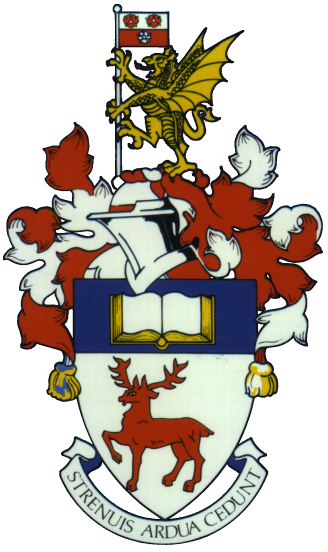
UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES

Mathematical Sciences



Forecasting User Roles In Online Communities

by

Edwin Tye

Thesis for the degree of Doctor of Philosophy

April 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES
Mathematical Sciences

Doctor of Philosophy

FORECASTING USER ROLES IN ONLINE COMMUNITIES

by Edwin Tye

There is a growing interest in setting up and modeling of online social networks, as there are major incentives (popularity, monetary) for owners and managers to understand their own communities. This is especially true for communities that were set up by businesses because of the time and money invested into building and maintaining such online platforms. We take the approach that these online communities operate similarly to an offline environment such that members of these social networks can be classified into several (social) roles, each asserting a different impact on the community.

This project focuses on forecasting the number of users in each of the roles of an online community. The forecasting model is split into two different parts. Part I models the movement of existing users between the roles, which was formulated as a linear difference equation upon time discretization. It is treated as an optimization problem where the objective function can take either the form of a least squares or a non–linear iterative update based on the difference equation, both with box and linear constraints. Part II predicts the number of new users joining and currently inactive users returning to the community. It is examined from a statistical point of view, where we postulate that the number of new users joining is akin to arrivals in a queue. In order to find the driving factor behind the numbers of new user joining, a series of models are explored using different independent variables. Models are built and tested for the two problems separately at first, then later combined to produce forecasts and their performances were accessed.

Contents

Nomenclature **xv**

Declaration of Authorship **xvii**

Acknowledgements **xix**

1 Introduction **1**
1.1 Motivation 1
1.2 Problem Statement 2
1.3 Description of Tasks 3
1.4 Outline of the Thesis 4

2 Background **7**
2.1 Introduction 7
2.2 Data Structure 7
2.3 Data Overview 10
2.4 Health of an online community 12
2.5 Roles in Online Communities 14
2.5.1 Clustering Methods 15
2.5.2 Usage in Role Analysis 19
2.6 Modeling of Online Communities 21

3 Main Methodology **23**
3.1 Compartment Model Introduction 23
3.2 Mathematical Formulation 24
3.2.1 Expected Transition 26
3.2.2 Last Transition 26
3.2.3 Contribution from External Sources 26
3.2.4 Comparing Formulation of Inactive Users 28
3.3 Estimation of Transition Matrix 29
3.3.1 Single Step Update 31
3.3.2 Iterative Update 33
3.3.3 Comparison of Update 34
3.4 Combination of Transition Matrices 37
3.4.1 Weighted Mixture of Matrix 37
3.4.2 Exponential Penalty 38

3.5 Stochastic Transition Matrix . . . 42
3.5.1 Parametric . . . 43
3.5.1.1 Truncated Multivariate Normal . . . 43
3.5.1.2 Truncated Univariate Normal With Copula . . . 45
3.5.1.3 Binomial Formulation . . . 46
3.5.2 Mixture of Historical Transition Matrices . . . 47
3.6 Results . . . 48
3.6.1 Deterministic Models . . . 49
3.6.2 Stochastic Models . . . 54

4 Auxiliary Problem **57**
4.1 Introduction . . . 57
4.2 Regression of Count Data . . . 58
4.3 Univariate Poisson Introduction . . . 60
4.4 Estimation Procedures . . . 62
4.4.1 Maximum Likelihood Estimate (MLE) . . . 63
4.4.2 Maximum–a–posteriori (MAP) . . . 65
4.4.3 Regularization Methods . . . 66
4.4.4 Monte Carlo Markov Chain (MCMC) . . . 68
4.4.5 MCMC Estimation For Poisson Regression . . . 70
4.5 Overdispersed Poisson . . . 73
4.5.1 Quasi–Poisson and Negative Binomial . . . 74
4.5.2 Poisson–Lognormal (PLN) . . . 76
4.5.3 Poisson–Lognormal With Autoregressive (PLNAR) . . . 78
4.5.4 Poisson–Multivariate Lognormal (PMLN) . . . 80
4.6 Results . . . 82
4.6.1 Quality of Fitted Models . . . 83
4.6.2 Out–of–Sample Performance . . . 86
4.6.2.1 Summary . . . 88

5 Main Results **91**
5.1 Introduction . . . 91
5.2 Results . . . 92
5.3 Discussion . . . 98

6 Concluding Remarks **109**
6.1 Summary . . . 109
6.2 Difficulties and Future Research . . . 110
6.2.1 Compartment Model . . . 110
6.2.2 Prediction of External Factors . . . 112

Appendix A **115**
A.1 Data summary for some forums . . . 115

Appendix B **121**

B.1 Estimated probability matrix under different constraints 121
B.2 Comparison of Formulations 122

Appendix C **127**
C.1 Comparison Between Different Proposal For Poisson Regression . . 127
C.2 Conditional Mean and Variance of Overdispersed Poisson 128
C.3 Poisson–Lognormal Sampling Information 129
C.3.1 Gradient and Hessian 129
C.3.2 Sampling Variance Parameter Under Uniform Prior 130
C.4 AR(1) Conditionals 131
C.4.1 Dispersions 131
C.4.2 Sampling Variance and Autoregressive coefficients 133

Appendix D **135**
D.1 Graphs of Combined Result 135

References **139**

List of Figures

2.1 A reply tree structure where each box corresponds to a message, the number and letters indicating the arrival order and the users who posted the messages respectively 9
2.2 Time series of the number of messages per week for the top 5 forums in terms of total messages 12

3.1 An example of a compartment model with 4 different user roles in a system 25
3.2 Scatter plot between the MSE obtained under the two different formulations over a 90 week period for forum 353. 30
3.3 Forecast using $\mathbf{P}_{\dagger}$ estimated using the linear single step update and the non–linear iterative update under the **I** formulation. Forecasted for a 99 weeks with initial value as the first observation and using the observed number of users at each time period 36
3.4 An example of two minima for the penalized transition matrix obtained in forum 264 under **I** at week $N = 94$, evaluated over 1000 equally spaced point over the interval $[0, 1]$ 40
3.5 The $\hat{\alpha}$ value obtained between **S1** and **S3** under the two different formulations. Over 90 consecutive weeks for forum 353 using the observed y_{Join} and y_{Return} and $t_0 = N - 10$ 41
3.6 The $\hat{\alpha}$ value obtained between **I** and **W** under the two different scenarios. Over 90 consecutive weeks for forum 353 using the observed y_{Join} and y_{Return} and $t_0 = N - 10$ 41
3.7 MSE of different methods across time for forum 264. 50
3.8 The number of users joining and returning for all the compartments apart from the inactive compartment. The inactive compartment shows the number of inactive users and the migration rate going out through time. 53

4.1 Observed (blue) and predictions (red) by the fitted model of the first 19 observations, using both the lagged and time dependent covariates 70
4.2 Example regression for a lag of 2 using lagged exogenous regressors 82
4.3 Time series plot for the out–of–sample MSE for both y_{Join} and y_{Return} over the 80 weeks for forum 353. All three models are under a Laplace prior. 87

4.4 Density plot for the out–of–sample MSE for both y_{Join} and y_{Return} over the 80 weeks for forum 353. All three models are under a Laplace prior. 88

5.1 The coverage against MSE of different methods and formulations over all 90 forecasts for forum 353 95
5.2 The coverage against MSE of different methods and formulations over all 90 forecasts for forum 256 96
5.3 The coverage against MSE of different methods and formulations over all 90 forecasts for forum 264 97
5.4 The MSE over time between **SB**, **SE** and **SP** for 3 forums for both formulations 100
5.5 The coverage over time between **SB**, **SE** and **SP** for 3 forums for both formulations 101
5.6 The fraction of time a method achieved the lowest MSE in forum 353. Each formulation sum to 100. 102
5.7 The fraction of time a method achieved the lowest MSE in forum 256. Each formulation sum to 100. 102
5.8 The fraction of time a method achieved the lowest MSE in forum 264. Each formulation sum to 100. 103
5.9 Kernel density plot of the raw error for each of the roles in forum 353 under the **I** formulation for 3 methods over all time 103
5.10 Kernel density plot of the raw error for each of the roles in forum 256 under the **I** formulation for 3 methods over all time 104
5.11 Kernel density plot of the raw error for each of the roles in forum 264 under the **I** formulation for 3 methods over all time 104
5.12 Kernel density plot of the raw error for each of the roles in forum 353 under the **W** formulation for 3 methods over all time 105
5.13 Kernel density plot of the raw error for each of the roles in forum 256 under the **W** formulation for 3 methods over all time 105
5.14 Kernel density plot of the raw error for each of the roles in forum 264 under the **W** formulation for 3 methods over all time 106
5.15 Kernel density plot of the raw error between the predicted and the actual active users at week 1 (Left panels) and week 10 (Right panels) of the forecast over 3 forums, two methods **SE** and **SP** under both formulations are shown 107

A.1 The number of users in each of the compartments through time for forum 353 115
A.2 The number of users in each of the compartments through time for forum 264 116
A.3 The number of users in each of the compartments through time for forum 256 116
A.4 The number of users in each of the compartments through time for forum 50 117

A.5 Network graph for forum 264 with data collected under a 13 week time window ending at time N=120. The bigger the node, the higher the betweenness centrality the node has and an edge with higher weight is darker 117
A.6 Network graph for forum 353 with data collected under a 13 week time window ending at time N=120. The bigger the node, the higher the betweenness centrality the node has and an edge with higher weight is darker 118

B.1 Autocorrelation plot of the first 20 lags for the transition $\mathbf{P}_{0,0}$, the number of user remains in the inactive compartment. 123
B.2 MSE over a set of α values using the inactive compartment for a 10 step in–sample tuning using 100 observations for forum 353 123
B.3 MSE over a set of α values without the inactive compartment for a 10 step in–sample tuning using 80 observations for forum 264 124
B.4 Forecast using $\mathbf{P}$ estimated under absolute loss and square loss using the non–linear iterative formulation. The 99 steps forecasts were generated using the first observation as the initial value and the observed number of users at each time period 125
B.5 Difference in the squared error between the forecast using $\mathbf{P}_{\dagger}$ estimated using the linear single step update and the non–linear iterative update under $\mathbf{I}$ formulation. A positive value indicates that the linear version has more squared error (against the observation) as compared to the non–linear version. Labels at the top right hand corner of each subplot shows the number of times the linear version is positive. 126

C.1 Autocorrelation plot over 50 lags for σ^2 under Normal prior 127
C.2 Autocorrelation plot over 50 lags for λ under Laplace prior 128
C.3 Difference in autocorrelation between the two different sampling scheme for the Multivariate Poisson Log–normal 128

D.1 The MSE over time between the two formulation for all three forums. Only the non–linear estimation of $\mathbf{SP}$ is demonstrated. . . . 135
D.2 MSE under the two formulation of $\mathbf{W}$ and $\mathbf{I}$ using $\mathbf{SP}$ for 3 forums over time 136
D.3 Plots of all 10 roles of interests for the observed and prediction number of users using $\mathbf{DP}$ and $\mathbf{SP}$. Time on the x–axis is relative to the last observed time point t_0, at week 110 such that 0 is observed. 136
D.4 The coverage against MSE of different methods and formulations over all 90 forecasts for forum 264 137

List of Tables

1.1 The roles (including the inactive users) that best represent the dataset under study by Rowe et al. (2013) 3

2.1 Basic summary statistics for the dataset (in the order of 10^4) 11
2.2 The 8 different type of users as summarized by (Brandtzæ g, 2010) 15

3.1 The average error and the number of times a scenario has achieved the lowest error over a 90 week period for forum 353 42
3.2 Summary of error over 90 consecutive weeks for all methods, each with a 10 week ahead forecast for three different forums. Both formulations of the inactive users were used with initial observation in the estimation either at $\mathbf{m}(1)$ or $\mathbf{m}(N-10)$, where N is the total number of observed mass. Numbers highlighted in green is the lowest MSE achieved between the methods under the same number of observations. 49
3.3 Forum 50, with additional information after splitting the time period at 90. Summary of error over 90 consecutive weeks for all methods, each with a 10 week ahead forecast for three different forums. Both formulation of the inactive users were used with initial observation t_0 in the estimation at either $\mathbf{m}(1)$ or $\mathbf{m}(N-10)$, where N is the total number of observed mass at the current week. 51
3.4 Summary of error over 90 consecutive weeks for the three stochastic methods as described in Section 3.5, each with a 10 weeks ahead forecast for three different forums. 54

4.1 Forum 353 with first 100 observations. Demonstrating the difference in inefficient factor (Ineff) and the time taken (seconds) between the 3 methods for regression coefficients for 1×10^5 iteration 73
4.2 Variables thought to be useful in predicting the number of users moving in and out of communities 83
4.3 The average DIC per observation over 90 weeks for both the number of users joining and leaving for forum 353 using different methods and priors 84
4.4 The average DIC per observation averaged over 90 weeks using PMLN that model both y_{Join} and y_{Return} simultaneously with different prior for forum 353 84
4.5 The average DIC per observation over 90 weeks using the PLN–AR(1) model under different prior for forum 353 85

4.6 The MSE (top) over a 10 weeks ahead forecast averaged over 80 weeks for both the number of users joining and leaving for forum 353 using different methods and priors 86

5.1 Average MSE of the out–of–sample forecast error over 90 consecutive weeks for all methods, each with a 10 weeks ahead forecast for three different forums. Both formulations of the inactive users were used with initial observation t_0 in the estimation at either $\mathbf{m}(1)$ or $\mathbf{m}(N-10)$, where N is the total number of observed mass. 92
5.2 Summary of error over 90 consecutive weeks for the three stochastic methods as described in Section 3.5, each with a 10 weeks ahead forecast for three different forums. 93
5.3 Percentage of times **SP** achieved a lower MSE than **SE** when comparing within each of the formulations, i.e. 45 in forum 353 under **W** indicate that **SP** has a lower MSE for 45% of time 94

