad to good dynamically. The proposed models also have the benefit of creating dependence between count and severity. Often in practice these two quantities are estimated independently and then multiplied together. Some researchers find this controversial as there could be dependence at play (see Garrido et al. [2016]). Another common problem in auto insurance data is there are too many zero claim counts to model effectively with the Poisson distribution. HMMs help alleviate this by spreading the zeros across multiple states. Thus the HMM-GLM could provide a better fit to reality.

Unfortunately we were unable to procure real data as insurance companies consider this material very sensitive. Given the simulations in Chapter 5 we think that the HMM-GLM could provide more accurate forecasts. In addition to the described models, one might want to consider HMM type models with longer time period dependencies. Also one could consider other emissions, such as Tweedie or Erlang. Another extension to GLMs are generalized linear mixed models which one could consider as emissions instead of GLMs. This concept has been researched before in Altman [2007]. Dr. Altman develops a mixed hidden Markov model, a hybrid HMM and GLMM, to model lesion counts in patients with multiple sclerosis. In conclusion we believe that HMM-GLMs can provide a more realistic interpretation of reality and thus more accurately depict insurance portfolios.

Bibliography

Rachel MacKay Altman. Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102(477):201–210, 2007.

Necdet Batir. On some properties of digamma and polygamma functions. *Journal of Mathematical Analysis and Applications*, 328(1):452–465, 2007.

Piet De Jong, Gillian Z Heller, et al. *Generalized linear models for insurance data*, volume 136. Cambridge University Press Cambridge, 2008.

Jieyu Fan. *On Markov and hidden Markov models with application to trajectories*. PhD thesis, University of Pittsburgh, 2015.

Gernot A Fink. *Markov models for pattern recognition: from theory to applications*. Springer Science & Business Media, 2014.

Jose Garrido, Christian Genest, and Juliana Schulz. Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, 70: 205–215, 2016.

Peter J Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, pages 799–821, 1973.

Silvie Kafková, Lenka Křivánková, et al. Generalized linear models in vehicle insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 62(2):383–388, 2014.

Xiaolin Li, Marc Parizeau, and Réjean Plamondon. Training hidden Markov models with multiple observations-a combinatorial method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):371–377, 2000.

Yi Lu and Leilei Zeng. A nonhomogeneous Poisson hidden Markov model for claim counts. *Astin Bulletin*, 42(01):181–202, 2012.

Peter McCullagh and John A Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

Nasser Mohammadiha, W Bastiaan Kleijn, and Arne Leijon. Gamma hidden Markov model as a probabilistic nonnegative matrix factorization. In *21st European Signal Processing Conference (EUSIPCO 2013)*, pages 1–5. IEEE, 2013.

Roberta Paroli, Luigi Spezia, and Giovanna Luisa Redaelli. Hidden Markov models for time series of overdispersed insurances counts. In *XXXI International ASTIN Colloquium*, pages 461–474. Proceedings of the XXXI International ASTIN Colloquium, 2000.

Oscar Alberto Xacur Quijano and José Garrido. Generalised linear models for aggregate claims: to Tweedie or not? *European Actuarial Journal*, 5(1):181–202, 2015.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Pascal Sebah and Xavier Gourdon. Introduction to the gamma function. *numbers. computation. free. fr/Constants/constants. html*, 2002.

Dawei Shen. Some mathematics for hmm. *Massachusetts Institute of Technology*, 2008.

Karline Soetaert. *rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations*, 2009. R package 1.6.

Karline Soetaert and Peter M.J. Herman. *A Practical Guide to Ecological Modelling. Using R as a Simulation Platform*. Springer, 2009. ISBN 978-1-4020-8623-6.

CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

Hao Zhang, Weidong Zhang, Ahmet Palazoglu, and Wei Sun. Prediction of ozone levels using a hidden Markov model (HMM) with gamma distribution. *Atmospheric environment*, 62: 64–73, 2012.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.