A.1 The number of total and active users at two different time point for the 3 forums of interest 115
A.2 Summary statistics of those forums that contain 80 % of the total number of threads generated. Ordered by the number of threads generated per month since the starting date 119
A.3 Summary statistics for the same list of forums as in Table 2 on a thread level, where % No reply indicates the percentage of thread within the forum that has gotten no reply up until the last observed time as of data extraction. The column "% Solved in same day" is only a percentage for threads that has been solved up until the last observed time. 120

B.1 The expected value of **P** for forum 353 using 100 observations. A 0 indicates that the edge does not exist and 0.0000 is some small value 121
B.2 The estimated **P** matrix using the natural bounds of $[0, 1]$ for forum 353 using 100 observations of **P** 122
B.3 The estimated **P** matrix using bounds defined by the past observed rate $\min_{t:1,\ldots,T}\{\mathbf{P}_{i,j}(t)\}$ and $\max_{t:1,\ldots,T}\{\mathbf{P}_{i,j}(t)\}$ for forum 353 using 100 observations of **P** 122
B.4 The average error and the number of times a scenario has achieved the lowest error over a 90 week period for forum 56 122
B.5 The average error and the number of times a scenario has achieved the lowest error over a 90 week period for forum 256 122
B.6 The average error and the number of times a scenario has achieved the lowest error over a 90 week period for forum 264 123

Nomenclature

W	Compartment model formulation without the inactive compartment
I	Compartment model formulation using the inactive compartment
DL	Deterministic - Last transition matrix
DE	Deterministic - Expected transition matrix
PEL	Deterministic - Transition matrix estimated using single step update
PENL	Deterministic - Transition matrix using estimated multi step update
DPL	Deterministic - Penalty factor estimated using single step update
DPNL	Deterministic - Penalty factor estimated using multi step update
DWL	Deterministic - Weight vector estimated using single step update
DWNL	Deterministic - Weights vector estimated using multi step update
SB	Stochastic - Binomial formulation
SG	Stochastic - Gaussian formulation
SE	Stochastic - Expected transition matrix
SP	Stochastic - Penalized transition matrix
SW	Stochastic - Weighted transition matrix

Declaration of Authorship

I, Edwin Tye , declare that the thesis entitled *Forecasting User Roles In Online Communities* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission

Signed:.....

Date:.....

Acknowledgements

First, let me thank my supervisors Thanos Avramidis and Jörg Fliege who managed to put up with an idiot such as myself. Without their guidance and patience, I could never able to complete my doctoral study. I would also like to thank ROBUST (EU–FP7 project grant no. 257859), which provided the funding for my study. Special thanks to Adrian Mocan and Matthew Rowe who provided me with the data that I used extensively in this thesis. I would also like to thank all the partners in the ROBUST project as they increased my exposure to the cutting edge research in information and communications technology.

Many thanks to those that shared an office with me, especially James Heppell and Mushota Kabaso for the numerous colorful (and definitely not politically correct) discussions. Phillipa Hiscock who worked in the same project with me, where we embarked on a journey that any description with my limited vocabulary simply fails to do it justice. Thanks to Amin Saied, Arash Gourtani, Chris Cave, Mark Bass and other fellow PhD students who have educated me on the countless topics in which I am simply a moron. Last but not least, Charles Heppell, Dirk Banholzer, Jamie Foster and Stuart George who kept me well entertained during those lunch times when I just needed an escape from work.

Moreover, I would also like to thank my parent, because it is customary to do so and my relatives, because they are all related to me. Finally, my badminton clubs, as they essentially served as safe–houses because none of the other members cared about my research.

Chapter 1

Introduction

1.1 Motivation

As there are an ever increasing number of people using the Internet, companies (supermarkets, advertising agencies and telecommunication providers etc) have begun collecting information about us while we consume information and products. We are also sharing our personal information willingly on social network websites like Facebook, that has grown to over one billion users (Facebook, 2013).

The explosive growth of online communities will inevitably continue in both social and business domains (Ovadia, 2013) where a lot of these communities have long lasting values. Anderson and Huttenlocher (2012) recognized that Question and Answers (Q&A) web sites form a knowledge repository similar to Wikipedia, and Brandtzæ g (2012) concluded that people who used social networking sites enhanced their existing interactions and built stronger relationships in the long run – when compared to those who did not.

So, how is an online community defined? Preece (2000) noted that the term *"community"* has changed throughout the years in sociology literature and gave a working definition; an online community consists of a set of people who interact with one another by sharing information or exchange of service via an online platform.

Naturally, there are many types of communities as they can be work related or of personal interests. The Stack Exchange Network (`stackexchange.com`) is an example where both types of communities exist on the same platform. This means that people can participate in multiple communities online just as they do offline.

Each online community is managed by a *community manager*, who may also be the owner of the community. Community managers oversee the structure and growth of an online community, ensuring that the community is serving its purpose. So it is important for community managers not only to understand the current structure and behavior of their communities, but also in the future. We aim to provide forecasts and predictions that will enable community managers to make informed decisions, more precisely, the future role composition of a community.

Role composition represents the number of users of a community in each of the possible role, where a *role* categorizes the type of behavior a user exhibits. The behavior of an individual in a particular community dictates his social role and (s)he may interact differently with members in different communities (Delamater and Myers, 2010). As time goes on, the role of a user in a specific community may change. These changes then affect the role composition and therefore the structure of the community. A community manager has the ability to act if he knows the future role composition of his community, depending on his own view of what an undesirable role composition may look like. Not only does the role composition act as an indicator to how *"healthy"* a community is, it has also been used to predict the number of posts (a common measure of health) in a community using the number of users in each of the role as covariates (Rowe et al., 2013).

Our task is based on this overview of an online community and we propose a method for forecasting future role composition. This information will then allow the community mangers to see the progression path of these groups of individuals in their respective roles. It is assumed here that a manager of an online community will be interested in the medium to long term prediction, and have in mind what an undesirable composition of member types looks like as per the goal of the community.

1.2 Problem Statement

Assuming that it is possible to characterize the participants (users) in an online community into their respective social roles. Our aim is to forecast the number of users in all the roles for some future time.

More concretely, we make forecasts for online communities found in the forums of SAP Community Network (SCN). These forums were analyzed by Rowe et al. (2013) who classified all the users into one of the 11 roles in Table 1.1 on a weekly

basis. Given the historical role labels for all the users for time $t = 1, 2, \ldots, N$ with

0	Inactive
1	Mixed Novice
2	Distributed Novice
3	Focused Novice
4	Knowledgeable Member
5	Knowledgeable Sink
6	Focused Expert Participant
7	Focused Expert Initiator
8	Mixed Expert
9	Distributed Expert
10	Unclassified

TABLE 1.1: The roles (including the inactive users) that best represent the dataset under study by Rowe et al. (2013)

$t = N$ being the current time, we wish to make q time step head forecasts for the number of users in all the roles.

1.3 Description of Tasks

Denote $m_j(t)$ as the number of users in role $j \in \{0, 1, \ldots, 10\}$ at time t, computed by summing over the class label of n observed users

$$m_j(t) = \sum_{s=1}^{n} \mathbf{1}\left\{\mathbf{z}_s(t) = j\right\} \tag{1.1}$$

where $\mathbf{z}_s$ is the k dimension indicator vector of role for user s. The quality of the forecasts are based on the Mean Squared Error (MSE) between the forecasts $\hat{\mathbf{m}}(t)$ and observation $\mathbf{m}(t)$

$$\text{MSE} = (q(k-1))^{-1} \sum_{t=1}^{q} \sum_{j=1}^{k-1} (\hat{m}_j(t) - m_j(t))^2, \tag{1.2}$$

which is summed over the k active roles. The inactive compartment is not included because we assume that the health of a community is only based on the active users and the roles they take on. When MSE alone is insufficient to separate the performance of the models, our secondary concern, the deviation between the forecasted and actual number of active users, will also be used.

The total number of active users changes as time progresses. This is due to new users joining the online community as well as active users becoming inactive. Number of users in role j at a particular time point is dependent on:

Problem 1
The number of users not in role j take on the role j

Problem 2
The number of new users joining role j from outside the community

We treat these two problems separately because the number of users with role i becoming role j between two time point is bounded $m_i(t)$, the current number of people in role i at time t. Whereas the maximum number of new users joining a community is the population of Earth minus the number of currently registered users for that community. We assume that this quantity is unbounded. The number of forecasting steps were chosen to be 10, after consultation with SAP.

Our focus in on three of the most active forums, Forum 353, 264 and 256, which represents the different dynamics observed. A summary on the total and active number of users at the start and end of the dataset for the three forum of interest can be seen in Table A.1. Forum 256 has nearly doubled its active user in just over 2 years (120 weeks) while the other two only has a slight increase.

The dynamics for each of the compartments for the three forums are displayed in Figure A.1, Figure A.2 and Figure A.3. The graphs show that forum 353 are relatively stable with forum 256 increasing in certain roles while forum 264 behave similarly to forum 353 but with bigger changes. Further differences can be seen in the network graph for forum 264 (Figure A.5) and 353 (Figure A.6) where the former is much more connected where 3 of the most influential user in forum 264 communicate with each other extensively.

1.4 Outline of the Thesis

This thesis is organized as follows: We begin by introducing the background of the problem in Chapter 2 where some general information regarding online communities is presented. Details on the origin and structure of the dataset we use throughout this thesis is described in Section 2.2 before moving on to explore the

literature on social networks as well the clustering. We then proceed to describe our modeling efforts in the next two chapters.

The two subproblems mentioned in the previous section are isolated and tackled separately in Chapter 3 and Chapter 4 respectively. Results for the individual problems can be found locally within their respective chapters, where Chapter 5 presents the results when the models developed in Chapter 3 and 4 are used in parallel. More concretely, the remainder of this thesis is organized as follows:

We begin by introducing the basis of our model in Chapter 3. This chapter introduces a compartment model with multiple implementations that generate forecasts based on observing user roles over time. The focus of this chapter is to model the movement of existing users between the roles (**Problem 1**). We defer the fluctuation of active users (**Problem 2**) to chapter Chapter 4 by assuming that the number of users joining can be observed even for the forecast. Additionally, we also study the possibility of modeling the number of inactive users returning from inactivity as an unbounded count rather than the limited by the number of inactive users. Both deterministic and stochastic compartment models will be presented with the rationale behind each model discussed.

Chapter 4 looks at the prediction on the number of new users joining (**Problem 2**) and returning from inactivity. We hypothesize that the number of new users come from a Poisson process. This chapter goes through the existing literature on count data modeling and especially on Poisson regression. A number of variants of Poisson regression are presented along with the estimation techniques before showing the performance of the various models.

The two are then combined together in Chapter 5 to provide a multi–step forecast where the results will also be discussed. We finish off by summarizing the work of this thesis, discussing the limitations of our models and future research directions in Chapter 6.

Chapter 2

Background

2.1 Introduction

An online community is formed when people interact with one another via the Internet. Preece (2000) noted that the definition of an online community has changed through the years which is unsurprising given the new ways of communication and reinvention of the existing platforms.

The studies of online communities employ techniques used in networks analysis (Wasserman and Faust, 1994), and have been used in areas such as preventing the spread of disease (Christakis and Fowler, 2010), obesity (Bahr et al., 2009), computer virus (Newman et al., 2002), marketing (Trusov et al., 2009) and information spread (Cha et al., 2009). Our focus is on the roles of a community, specifically in predicting the number of users in each of the roles at some future time.

The structure of the dataset will be described first in Section 2.2. Then a summary of the statistics in Section 2.3 before going onto defining the health of a community in Section 2.4. Finally, the notion of *role* as well as its implication in online communities in Section 2.5.

2.2 Data Structure

SAP Community Network (SCN) is a social networking platform. Primarily designed for users and developers of SAP products to discuss and share ideas, as

well as interact with the company itself. The term *user(s)* will refer to anyone that participated in SCN from here onwards, where each individual user is identified by a *unique registration id* on the platform. There are a variety of ways for users to interact with each other, i.e. blogs, starting a poll, private messages, forums and document collaborations etc. Also, Webinars and videos by SAP introduce/teach different solutions and usage of their products to current as well as potential customers.

The dataset under study is derived only from the forum sections of SCN; a total of 95 forums with data from Feburary–2004 to November–2010, which is approximately a third of the total size in terms of information stored (measured in Gigabytes) at the time of extraction. The purpose of the forums is primarily to serve as an arena for *Question and Answers* (Q&A) when the answers cannot be found elsewhere, but tips and information about specific products can also be found.

SAP has designed the SCN platform such that each forum only represents a specific topic, usually a product of SAP. Furthermore, a moderator of SCN has the authority to move inappropriate topics to a more suitable forum, as a way to ensure that relevant information are directed to people that might be interested. Therefore, we assume that each online community is established based on an individual forum. This assumption simplifies the analysis to allow for forecasts be performed for each individual forum while ignoring the rest. It also places a limitation on the amount of information available when making the forecasts as some forums may be correlated.

As we use the independence assumption on the forums, a user joins a community at the point where they post their first message. This approach is taken because we have no data regarding passive participation of the users, i.e. when (s)he visited the forum previously without posting a message. Furthermore, the assignment of roles (Section 2.5) to users is solely based information derived from these active participation via messages.

A collection of messages that all links to the same parent is called a thread, and each thread belongs to a specific forum. The parent of a thread is the first message that was created by a user. This thread is related to the other messages either directly or indirectly through a reply tree structure as seen in Figure 2.1.

The user that posted this first message is the owner of the thread and (s)he can award *points* of value 2,6,10 (SAP, 2012) for any message within the thread posted by a respondent, but with an upper limit of one 10, two 6's, and an unlimited

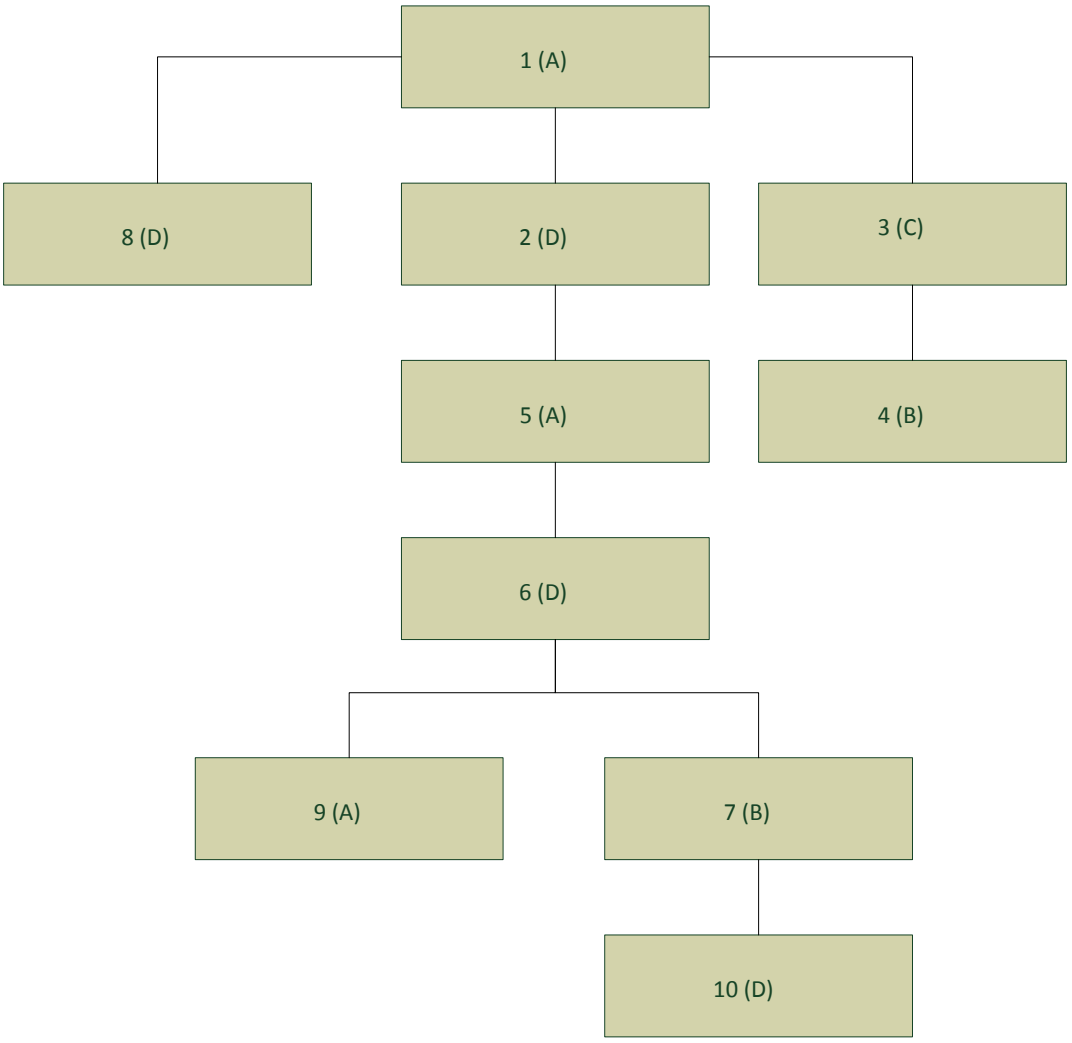


FIGURE 2.1: A reply tree structure where each box corresponds to a message, the number and letters indicating the arrival order and the users who posted the messages respectively

number of 2's in a single thread. General guideline for the points awardation are 10 – problem solved, 6 – very helpful, 2 – helpful. The thread owner can award the points as (s)he sees fit after considering the value of each message.

This point system is the basis of the SCN "Contributor Reputation Program" that keeps track of the points scored by individual users over a rolling 12 month period. It encourages users within the community to help each other, similar to other Q&A style sites like `stackexchange.com` or `answers.yahoo.com`. It has to be noted here that the current system has changed slightly from one that was used whilst the data were generated. The newly implemented scoring system on SCN has been simplified to 0,5,10.

Using standard notation, the graph $G = (V, E)$ contains the set of vertices V

that represent users in a community and E the set of edges between users. The existence of an edge depends on the different interpretations, some authors only treat a direct reply (message 3 to message 1) as an edge while others takes into account preceding messages (message 4 to message 1 also forms an edge) instead of individual messages (Toral et al., 2009). The difference usually stems from the construction of the forums, as some platforms only allow reply to the whole thread rather than to specific messages within a thread. For example, a reply at any point of the thread may imply that (s)he has read previously connected messages i.e. assuming that when user B posted message 7, (s)he has read messages 1,2,5 and 6 but not message 3. An example of what G looks like with E weighted according to the number of messages can be seen in Figure A.5 and Figure A.6.

There are also different ways to count the number of exchanges between the same two users. Directed edges are most commonly used and it can be binary or weighted. For example, a directed edge was created from user D to user A due to the reply of message 2 to message 1 and future interaction (messages 8 to 1 and 6 to 5) do not contribute further. Alternatively, weights are assigned to the edges according to the frequency of contact, which will be 3 as user D has replied to user A three times as seen in Figure 2.1. Both approaches can be modeled by distributions from the exponential family (Holland and Leinhardt, 1981; Wang and Wong, 1987; Handcock et al., 2007; Krivitsky, 2012), the logistic and Poisson distribution respectively and the corresponding regression formulations have been developed in the statistics literature.

2.3 Data Overview

A brief summary of the whole dataset and the 6 full years (Jan–Dec) can be seen in Table 2.1 with a more detailed version in Appendix A.1. Unsurprisingly, the number of users as well as the number of messages and threads increase year on year as Internet has become more readily available and affordable. As previously mentioned, the total number of forums have increased and they contribute to the additional volume in all areas for SCN. Some existing forums have been attracting new users and more content has been generated as time progresses, while others have more or less remained at the same level, see Figure 2.2. It is important to note that users cannot delete their accounts on SCN as part of the platform design, so the total number of users is constantly increasing.

Type	Total	2005	2006	2007	2008	2009	2010
Users	107	2	8	18	33	38	49
Threads	450	4	17	41	104	136	163
Threads Solved	104	2	4	8	25	33	31
# Users that solved a thread	11	0.2	0.7	1.5	3.5	4.3	4.4
Messages	1900	16	61	158	422	564	629

TABLE 2.1: Basic summary statistics for the dataset (in the order of 10^4)

A more detailed summary of 25 forums, selected based on the activity in terms of new threads per month, can be seen in Appendix A.1 in decreasing order of activity. These 25 forums combined made up of just over 80% of the total thread volume. Even though the forums shown in Appendix A.1 were the most active, they still varied significantly between them as the highest one generated over 682 threads per month in comparison with the smallest of 93. The total number of users includes all those who have ever posted in the forum which does not reflect the total number of active users at any given time. This is because active users are those who have posted within some time period and most of the users only contribute sparingly.

There are about 42% of the total users who can currently be considered a *"One post wonder"* (people that only posted a single message), where 85% of them started the only thread they posted in. This is indicative of Q&A type forums where people go when seeking answers, and it can be seen from Table A.2 that nearly 24% of messages start a thread. At the same time, 20% of users posted 85% of the total messages, and 20% started 76% of threads. Whittaker et al. (1998) had similar findings on newsgroups where most messages were attempts to start a conversation and activities are dominated by a small set of members in the community. Only 19% of users have scored any points and just over 10% of the population had solved a thread, further indicating that contributions came from a minority of users.

The highest number of forums any user has taken part in is 33, but only 0.4% of total users have participated in more than 10 forums. Additionally, 64% of the users only ever participated in a single forum, and the number increases to 82% when including those who posted in two forums. This shows that in addition to the *"One post wonder"*, most other people only concentrate in a single forum as well, which further demonstrates the separation between forums.

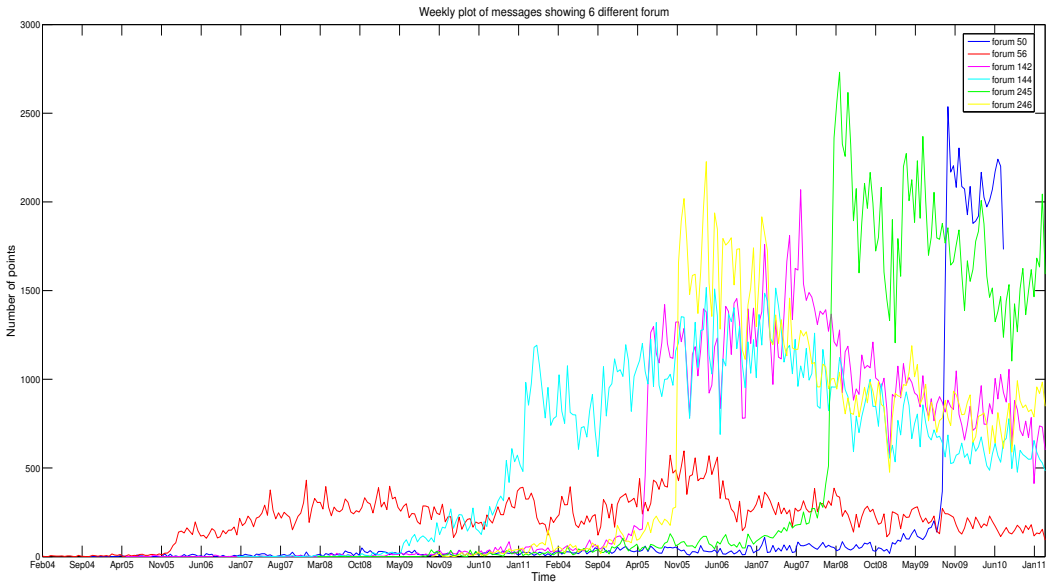


FIGURE 2.2: Time series of the number of messages per week for the top 5 forums in terms of total messages

2.4 Health of an online community

For a business organization like SAP which has invested time and money into setting up a platform for people to interact, it is important that the invested resources do not go to waste. This means that the communities they maintain should be *"healthy"*. But how do we tell whether a community is healthy or not? And what are the factors that affect such health indicators? One of such models is the "Information System Success Model" (ISSM), originally proposed by DeLone and McLean (1992) and later revised (DeLone and McLean, 2003). ISSM discusses appropriate measures of success on many levels including the quality of the system, information and user satisfaction to name a few. However, the authors also noted that the model only acts as a framework and researchers should select the appropriate dependent variables based on their own objective and interpretation of success.

ISSM recognizes that user surveys are a good way to measure user satisfaction. An obvious pair of answer and question is, "Are the participants of the community satisfied?", and if the answer is "yes", then there is confidence in saying that the community does provide value to users. Unfortunately, it requires asking every single user the same question at regular intervals. Survey response also assumes that the users are not lying, although it can be argued that the investment in time and energy when answering questionnaire alone is proof that the users value the community. Still, surveys of this kind create a bias when a community has many

lurkers, a person (may not be a registered user) who never posts despite following the forum. It can be argued that because lurkers do not post and generate content, they do not affect the success of a community even though they may make up a large proportion of the population. But at the same time, revisiting the same community implies some sort of attachment or satisfaction gained. Nonnecke and Preece (2000) had found that more than half of members in discussion lists were lurkers and provided a detailed list of reasons for why they possessed such behavior.

When users of a forum stop gaining gratification from active participation, new content will no longer be generated. This lack of interaction signifies the end of an online community under the definition of an online community by Preece (2000). But at the same time, existing threads may already provide valuable information for an outsider (a person not previously involved in the community) looking for answers to their questions. Anderson and Huttenlocher (2012) tried to predict this long term value of a thread, based on the number of views a thread obtains in the future. Such measure is similar to the ones mentioned in ISSM where it suggested that the frequency and the total number of times a user visits are more accurate representations of satisfaction instead of data obtained by a user survey. Similarly, a suitable answer for the thread owner signifies user satisfaction as (s)he has achieved his/her original goal. Statistics of SCN as seen from Table A.3 shows that on average 62% of threads were either *answered* or *solved*. This suggests that the forums are generally healthy in this aspect; compared to 69% in Stack Exchange Network (Anderson and Huttenlocher, 2012) and 50% in Yahoo Answers (Agichtein et al., 2009).

Content generation in terms of total number of threads as well as messages is a good indicator for the active participation of a forum, but it is not a good measure of interaction between users given that a new thread may not have any replies to it. Similarly, Nolker and Zhou (2005) found "chatters" in online communities, where two users reply to one another in the same thread or different threads.

This can result in a high volume of both messages and threads not only for each of the individuals but also for the whole community. Therefore, the health of a community based on volume activity measures alone are not sufficient, because the number of people participated should also be taken into consideration. In fact, Morris and Ogan (1996) noted that the Internet can be considered a mass medium, hence the theory of critical mass (Rogers, 2003) should also apply. A critical mass occurs when there are enough people involved (say using a service) for it to be self–sustainable. When the number of people using the service exceeds

the critical mass, it creates a snowball effect such that it becomes beneficial for other people to also use the service. Therefore, the number of user participating in a community also reflects on the maturity of the community (Iriberri and Leroy, 2009).

Since the forums on SCN are product specific, interest levels are usually lower for niche products or those at the end of their lifecycle when compared to a newly launched products or a core program. These forums are also set out to be Q&A forums, which means that a new thread is most likely created by a current consumer who encountered difficulties while using the product. Therefore, a lot of new threads may be due to a failure in the product itself or users facing difficulties after an update in addition to the natural increase of consumers.

2.5 Roles in Online Communities

Social role is "rights and duties attached to a given status" as defined by Goffman (1959), and every person carries a different role depending on the situation/community they are in at any particular moment. The role of a person determines their behavior, for example, a person can be a mother, daughter, wife and teacher etc and she will present herself in a suitable manner depending on who she is dealing with and what she is doing. The research on roles have gained a lot of interest in areas like diffusion processes for example (Rogers, 2003), a study on the spread of information or disease in different types of networks. Information flow is also related to the idea of weak ties (Granovetter, 1973) where the nodes that connect two cliques for example, are central to the process. These nodes can be interpreted as the "leader" of a clique (community) who have high influence and form connections with other cliques. The identification of these nodes can help in stopping viruses spread on computer networks, epidemic outbreaks and better marketing strategies just to name a few. But of course, there are different measures of importance and the usage of such information is network dependent where the number of edges going in and out of a node may be a good indicator to the role type.

Role analysis helps in identifying what type of community it is by 1.) The number of different roles and 2.) The proportion of users in each of the roles (Rogers, 2003). Himelboim et al. (2009) found certain roles amplified the amount of messages in a thread where others help and Rowe et al. (2011) found that the current role information (number of users in each of the roles) was useful in predicting the

volume of contents generated in the future. Other authors have also found that certain roles were important for certain communities to succeed, such as discussion bulletin board (Nolker and Zhou, 2005) or an innovation forum (Hautz et al., 2010). Therefore, it is beneficial for a community manager to know not only the current composition of the roles, but also in the future. For example, assume that the "answer" type of user is important for forums as they deposit their knowledge that is accessible to all users (Welser et al., 2007), a community manager may ask "Will the number of people who answer questions decrease in the future?" So, how do we find out the role of a user?

Given that a role determines a user's behavior, the reverse is also true and it should be possible to infer the role given the behavior. Brandtzæ g (2010) summarized 22 studies on the behavior in the Internet media market and provided a unified typology. Even though many of the roles were proposed without any naming consistency, the paper grouped them into 8 distinctive type by the user usage pattern (of each role) as follows

Non–users	Sporadics	Debaters	Entertainment users
Socializers	Lurkers	Instrumental users	Advanced users

TABLE 2.2: The 8 different type of users as summarized by (Brandtzæ g, 2010)

2.5.1 Clustering Methods

The most common technique in those 22 studies analyzed by Brandtzæ g (2010) is cluster analysis. Although all 22 studies were focused on online communities, the notion of clustering communities (in general) can be found as early as Davis (1967). A nice review on this topic can be found in Jain et al. (1999).

Some of the most common data clustering techniques will be described here as they will be referred to in Chapter 3. First, it is important to note that the term *clustering* can be interpreted as data clustering, graph clustering or even clustering coefficients on a graph, depending on the discipline. Newman (2003) remarked that data and graph clustering often get confused and that the algorithm for one can be used on the other in some situations even though it may not work well. The goal of data clustering is to find a useful pattern in the data, especially in high dimension multivariate data that is hard to visualize.

Both graph and data clustering are automated searches that aim to group observations in a dataset from a similarity perspective, but the way they do are not

necessarily the same as the underlying hypotheses differs. Data clustering is based on the idea that a subset of observations is generated by the same process, which forms individual clusters. Graph clustering comes from a structural perspective on the whole or subset of vertices V. It can either be based on the quality of the clusters, i.e. change in connectivity before and after the clusters are formed, or on a similarity measures, i.e. the similarity of connection between the vertices (Schaeffer, 2007). For vertices in a clique, they all have similar connections (to one another) but different to the rest of the graph. This is similar to the setting when part of a dataset has observations very close to each other and indeed the algorithms for graph and data clustering are interchangeable at time (Dhillon et al., 2004).

The most well known data clustering method is the k–means, a term coined by Macqueen (1967). It has been used to find patterns in multivariate observations in many fields of study and is one of the most used algorithm in data mining (Wu et al., 2007). The main idea is that the set of observations can be split into a total of k number of groups. This is achieved by placing each observation into one of the k groups by minimizing the total squared distance

$$f(\mathbf{C};\mathbf{y}) = \sum_{j=1}^{k}\sum_{i=1}^{n} z_{i,j}(\mathbf{y}_i - \mathbf{c}_j)^\top(\mathbf{y}_i - \mathbf{c}_j), \tag{2.1}$$

where $\mathbf{y}_i$ is the observation of user i, $\mathbf{c}_j$ represents the centroid of group j and the latent variable $z_{i,j}$ is an indicator function for user i belonging to group j. Hence, $\mathbf{z}_i$ is a vector of length k that sum to 1 and it can be interpreted as a categorical random variable generated from a multinomial distribution

$$Z \sim \mathcal{M}_k(1, \boldsymbol{\pi}) \tag{2.2}$$

with mixing proportion $\boldsymbol{\pi}$ satisfying

$$\sum_{j=1}^{k} \pi_j = 1, \quad 0 \leq \pi_j \leq 1 \; \forall j. \tag{2.3}$$

The power and popularity behind k–means is that (2.1) can be applied to any measure of distance/similarity such as those produced via the kernel method (Schölkopf et al., 1997, 1998). The kernel method maps the raw observations into a higher dimension feature space in an attempt to discover non–linear relationships in the data, which the clusters generated using squared distance are unable to recognize.

Finding the optimal clusters in kernel k–means can be formulated as spectral clustering problem/normalized cut (on a graph) (Dhillon et al., 2004) and in terms of non–negative matrix factorization (Ding et al., 2005), where efficient algorithm exists.

Regardless of the distance formulation, the k–means algorithm outputs a single label for each observation regardless of how close it is to the centroid of its own group relative to the others. This type of method falls under the category of *hard clustering* as each data point only belongs to a single group. Conversely, *soft clustering* provides the probability of an observation being in any given group. Fuzzy k–means (Everitt et al., 2009) is an example of soft clustering. Other soft clustering techniques include model/distribution based approaches, for example, Gaussian Mixture Model (GMM) (Banfield and Raftery, 1993) produce probabilities of belonging to each of the cluster for all the observations through the use of Gaussian distributions. The general form (density function) of a mixture model for the i^{th} observation given the parameters $\boldsymbol{\theta}$ is expressed as

$$f(y_i; \boldsymbol{\theta}) = \sum_{j=1}^{k} \pi_j f_j(\mathbf{y}_i; \theta_j) \tag{2.4}$$

where $\boldsymbol{\pi}$ satisfies (2.3). GMM is a special case that assumes $f_j(\cdot)$ is a p dimension multivariate normal (MVN) probability density function with the corresponding mean $\boldsymbol{\mu}_j$ and covariance $\boldsymbol{\Sigma}_j$, expressed as

$$f(\mathbf{y}_i) = \sum_{j=1}^{k} \pi_j (2\pi)^{-p/2} \left|\boldsymbol{\Sigma}_j\right|^{-1/2} \exp\left\{ -\frac{(\mathbf{y}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_j)}{2} \right\}. \tag{2.5}$$

Let $\tau_{i,j} = \mathbb{P}(\mathbf{z}_i = j \mid \mathbf{y}_i; \theta_j)$ be the probability that observation $\mathbf{y}_i$ belongs to the j^{th} mixture with θ_j representing the parameters of the probability distribution, then estimating the mixing proportion from the current set of observation is $\pi_j = n^{-1} \sum_{i=1}^{n} \tau_{i,j}$. To assign a label to each individual user, i.e. converting it to hard clustering can be done by finding the mixture with the highest probability for each y_i, the mixture j corresponding to the largest $\tau_{i,j}$. Alternatively, if we only allow a scalar variance $\Sigma_j = \sigma^2 \mathbf{I} \quad \forall j$, then $\tau_{i,g} \to 1$ if cluster g is the closest to $\mathbf{y}_i$ (in terms of $\|\mathbf{y}_i - \boldsymbol{\mu}_j\|^2$) as $\sigma^2 \to 0$.

If the variance is fixed, say $\boldsymbol{\Sigma} = \mathbf{I}$, then k–means can be seen as a special case of GMM where the log–likelihood function (2.5) and the objective function of k–means (2.1) differ by a constant. The difference between the two method lies in

the method of determining the centroids c_j and parameters θ_j. In k–means, each data point only contributes once to the centroid of the group it belongs, but it contributes to all the parameters in every component of the mixtures $f_j(\theta_j)$ in mixture modeling. The results obtained in the estimation process are not necessarily optimal as there will be many local minima for both (2.1) and (2.4), especially when the data are of high dimension.

Fraley and Raftery (2002) further detailed the difficulties and limitations of using GMM, especially when the data are not Gaussian, linear or of quantitative measure. If the data are not Gaussian or linear, it may be possible to solve the issue by normalizing or by using other distributions (McLachlan and Peel, 2000). But the challenge remains if the explanatory variables are discrete or ordinal as converting integers to reals does not necessarily reflect the true difference. Therefore, the measure of distance between data point is the most important aspect of clustering, and one that usually requires expert knowledge given the data on hand. The choice of model and method reflects a prior believe on the data (McLachlan and Peel, 2000), i.e. standard k–means assumes that the data have Euclidean distance separation and variables are of similar measure, where GMM assumes that the data are all normally distributed.

Another disadvantage is that the number of clusters have to be pre–determined before running the algorithm. Therefore, a range of values, $k = 1, 2, \ldots, n$ needs to be evaluated in order to determine the most suitable number of total clusters by using some model selection criteria such as AIC, BIC, Silhouette Coefficient, etc (Kaufman and Rousseeuw, 2005). This is because over–parameterization the number of clusters does not yield clusters with zero observation. Using k–means as an example, if k is larger than the true number of cluster G, then ideally $\pi_j = 0$ for $j = k + 1, \ldots, G$. But because a lower objective value can be obtained from (2.1) by putting the currently empty cluster centroids $c_j, j = k+1, \ldots, G$ onto the observations, such that $c_j = y_i$ for some i, j such that $z_{i,j} = 1$. Hence, a cluster cannot be empty. Advanced methods that allow a more "automated" approach to select the number of clusters in GMM can be found in the literature, but they are all computationally intensive, see Marin et al. (2005); Lee et al. (2008) and references therein.

Sneath (1957) was one of the first to form clusters by linking similar data points one–by–one which is now known as *Hierarchical clustering*, a popular method in graph clustering literature. The basic idea is to divide or aggregate the observations repeatedly using a similarity measure until all observations have been

split/joined up. The data points in each cluster can then be recovered immediately given a pre–specified number of clusters that is required. This similarity measure can be any metric (Euclidean/squared Euclidean/Manhattan distance) that is deemed suitable for the dataset and faces the same difficulty given non–quantitative variables.

A major difference between mixture model and hierarchical clustering is that clusters in the latter approach are formed by combining lower level clusters on the dendrogram, i.e. a new cluster is created by splitting a current cluster while ignoring the rest of the dataset. This is different to k–means or mixture model where an addition cluster on the current partition can take observations from any of the existing clusters. A model (probability distribution) based approach can be incorporated into hierarchical clustering and more details can be found in Fraley and Raftery (2002).

2.5.2 Usage in Role Analysis

Golder and Donath (2004) observed that members of an online community can be classified into several roles, the role that each user takes also carry certain constraints or freedom in their actions. The impact on the community also differs between the roles as the community tends to be defined by the knowledge and belief of the more well known members, who usually generate a lot of posts. But a user who generates a significant amount of contents can also be unimportant to the community if the information can be regarded as spam. Hierarchical clustering and k–means were used in Chan et al. (2010) and Maia et al. (2008) respectively. Visualization techniques are also popular in social network analysis especially for small datasets (Freeman, 2000). Welser et al. (2007) separated the users into groups based on visualization before confirming the feasibility of the groups by regression analysis. Mixture models were attempted by Handcock et al. (2007) which used GMM on the set of spatial distances by projecting the similarity between users using the latent space approach suggested by Hoff et al. (2002). As clustering is performed in a lower dimension (in the latent space) that allows easy visualization, it enables the end user of the algorithm to verify the result against their prior hypothesis of both the number and structure of the clusters.

A user is defined to have joined a community only if (s)he has posted a message. An inactive user is defined to be a user who have not posted for a significant period of time. “Lurkers” are those users who do not post in the forum but consume

contents, so a lurker who has stopped posting will be considered an inactive user as they are not separable in our data.

The classification of "inactive" user also depends on the length of the time period considered. If a dataset is constructed by aggregating all the actions each individual user has performed since the inception of a forum, then none of the users can be considered inactive by the previous definition of user joining a community. On the other hand, if the dataset is constructed by aggregating actions within the past year, then every user who have not within this one year period will be considered as inactive. Therefore, the length of time of participation to use for the classification of inactive users needs to be chosen carefully.

As SCN does not allow deregistration, one obvious problem is making the distinction between inactive users who are still consuming the contents (lurkers) and those who have left the community as they are identical from the data's perspective. Therefore, the time window for which to observe the community is very important because not only the dynamics (number/type of users) changes through time, the speed (frequency of post) can also be significantly different. There is usually no justification regarding the size of the time window apart from what appears to be suitable after initial analysis. Indeed, Nolker and Zhou (2005) mentioned in their conclusion that the choice of one year was arbitrary and further research needs to be done in this area.

Additionally, the nature of the clustering algorithms means that every single user is assigned to a cluster. This will either force outliers to be in a cluster of their own or join a nearby cluster. Therefore, explicit splitting criteria have been used to place user into pre–defined groups, similar to SAP's points recognition program (SAP, 2012) that is currently in place. This method is similar to a decision tree without the learning phase, where the biggest advantage is speed. Its simplistic structure allows easy interpretations, but at the same time, requires expert knowledge on the community with regard to the number of roles and the features exhibited by users under different roles.

The same dataset had been studied by Rowe et al. (2013) and found the roles in Table 1.1 to best represent the 33 forums analyzed , based on k–means clustering on the first six months of the dataset using silhouette coefficient as model selection criteria. After identifying the appropriate user type and the corresponding behavior pattern that derived such roles, users are placed in those roles for the rest of the dataset on a weekly basis using a rolling 6 month time window. This ensures that the total number of roles and their relative meaning stay consistent

over time. We perform our forecasts based on the assignment of the roles on each of the user over the two year period by Rowe et al. (2013). The focus is placed on the existing forums ($>$ 6 months of data) and ignore the newly created ones as they are deemed to have insufficient data. A forum is treated as a single community and each user can have a different role in different forums at the same time point. The role of a user can change between time points and is considered to be inactive when (s)he did not post any messages within the observed time window.

2.6 Modeling of Online Communities

Analysis on online communities benefited from the previous works on offline communities, where existing techniques developed in other fields (graph theory, social science, statistics etc) are also applicable. Graph growing models are probably the most well known method to model and infer an online community. Random graph (Erdös–Rényi model) (Erdös and Rényi, 1959) or preferential attachment (Barabási–Albert model) (Barabási and Albert, 1999) are widely used because they have been observed in real networks. An online community can then be forested by identifying a model that fit the structure and attributes of a community with the correct parameter. Forecasts are then obtained by via simulation using the graph growing algorithm and the current state of the community. However, we only wish to infer the "active" users and their respective roles and not all the users who have participated in a forum.

Granovetter (1973) originally and later revised Granovetter (1983) the theory of week ties, where he argued that the only possible connection between two strongly connected social networks is a week one. This idea of macro interaction of communities that are subsets of a larger communities fueled the introduction of blockmodels (White et al., 1976; Boorman and White, 1976). Blockmodels were an attempt to not only separate sub–groups (blocks) of a community but also the role structure within a block. This was then further developed by Holland and Leinhardt (1981) into the p_1 graph model, using distributions from the exponential family to model the existence of an edge between two people, and has since been generalized into the class of Exponential Random Graph Model (ERGM) (Wang and Wong, 1987; Wasserman and Pattison, 1996, 1999; van Duijn et al., 2004; Snijders et al., 2006).

ERGMs have been used to study communities on its own and in conjunction with other methods. Hoff et al. (2002) modeled the existence of an edge between

users in a social network using logistic regression and latent variables. Clustering were then performed on the latent variables to identify the different types of roles and the belonging of each user to his/her respective role (Handcock et al., 2007). But, like all clustering techniques, the algorithms works on static data that is time sliced data over some pre–determined time period. The assumption of an exponential family distribution on the edge provides a flexible formulation, such as the discrete time temporal extension to ERGM developed by Hanneke et al. (2010). This temporal extension allows analysis and prediction to be performed on the social network using a series of observations taken at equally space time points. But the model can only accommodate observations that all have the same number of users, as the model relies on having the same normalizing constant, which does not cancel out between time steps when the graph changes size, i.e. when users join or leave.

Both of these approaches only deal with certain aspect of our problem where a static ERGM does not forecast the role composition for some future time period while the temporal extension has fixed number of users. Furthermore, using a time sliced set of data implies that only the active users during the specified time period are observed. This in fact ignores the valuable information provided by the "inactive users" (those that have not participated in the community for a prolong period i.e. six months, one year), as we built our model based on the observed role of every user in a community over time.

Chapter 3

Main Methodology

3.1 Compartment Model Introduction

Compartmental models have been used in many disciplines (Godfrey, 1983) and are usually represented as a graph where compartments are the nodes. A compartment represents a collection of contents that is indistinguishable from each other, with the ability to travel into another compartment, this ability being represented by a directed edge.

A popular usage of compartment models is in the field of epidemiology where the SIR (susceptible, infectious, recovered) and its extension SEIR (E for exposed) model (Bailey, 1975) are used to model the progression of disease. In both of the models, each compartment represents the people at a particular stage of the disease and a set of differential equations are used to model the rate of people transitioning from one stage to the next. The modeling of SIR has also been applied to individuals in a network (Newman, 2003), where the infection is passed on from one node to the next. Such models are central to the understanding in spreading of diseases, which is important in both prevention and detection of outbreaks as demonstrated by Christakis and Fowler (2010).

Compartmental models is also a general case of *population models* that are commonly used in ecology to model the changes in the size of a population. Denote $N(t)$ as the population of a species at time t. Then the simplest population is

$$N(t + \Delta t) = N(t) + B(t, \Delta t) - D(t, \Delta t), \tag{3.1}$$

where the change in population size is governed via birth $B(t, \Delta t)$ and death $D(t, \Delta t)$ in the same time period (Turchin, 2003). This can be used to model say, the total number of active users in an online community where the number of new users joining and leaving the community is $B(t, \Delta t)$ and $D(t, \Delta t)$ respectively. A more complex population model, i.e. one that models a set of age groups in a population, is obtained by extending $N(t)$ in (3.1) to a vector with the corresponding movement between age groups. Population models were also used to model the change in size of multiple species (Law and Blackford, 1992; Scheffer et al., 2001; Gibson, 1998; Yoshida et al., 2003; Geritz and Kisdi, 2004; Scheffer and van Nes, 2006), most of them were based on the famous Lotka–Volterra equation (Lotka, 1925) which described the interactions between a predator and prey.

Our aim is to model the number of users in each of the roles through time, where the role of a user can change as time progresses. Here, each compartment is used to represent a specific role and the mass of a compartment is the total number of users in that role. An example can be seen in Figure 3.1 with 4 different roles as well as additional users joining the system via a single role. It also demonstrates that all possible paths need not exist as users in the *New User* compartment only have edges going in the outward direction. In the setup of Figure 3.1, none of the users are allowed to leave the community, so the total mass of the compartments increases with users joining the community. On the other hand, the size of the community is subject to fluctuation if the users are allowed to leave a community.

The remainder of this chapter is as follows. An overview of the compartment model formulation will be discussed first in Section 3.2, where the connection between a linear compartment model and Markov chain is made. In particular, Section 3.2.3 introduces a variant of the base model that does not model inactive users directly. Estimation of the parameters in the deterministic setting will be discussed in Section 3.3 and Section 3.4 before moving on to the stochastic versions in Section 3.5. Finally, Section 3.6 compares the result between the different estimation methods discussed throughout the chapter.

3.2 Mathematical Formulation

Given the classification of role on each of the user at (equally spaced) time points $t = 1, 2, \ldots, N$, observed weekly, the number of users transitioning from role i to

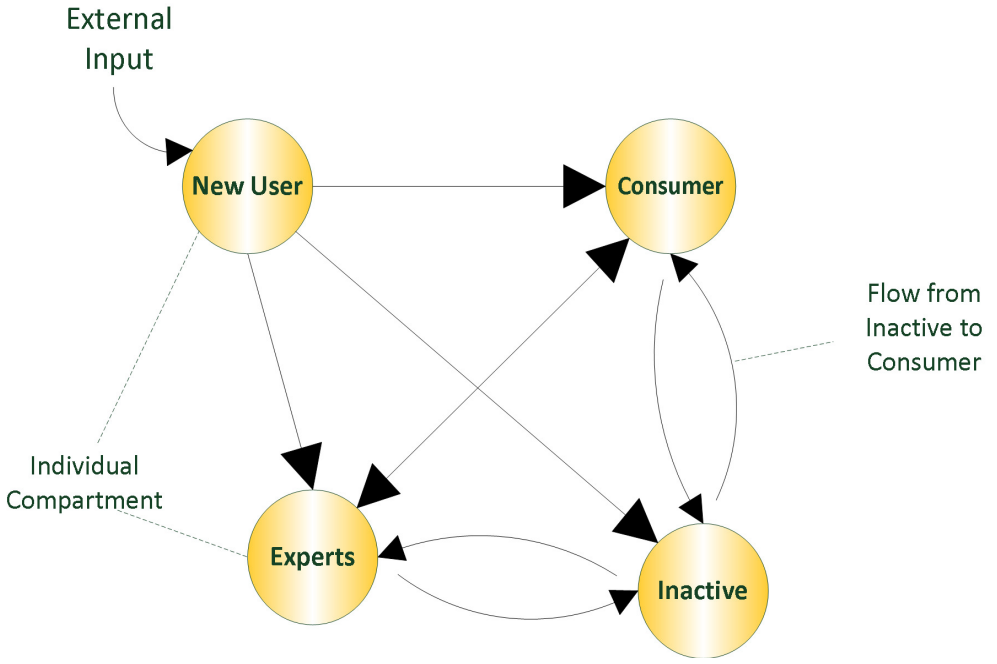


FIGURE 3.1: An example of a compartment model with 4 different user roles in a system

j between week t and week $t+1$ is

$$v_{i,j}(t) := \sum_{s=1}^{n} \mathbf{1}\left\{\mathbf{z}_s(t+1) = j, \mathbf{z}_s(t) = i\right\}, \tag{3.2}$$

where $\mathbf{z}_s(t)$ is the indicator vector of length k (from Section 2.5.1). Let $m_i(t)$ be the number of users (mass) for role i at time t and $\mathbf{m}(t)$ the corresponding vector for all roles . We define our compartmental model as a discrete time Markov chain (DTMC) with contribution from external sources $\mathbf{a}(t)$ as

$$\mathbf{m}(t+1)^\top = \mathbf{m}(t)^\top \mathbf{P}(t) + \mathbf{a}(t)^\top, \tag{3.3}$$

where $\mathbf{P}(t)$ the transition probability matrix with elements

$$\mathbf{P}_{i,j}(t) := \mathbf{m}_i^{-1}(t) v_{i,j}(t) \tag{3.4}$$

given the mass $\mathbf{m}_i$ and the flow $v_{i,j}$. The matrix $\mathbf{P}$ is also known as a Markov matrix or a right stochastic matrix, where the following conditions are satisfied

$$\sum_j \mathbf{P}_{i,j} = 1 \; \forall i, \quad 0 \leq \mathbf{P}_{i,j} \leq 1 \; \forall i, j \tag{3.5}$$

The external sources $\mathbf{a}(t)$ at time t is an aggregate from both the number of users joining $\mathbf{b}(t)$ and the number of users leaving $\mathbf{c}(t)$.

$$\mathbf{a}(t) = \mathbf{b}(t) - \mathbf{c}(t) \tag{3.6}$$

If we are willing to make assumptions about $\mathbf{P}(t)$ and $\mathbf{a}(t)$ for future time period $t > N$, then a q–steps ahead forecasts starting from N can be obtained by using (3.3) iteratively for a point estimate (expected value) or the full forecasting distribution generated via simulation. Simply baseline models constructed using the observed transition matrices $\mathbf{P}(t)$.

3.2.1 Expected Transition

First we assume that the future rates will stay constant over time, and estimate $\mathbf{P}$ by taking the average over the $N-1$ number of historical observations (given N observed mass vector)

$$\hat{\mathbf{P}} = \mathbb{E}(\mathbf{P}) = (N-1)^{-1} \sum_{t=1}^{N-1} \mathbf{P}(t) \tag{3.7}$$

This will be refereed to as **DE**, the *expected transition matrix*, under a deterministic formulation where $\hat{\mathbf{P}}$ is fixed during the forecast.

3.2.2 Last Transition

There may also be reasons to believe that the most recent observed rate $\hat{\mathbf{P}} = \mathbf{P}(N-1)$, denote it as **DL**, will produce the best forecasting as it reflects the current state of the system. Unless $\mathbf{P}$ is constant over time, our forecast diverge by a factor of $\delta = \mathbb{E}(\mathbf{P}) - \mathbf{P}(N-1)$ at each step.

3.2.3 Contribution from External Sources

As previously mentioned in Chapter 2, we cannot distinguish between the users who have left the community and those who are just inactive. Therefore, we may want to model the *"inactive"* users as a role, or simply assume that all of them have left the community with some maybe rejoining later.

Our model makes no restriction on which roles the inactive users are allowed to return to, but instead allow collected data to naturally provide such information, using the observed transitions $\mathbf{v}_{0,j}$ (3.2).

In the latter case, the number of users leaving the compartments can be modeled by $\mathbf{c}(t)$ (3.6), but the non–negative constraint makes it a very difficult task. This is because not only do the current masses define the bounds $\mathbf{c}_j(t) \leq \mathbf{m}_j(t) \forall j$, which might be small ($< 5$) for some $\mathbf{m}_j$, they can also change (possibly towards zero) in the forecast. So we use the transition probability to model the number of users leaving where the non–negativity condition is automatically satisfied.

Now the joining vector $\mathbf{b}(t)$ consists of both the number of new users joining and the additional users who are returning from inactivity, $\mathbf{b}(t) = \mathbf{b}_{Join}(t) + \mathbf{b}_{Return}(t)$. Even though both $\mathbf{b}_{Join}$ and $\mathbf{b}_{Return}$ are vectors of length k, equal to the number of active compartments, it was observed in the data that users joined and returned only to some and not all the compartments. The frequency of any users joined or returned to certain compartments were also low. Therefore, we simplify the modeling process by modeling the sum of the vectors as follows:

$$y_{Join}(t) = \sum_{j=1}^{k} b_{j,Join}(t), \quad y_{Return}(t) = \sum_{j=1}^{k} b_{j,Return}(t).$$

Given prediction $\hat{y}_{Join}(t)$, the number of users joining each compartment can be found using a proportional vector $\boldsymbol{\gamma}$. Evidently, the resulting prediction for a particular compartment $\hat{b}_j = \gamma_j \hat{y}_j$ is not guaranteed to be an integer.

However, note that the observations $\mathbf{b}_{Join}(t)$ can be interpret as a realization from some stochastic process such as

$$\mathbf{b}_{Join}(t) \sim \mathcal{M}_k(y_{Join}(t), \gamma_{Join}(t)) \tag{3.8}$$

where $\mathcal{M}_k$ is a k dimension Multinomial distribution. Using this stochastic representation for both $\mathbf{b}_{Join}$ and $\mathbf{b}_{Return}$, we have integer predictions for each simulation. However, the expected forecasts are not integers even though the other statistics such as the confidence interval, mode, median of the forecasted distributions are.

These two different interpretations will be referred to as **I** and **W** respectively, and can be summarized as follows

Inactive role as a compartment (I)

For the case where inactive users belong to a compartment, our external contribution $\sum_j a_j(t)$ (3.6) is simply the number of new users joining at time t, $y_{Join}(t)$, and $\mathbf{P}(t)$ is a $[k \times k]$ matrix.

Without inactive compartment (W)

For the other setup, let 0 be the index that represents the inactive compartment such that $\mathbf{m}_0$ is the inactive compartment, then $a_j(t)$ consists of both the number of new users joining and the number of users returning from the inactive state to compartment j at time t, $\sum_j \mathbf{a}_j(t) = y_{Join}(t) + y_{Return}(t)$. The migration rate matrix $\mathbf{P}_{-0}(t)$ is now of dimension $[k-1 \times k]$, where the subscript -0 signifies the removal of the row indexed 0.

The second case is a complement of the first, where the inactive users are effectively unobserved because the forecast does not depend on the number of inactive users even though $\mathbf{m}_0$ exist in the forecast. To eliminate the inactive compartment completely, we can let our migration rate matrix be of dimension $[k-1 \times k-1]$ where $1 - \sum_j \mathbf{P}i, j$ contains the extra flow out of compartment i (those originally flowing out of the system). This means that the row sum of $\mathbf{P}$ is not necessary one, but the diagonals still have the same bound $0 \leq \mathbf{P}_{i,i} \leq 1$.

Predictions of y_{Join} and y_{Return} will be the discussed in detail in Chapter 4. For now, we assume that both y_{Join}, y_{Return} are observed even in the forecast and demonstrate a (potential) addition benefit when the inactive users are not modeled as a compartment.

3.2.4 Comparing Formulation of Inactive Users

We demonstrate the difference in prediction, in terms of MSE, between the formulation **I** and **W** described previously. Recall that the "inactive" users are defined to include both the users who have left the community and those not making active contribution that is observable in our data. Given that $\mathbf{P}$ is governed by $\mathbf{m}$ (3.4), we have $\mathbf{P}_{0,0}(t) \to 0$ as $\mathbf{m}_0(t) \to \infty$ if $v_{i,i}(t)$ is constant over time. Therefore, $\mathbb{E}(\mathbf{P})$ will not be a good estimate for the inactive compartment when serial correlation exists.

The sample correlation of the forums were investigated and it was found that $\mathbf{P}_{0,0}$ had significant autocorrelation at the first lag for all forums. An example can be seen in Figure B.1 where it shows the sample autocorrelation of the first 20 lag of

$\mathbf{P}_{0,0}$ for forum 353. Autocorrelation was also found in $\mathbf{P}_{0,j}$ for some j elements but only for a limited number of forums. Therefore, the $\mathbf{W}$ formulation was an attempt to eliminate the difficulty in modeling the autocorrelation of $\mathbf{P}_{0,0}$ or any of those that might also exist in $\mathbf{P}_{0,j}$.

Assuming that y_{Return} can be predicted accurately, it is possible to greatly increase the forecast performance by eliminating the inactive compartment. We demonstrate this by producing forecasts for a 10 week period using $\mathbb{E}(\mathbf{P})$ with the observed y_{Join}, y_{Return} over 90 consecutive weeks for forum 353 and report their MSE.

Apart from a few points (6 to be exact) in Figure 3.2, all the others lie below the diagonal line, the area that corresponds to a higher MSE for the $\mathbf{I}$ formulation. The average MSE over all 10 time and 10 roles is 40 and 27 showing a major improvement in the forecast by removing the inactive compartment. This finding is similar to other forums as well, i.e average MSE of 18 and 17 for forum 256, 75 and 51 for forum 264, under $\mathbf{I}$ and $\mathbf{W}$ respectively. Such results suggest that the quality of the forecasts may be improved by removing the inactive user compartment and predict y_{Return} separately. The two formulations, $\mathbf{W}$ and $\mathbf{I}$, will be compared throughout and especially in Chapter 5 when predictions of both y_{Join} and y_{Return} will be used in the forecasts.

The two formulations, $\mathbf{W}$ and $\mathbf{I}$, will be compared throughout this thesis. We assume that both y_{Join} and y_{Return} are observable in the forecasts in this chapter, i.e. all the prediction error only comes from the transition matrix which allows easier comparison between the modelling of $\mathbf{P}$, and predictions $\hat{y}_{Join}, \hat{y}_{Return}$ used in Chapter 5 to capture the total uncertainty in the forecasts.

3.3 Estimation of Transition Matrix

In addition to taking the expectation, the transition matrix $\mathbf{P}$ as well as the proportional vector γ can be estimated using just the observed mass. Let $\boldsymbol{\gamma}_{Join}$ be the proportional input vector of the form (3.8), then by using augmented matrix and vector

$$\mathbf{P}_{\dagger}(t) = \begin{bmatrix} \mathbf{P}(t) \\ \boldsymbol{\gamma}_{Join}^{\top}(t) \end{bmatrix}, \quad \mathbf{m}_{\dagger}(t)^{\top} = \begin{bmatrix} \mathbf{m}(t)^{\top} & y_{Join}(t) \end{bmatrix}, \tag{3.9}$$

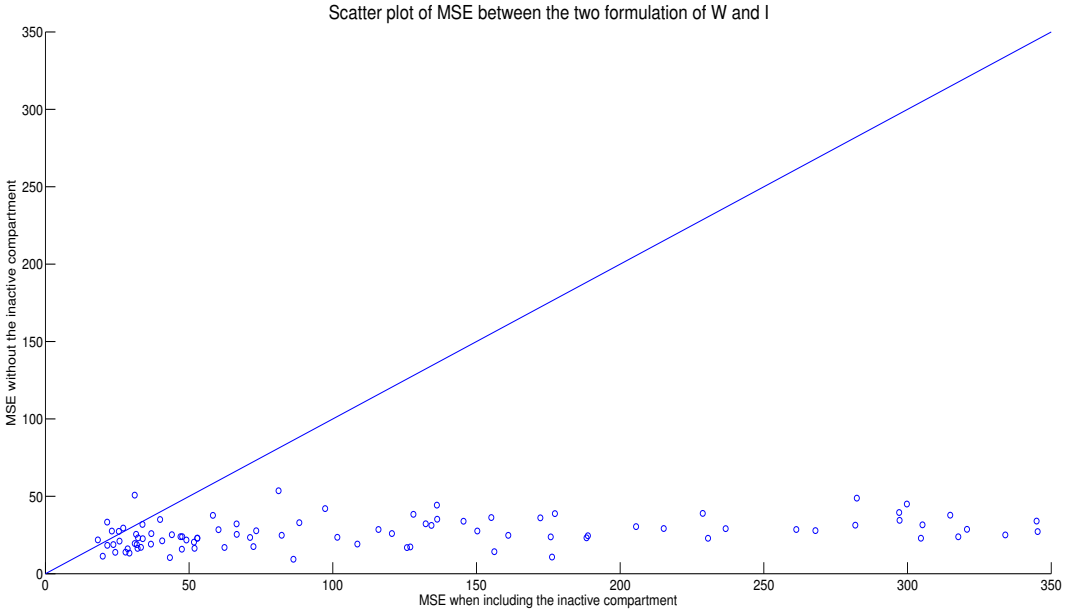


FIGURE 3.2: Scatter plot between the MSE obtained under the two different formulations over a 90 week period for forum 353.

where $y_{Join}(t)$ is the observed total number of users joining at time t, we can write (3.3) as

$$\mathbf{m}(t+1)^\top = \mathbf{m}_\dagger(t)^\top \mathbf{P}_\dagger(t) \tag{3.10}$$

The complementary case of $\mathbf{W}$ where the inactive users are effectively "unobserved" can be expressed as

$$\mathbf{P}_\dagger = \begin{bmatrix} \mathbf{P}_{-0} \\ \boldsymbol{\gamma}_{Join}^\top(t) \\ \boldsymbol{\gamma}_{Return}^\top(t) \end{bmatrix}, \quad \mathbf{m}_\dagger(t)^\top = \left[\mathbf{m}(t)_{-0}^\top \;\; y_{Join}(t) \;\; y_{Return}(t)\right]. \tag{3.11}$$

There are two different ways to formulate our problem for the estimation of $\mathbf{P}$. The first one is based on making a single step forecast at each stage, which is the same formulation as in a classical time series setting. While the other one takes a starting point, say the first observation, set it as an initial value and forecast using (3.3) through all observed time points.

The single step update can be interpret as a linear difference equation, or equivalently a vector time series. Even though parameter estimation in time series has simple expressions under certain formulation, such as vector autoregression (Hamilton, 1994, chap. 11), they are all based on the normality assumption which our data do not. Furthermore, we have to respect the constraints of a Markov matrix (3.5) and the additional constraints coming from the data (such as an

unobserved flow from state i to j). This result in a constrained optimization problem when estimating our parameters, the non–diagonal elements of $\mathbf{P}$, and is tackled in Section 3.3.1. Similarly, the same argument applies when estimating the parameters under an iterative update, which will be tackled in Section 3.3.2.

Although square loss is use almost exclusively in the following section, we have stated the problem under a generic loss function $L(\cdot, \cdot)$ whenever possible. This is because square loss can be interpreted as a maximum likelihood estimation under a Gaussian distribution and alternative distribution such as the Poisson can also be used. Our focus is place in comparing between the two formulations, linear and non–linear, as the resulting estimates are similar under different loss function. Furthermore, when the loss is measured using a distribution from the exponential family, the difficulty in the estimation is virtually identical as the objective function is still twice differentiable.

3.3.1 Single Step Update

First, we treat our parameter estimation problem as solving (3.3) with observations for both $\mathbf{m}(t+1)$ and $\mathbf{m}(t)$ at all time points. This is to say that the forecast is only for a single step

$$\hat{\mathbf{m}}(t+1)^\top = \mathbf{m}(t)^\top \mathbf{P} + \hat{\mathbf{b}}(t)^\top. \tag{3.12}$$

Estimating $\mathbf{P}_\dagger$ given the observed $\mathbf{m}(t), \hat{\mathbf{b}}(t)$ can be put in the form of a linear system $\mathbf{AX} = \mathbf{B}$ by combining the $N-1$ set of (3.10) where

$$\mathbf{A} = \begin{bmatrix} \mathbf{m}_\dagger(1)^\top \\ \mathbf{m}_\dagger(2)^\top \\ \vdots \\ \mathbf{m}_\dagger(N-1)^\top \end{bmatrix}, \quad \mathbf{X} = \mathbf{P}_\dagger, \quad \mathbf{B} = \begin{bmatrix} \mathbf{m}(2)^\top \\ \mathbf{m}(3)^\top \\ \vdots \\ \mathbf{m}(N)^\top \end{bmatrix}. \tag{3.13}$$

But the solution of (3.13) (say by least squares) is not guaranteed to satisfy the element constraints (3.5), and they need to be enforced. First, recognize that a linear system $\mathbf{AX} = \mathbf{B}$ can also be solved column–wise, which can be written as

$$\mathbf{C} = \mathbf{A} \otimes \mathbf{I}_k, \quad \mathbf{x} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_k \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \vdots \\ \mathbf{B}_k \end{bmatrix} \tag{3.14}$$

where $\mathbf{B}_i$ is the i^{th} row of the matrix $\mathbf{B}$, $\otimes$ is the Kronecker product and $\mathbf{I}_k$ is an identity matrix of k dimensions. Then our constrained optimization problem under some loss function $L(\cdot,\cdot)$ can be written in standard form

$$\begin{aligned} \arg\min_{\mathbf{x}} \quad & L(\mathbf{Cx}, \mathbf{d}) && (3.15)\\ \text{s.t.} \quad & \mathbf{Qx} = \mathbf{u} && (3.16)\\ & \mathbf{Rx} \leq \mathbf{v} && (3.17)\\ & \mathbf{lb} \leq \mathbf{x} \leq \mathbf{ub} && (3.18) \end{aligned}$$

with $\mathbf{C}$ and $\mathbf{d}$ previously defined. Given that $\mathbf{P}$ is a Markov matrix, it has to satisfy

$$\sum_j \mathbf{P}_{i,j} = 1, \quad 0 \leq \mathbf{P}_{i,j},$$

with the first contributing to (3.16) and the latter to (3.18). Note that the combination of the equality and non–negative constraints implies that $\mathbf{P}_{i,j} \leq 1 \forall i, j$. Similarly, the proportional vectors also has the same set of constraints

$$\sum \boldsymbol{\gamma}_i = 1, \quad 0 \leq \boldsymbol{\gamma}_i \; \forall \; i.$$

When the forecast of the inactive users are not of concern, then (3.15) becomes

$$\sum_{t=2}^{N} L(\hat{\mathbf{m}}_{-0}(t), \mathbf{m}_{-0}(t))$$

where estimated transition matrix has the column 0 removed and the equality is now an inequality $\sum_{j:j\neq 0} \mathbf{P}_{i,j} \leq 1$ to keep a nice expression with a linear objective function. Additionally, we assume that the observed transitions are the results of some true transition plus/minus some additional error, so the true transition probability should fall within the range of the ones already observed. Our restricted minimum becomes

$$\mathbf{P}_{i,j} \geq \max\left\{(1-\delta)\min\left\{\mathbf{P}_{i,j}(1), \mathbf{P}_{i,j}(2), \ldots, \mathbf{P}_{i,j}(N-1)\right\}, 0\right\} = \hat{lb}_{i,j} \qquad (3.19)$$

and maximum

$$\mathbf{P}_{i,j} \leq \min\left\{(1+\delta)\max\left\{\mathbf{P}_{i,j}(1), \mathbf{P}_{i,j}(2), \ldots, \mathbf{P}_{i,j}(N-1)\right\}, 1\right\} = \hat{ub}_{i,j}, \qquad (3.20)$$

both with a compensation factor δ to account for tail values that have yet to be observed. We use the term *"observed bounds"* when referring to equations (3.19)

and (3.20) while setting $\delta = 0.05$ for our estimation.

Using the observed bounds ensure that an unobserved edge between two compartments will result in an estimated transition probability of zero. An example of this can be seen in Appendix B.1 where the transition from state 2 to state 1 has never been observed, as shown by $\mathbb{E}(\mathbf{P})$ (Table B.1), but the estimation performed using the natural bounds (Table B.2) has $\hat{\mathbf{P}}_{2,1} = 0.018$. This is corrected when using the observed bounds (Table B.3) as both $\hat{lb}, \hat{ub}$ is equal to zero. Therefore, the observed bounds is preferred because it respects the data rather than the model definition and we denote this formulation as **PEL**. Finding the solution to **PEL** (3.15, 3.16, 3.17, 3.18) was done by using the MATLAB function **lsqlin** which required only a few seconds and less than 100 iterations.

3.3.2 Iterative Update

We can also use a long term forecasting interpretation where a $\mathbf{P}$ is used for a q steps forecast starting at some initial time point t_0. Let the initial value at time point t_0 be the observed mass

$$\hat{\mathbf{m}}(t_0) = \mathbf{m}(t_0)$$

and forecast using the update equation

$$\hat{\mathbf{m}}(t+1)^\top = \hat{\mathbf{m}}(t)^\top \mathbf{P} + \hat{\mathbf{b}}(t)^\top, \tag{3.21}$$

such that the forecast at time $t+1$ depends on the forecast at time t. Which is to say, given some initial conditions and total number of forecasting steps, we wish to find a $\hat{\mathbf{P}}$ that forecast/interpolate as closely to the historical observation as possible. We denote this formulation as **PENL**.

This can also be written in the form of (3.3) by augmenting the matrices like (3.9), i.e. for the $\mathbf{I}$ formulation, we replace the observed mass $\mathbf{m}(t)$ by the predicted mass $\hat{\mathbf{m}}(t)$ and let the predictions $\hat{\mathbf{b}}(t)$ be (3.8) where we wish to infer $\boldsymbol{\gamma}$ using actual observations $y_{Join}(t)$

$$\mathbf{m}_\dagger(t) = \left[\hat{\mathbf{m}}(t)\ ;\ y_{Join}(t)\right].$$

The augmented probability matrix $\mathbf{P}_\dagger$ is the same as the one in (3.9) but our objective function does not have a nice form as it is non–linear due to the iterative forecast process. The constraints are the same as previous formulation and our

optimization problem under the update (3.21) for q number of steps now become

$$\begin{aligned} \underset{\mathbf{P},\boldsymbol{\gamma}}{\arg\min} \quad & \sum_{t=1}^{q} L(\mathbf{m}(t_0+t), \hat{\mathbf{m}}(t_0)) \\ \text{s.t.} \quad & \sum_{j} \mathbf{P}_{i,j} = 1 \quad \forall\, i \\ & \hat{lb}_{i,j} \leq \mathbf{P}_{i,j} \leq \hat{ub}_{i,j} \quad \forall\, i,j. \end{aligned} \tag{3.22}$$

Any initial value/ starting point can be chosen, such as the start of the historical data at $t_0 = 1$ such that $q = N - 1$, or say only on the last 10 observations with $t_0 = N - q$ and $q = 10$.

The formulation (3.22) assumes that all the elements in $\mathbf{P}$ are variables that needs to be estimated. This is in fact not necessary because the diagonals are deterministic functions of the non–diagonals

$$\mathbf{P}_{i,i} := 1 - \sum_{j:j\neq i} \mathbf{P}_{i,j}.$$

We have an equivalent formulation by changing the equality to an inequality constraint

$$\sum_{j:i\neq j} \mathbf{P}_{i,j} \leq 1 \quad \forall\, i.$$

Both of these formulations were attempted using the MATLAB function **fmincon** using their in built interior–point algorithm and ran 50 times using the built–in **MultiStart** function. Majority ($\geq 80\%$) of the attempts converged to the same solution and the first formulation with equality constraint was found to be significantly faster.

3.3.3 Comparison of Update

A forecast of 99 weeks under the $\mathbf{I}$ formulation with initial value at the first observation $\mathbf{m}(1)$ using the observed y_{Join} at each time period can be seen in Figure B.4. The plot shows the in–sample forecast for two different $\hat{\mathbf{P}}_\dagger$, estimated under an absolute loss and square loss function over all roles, as well as the forecast under $\mathbb{E}(\mathbf{P}_\dagger)$.

Both sets of estimated $\hat{\mathbf{P}}_\dagger$ produced very similar forecasts, but the elements in $\hat{\mathbf{P}}_\dagger$ are very different apart from the proportional vector $\hat{\boldsymbol{\gamma}}$. Both of the estimated

transition matrix also look very different to $\mathbb{E}(\mathbf{P}_\dagger)$. A first look suggests that the $\hat{\mathbf{P}}$ under square loss is not aperiodic because $\mathbf{P}_{1,1} = 0$ and $\mathbf{P}_{1,2} = 1$, but in fact this is not true when looking at the eigenvalues. In fact, the stationary distribution has all the mass in the inactive role/state when under square loss, whereas it spreads over 3 different states when it is under the absolute loss estimation. Both sets of eigenvalues of $\hat{\mathbf{P}}$ contain elements which are complex, which is not the case for $\mathbb{E}(\mathbf{P})$.

It is also of interest to compare the two different formulations, in the linear setting **PEL** (3.12) and the non–linear update **PENL** (3.21). We would normally expect to see the estimated $\hat{\mathbf{P}}$ from the linear formulation to fluctuate less than estimate from the non–linear formulation. This is because, as the name suggests, the change is linear between two consecutive time steps in the former, i.e. does not have to account for change of direction, such as oscillation if it exists.

We check this hypothesis using our dataset. We estimated the transition matrix under both formulation using the first 100 observed mass with both under the square loss function to provide an appropriate comparison. We first look at a 99 steps forecast using the two different $\hat{\mathbf{P}}_\dagger$ with $t_0 = 1$ under $\mathbf{I}$ formulation. The forecast can be seen in Figure 3.3 which is also a replicate of the interpolation in the estimation for the **PENL** case. There are significant differences between the forecast where **PENL** fluctuates a lot more than **PEL** in all the compartments. While both estimation out–performed $\mathbb{E}(\mathbf{P})$, **PENL** had better performance as shown in Figure B.5, where **PENL** dominates **PEL** in error for more than three quarters of in–sample forecast. Out–of–sample forecast performance are compared in Section 3.6.

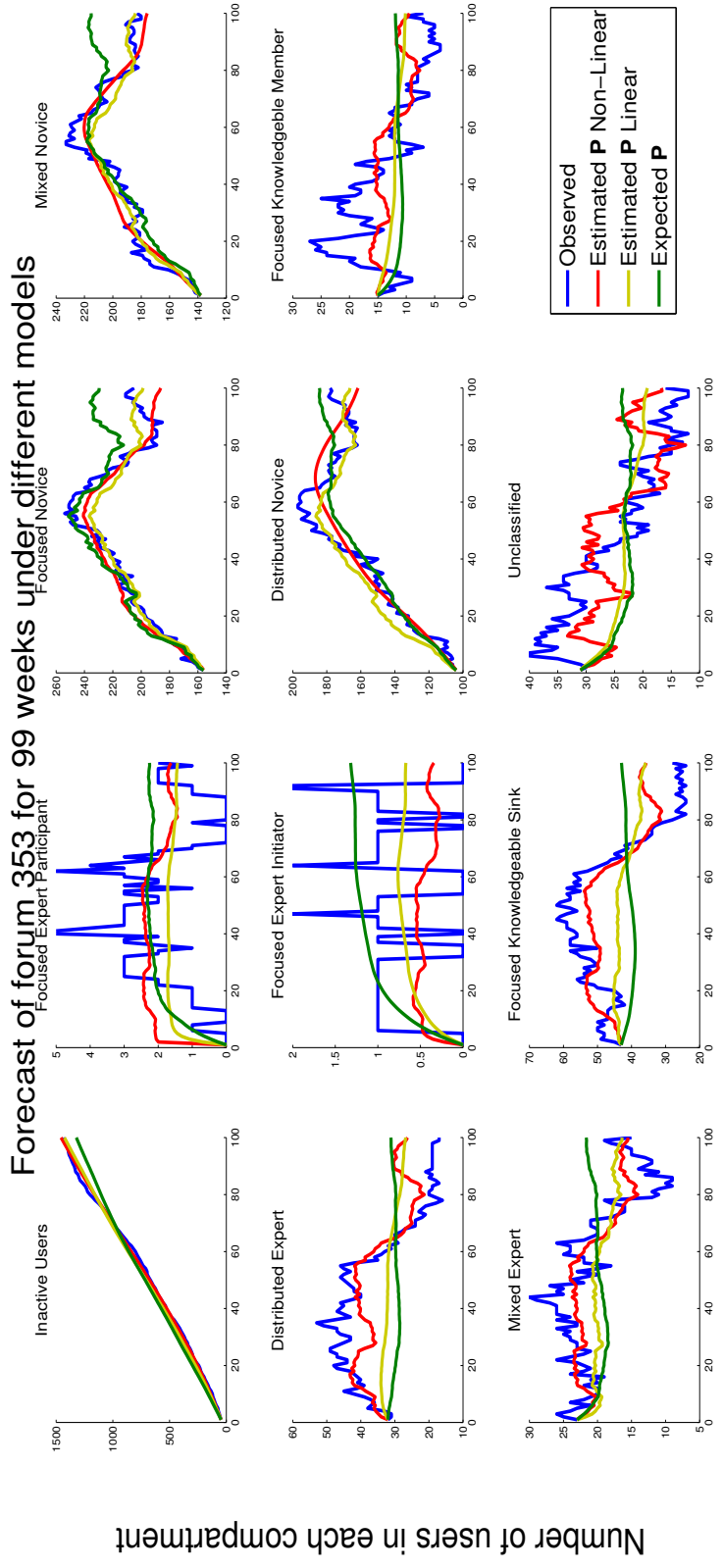


FIGURE 3.3: Forecast using $\mathbf{P}_\dagger$ estimated using the linear single step update and the non–linear iterative update under the **I** formulation. Forecasted for a 99 weeks with initial value as the first observation and using the observed number of users at each time period

3.4 Combination of Transition Matrices

The methods mentioned above aim to find all the elements of $\mathbf{P}_\dagger$ using the observed mass as well as the historical $\mathbf{P}_\dagger(t)$ to obtain the lower and upper bounds of each element, which is a stark contrast to $\mathbb{E}(\mathbf{P}_\dagger)$ where the estimation ignores the observed mass completely. A compromise between the two would be to estimate $\mathbf{P}_\dagger$ using both the observed mass and transition matrix, which has the potential to generate better forecasts as it tries to use both sets of information. Let the $N-1$ observed $\mathbf{P}_\dagger$ be

$$\mathbf{P}_\dagger \sim \hat{F}_{N-1}(\mathbf{w}) \tag{3.23}$$

where $\hat{F}_{N-1}$ is the empirical distribution and $\mathbf{w} = (w_1, w_2, \ldots, w_{N-1})$ the corresponding weight vector for each of the observations. Now, we describe two methods that uses the historical observed transition matrices.

3.4.1 Weighted Mixture of Matrix

A simple extension to (3.7) is to find a weighted average of (3.23)

$$\hat{\mathbf{P}}_\dagger = \sum_{t=1}^{N-1} w_t \mathbf{P}_\dagger(t), \quad \sum_{t=1}^{N-1} w_t = 1,\ w_t \geq 0, \tag{3.24}$$

we denote this weighted approach as **DW**. Taking the sample average of $\mathbf{P}_\dagger$ is a special case using uniform weights $w_t = (N-1)^{-1}, \forall t$. The constraints (3.24) automatically ensures that (3.5) are satisfied.

Both of the formulations mentioned in Section 3.3, where the forecasts are (3.12) or (3.21), can be used. For the linear case, the objective function is

$$\arg\min_{\mathbf{w}} \sum_{t=t_0}^{N-1} \left\| \mathbf{m}(t+1) - \sum_{i=1}^{N-1} w_i \mathbf{P}_\dagger^\top(i) \mathbf{m}_\dagger(t) \right\|^2, \tag{3.25}$$

which can also be written in the constrained least squares form

$$\begin{aligned} &\arg\min_{\mathbf{w}} \|\mathbf{A}\mathbf{w} - \mathbf{b}\|^2 \\ &\text{s.t.} \sum_{i=1}^{N-1} w_i = 1 \\ &w_i \geq 0\ \forall i, \end{aligned} \tag{3.26}$$

by computing $\hat{\mathbf{m}}^i(t+1) = \mathbf{P}_\dagger^\top(i)\mathbf{m}_\dagger(t)$, and let

$$\mathbf{A} = \begin{bmatrix} \hat{\mathbf{m}}^1(t_0+1) & \hat{\mathbf{m}}^2(t_0+1) & \cdots & \hat{\mathbf{m}}^{N-1}(N) \\ \hat{\mathbf{m}}^1(t_0+2) & \hat{\mathbf{m}}^2(t_0+2) & \cdots & \hat{\mathbf{m}}^{N-1}(N) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{m}}^1(N) & \hat{\mathbf{m}}^2(N) & \cdots & \hat{\mathbf{m}}^{N-1}(N) \end{bmatrix}$$

with the corresponding response as

$$\mathbf{b} = \left[\mathbf{m}^\top(t_0+1)\ \mathbf{m}^\top(t_0+2)\ \cdots\ \mathbf{m}^\top(N)\right]^\top.$$

When $\mathbf{A}$ is full rank, (3.26) is strictly convex. The vector $\mathbf{w}$ is of length $N-1$, the same number of observed $\mathbf{P}$. Dimension of the vector $\hat{\mathbf{m}}^i(t)$ depends on the number of roles we are interested in, which in our case is 10 excluding the inactive role. In order for $\mathbf{A}$ to have more rows than columns, the number of steps $N - t_0$ needs to be greater than $(N-1)/10$. If there are identical transitions matrices, i.e. $\mathbf{P}_\dagger(i) = \mathbf{P}_\dagger(j)$ for some i, j then the number of variables as well as columns in $\mathbf{A}$ decreases by the number of repeated $\mathbf{P}_\dagger(t)$.

3.4.2 Exponential Penalty

Another weighted approach using a time discount factor similar to exponential smoothing in the time series literature (Holt, 2004), denoted as **DP**. This method aims to find a $\hat{\mathbf{P}}_\dagger$ between **DE** and **DL** that has the optimal amount of local trend. We introduce a time discount factor (penalty) $\alpha \in [0, 1]$ that controls the contribution of $\mathbf{P}_\dagger(t)$ based on how far back the observed matrix is from the last observation at time t_0. Denote by $\hat{\mathbf{P}}_\dagger(\alpha)$ the expected value of (3.23) under penalty α,

$$\hat{\mathbf{P}}_\dagger(\alpha) = \sum_{j=0}^{N-2} w_j \mathbf{P}_\dagger(N-1-j), \quad w_j \propto \begin{cases} 1 & \text{if } j = 0 \\ \alpha^j & \text{if } j = 1, 2, \ldots N-2. \end{cases} \tag{3.27}$$

such that α_j follows a finite geometric series with the normalized weight w_j, $\sum w_j = 1$. This weighted approach can be understood as the following three

settings;

$$\alpha = 0 \;\Rightarrow w_j \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.28}$$

$$\alpha \in (0,1) \;\Rightarrow w_j = \frac{1-\alpha}{1-\alpha^{N-1}}\alpha^j, \quad j = 0, 1, \ldots, N-2 \tag{3.29}$$

$$\alpha = 1 \;\Rightarrow w_j = 1/T, \quad j = 0, 1, \ldots, N-2 \tag{3.30}$$

where (3.28) and (3.30) represent the previous two deterministic method, namely **DL** and **DE** respectively. As this penalized approach also contains the two baseline models, it provides a hint on expected behavior based on the recent history (subject to the number of in–sample steps used) where $\hat{\alpha}$, the estimated penalty is an indication of the strength of the local trend. Assuming that the future behavior is similar to the current trend, our tuning should guarantee the performance to be at least as good as the two baseline models of **DE** and **DL**. The optimization problem under linear formulation is

$$\begin{aligned} \arg\min_{\alpha} \quad & \sum_{t_0=1}^{N-1} \left\| \mathbf{m}(t_0+1) - \frac{1-\alpha}{1-\alpha^{N-1}} \sum_{j=0}^{N-2} \alpha^j \mathbf{P}^\top (N-1-j)\mathbf{m}(t_0) \right\|^2 \\ \text{s.t.} \quad & 0 \leq \alpha \leq 1 \end{aligned} \tag{3.31}$$

and multiple minima may exist as demonstrated in Figure 3.4. Similarly, the same can be found for the non–linear formulation (Figure B.2). Although the objective function is not guaranteed to be convex, the optimal value $\hat{\alpha}^{optim}$ ($\hat{\alpha}$ from herein) is bounded and is easily found by evaluating a set of equally spaced points over the interval $[0, 1]$.

Given $\hat{\alpha}$, this value determines the relative contribution of $\mathbf{P}_\dagger(t)$ and there is a decision on how to use the observed data. More specifically, the number of $\mathbf{P}_\dagger$ to use in the estimation and forecast stage when $t_0 \neq 1$. The objective function is to minimize the MSE of the observed masses from $\mathbf{m}(t_0)$ to $\mathbf{m}(N)$ and obtain $\hat{\alpha}$, where the following three scenarios arise;

Scenario 1 (S1)

Estimate using the observed $\mathbf{P}_\dagger(t)$ from 1 to t_0 and forecast with $t = 1, \ldots, t_0$.

Scenario 2 (S2)

Estimate using the observed $\mathbf{P}_\dagger(t)$ from 1 to t_0 and forecast with $t = 1, \ldots, N-1$.

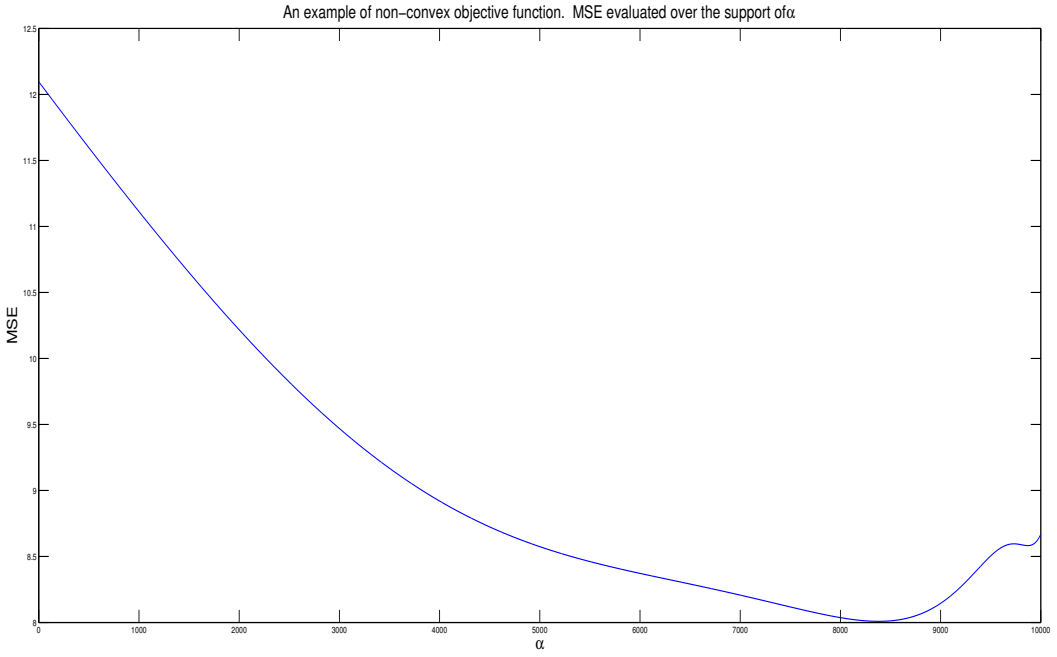


FIGURE 3.4: An example of two minima for the penalized transition matrix obtained in forum 264 under **I** at week $N = 94$, evaluated over 1000 equally spaced point over the interval $[0, 1]$

Scenario 3 (S3)

Estimate and forecast using observations of $\mathbf{P}_\dagger(t)$ for $t = 1, \ldots, N-1$.

Two of these scenarios, **S1** and **S3**, are also possible for the weighted vector **w**. But **S2**, which has the same $\hat{\alpha}$ as **S1**, exists because the penalty α applies regardless of the total number of observations where the vector **w** is of a fixed length. In both **S1** and **S2**, there was no overlap of information, i.e. the error were measured on $\mathbf{m}(t_0+1)$ to $\mathbf{m}(N)$ whereas $\hat{\alpha}$ was estimated based on observations from $\mathbf{P}_\dagger(1)$ to $\mathbf{P}_\dagger(t_0)$.

We demonstrate the differences among the three choices using the same method as those in Section 3.2.4 under a non–linear formulation. Estimation and forecast for a 10 week period were performed over 90 consecutive weeks. Both the **W** and **I** formulation were tested using the observed y_{Join} and y_{Return} for both the estimating and forecasting stage. For Scenario 1 and 2, a suitable end point, $t_0 = N - 10$, was chosen to be the same period as the desired forecast as an attempt to mimic the predictions within the historical observations.

The scatter plot between the set of $\hat{\alpha}$ between **S1** and **S3** for forum 353 can be seen in Figure 3.5 with the diagonal line representing $x = y$, i.e. $\hat{\alpha}$ is equal under both scenarios. The plot shows that $\hat{\alpha}$ can differ significantly between **S1** and **S3**

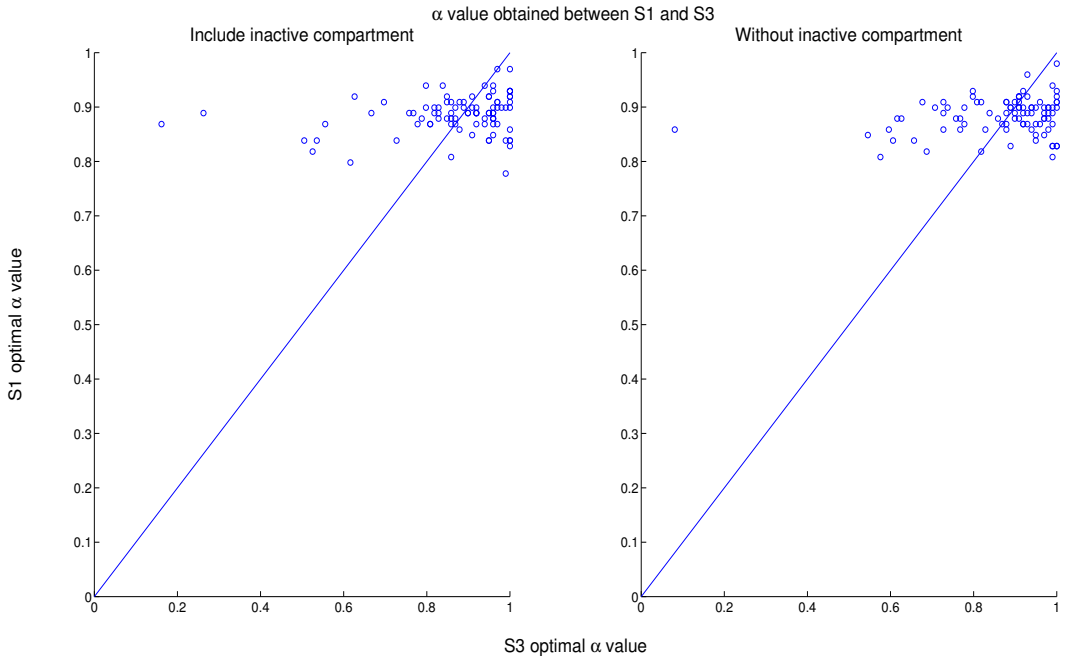


FIGURE 3.5: The $\hat{\alpha}$ value obtained between **S1** and **S3** under the two different formulations. Over 90 consecutive weeks for forum 353 using the observed y_{Join} and y_{Return} and $t_0 = N - 10$

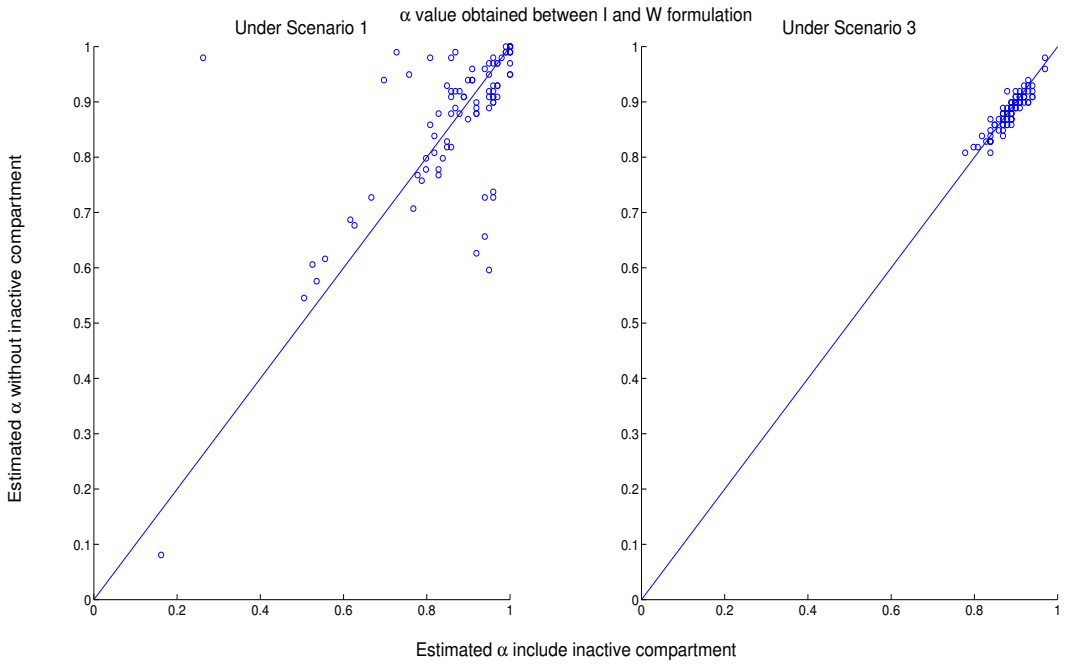


FIGURE 3.6: The $\hat{\alpha}$ value obtained between **I** and **W** under the two different scenarios. Over 90 consecutive weeks for forum 353 using the observed y_{Join} and y_{Return} and $t_0 = N - 10$

	I			**W**		
	S1	**S2**	**S3**	**S1**	**S2**	**S3**
Average MSE	91	115	90	29	31	25
Number of lowest MSE	35	9	47	28	13	50

TABLE 3.1: The average error and the number of times a scenario has achieved the lowest error over a 90 week period for forum 353

with virtually no correlation. This is true whether the inactive compartment was included or not, demonstrating the number of observations affects $\hat{\alpha}$ significantly. It is further demonstrated by Figure 3.6, especially in the case of **S3**. Where the two sets of $\hat{\alpha}$ under **W** and **I** have a strong positive correlation, all the points only deviate slightly from the diagonal line. Although the majority of the points in **S1** also fall near the diagonal line, a number of $\hat{\alpha}$ differs by a large margin between the **I** and **W** formulation.

Table 3.1 shows the MSE as well as the number of times a scenario achieved the lowest MSE over the 90 separate forecasts. Clearly, **S3** produced the best result out of the three, and Table 3.1 demonstrates the superior performance when not using the inactive compartment as the average MSE is lower for all three scenarios. The performance of **S2** suffered while using more observations than **S1**, but this does not generalize to other forums as well (Table B.4, B.5, B.6) with **S2** beating **S1** dependent on both the forum and the formulation for the inactive users. Nevertheless, **S3** dominates both **S1** and **S2** consistently.

These results suggest that **S3** should be used, because it has the best performance in addition to the natural usage of the available data – all observed $\mathbf{P}_{\dagger}(t)$ from 1 to $N-1$ are available for both the estimation of $\hat{\alpha}$ and forecasting. Whereas **S1** and **S2** only use $\mathbf{P}_{\dagger}(t)$ up to $N-1$ and ignore the most recent observations, those exact observations we think are more important when forecasting and attempt to place more weights on them through the use of $\hat{\alpha}$.

3.5 Stochastic Transition Matrix

Under the previous assumption that the observed classification of users $\mathbf{z}(t)$ is correct at all time t, the difference in transition probability $\mathbf{P}(t)$ between time points implies some form of stochasticity. A common assumption used on compartmental

models is a Gaussian error (Jacquez, 1972)

$$\mathbf{P}_{i,j}(t) \sim \mathcal{N}(\mathbf{P}_{i,j}, \sigma^2_{i,j}),\ \forall\ i, j \text{ and } i \neq j. \tag{3.32}$$

When the underlying system is actually stochastic, it is important to correctly model this, as a deterministic formulation can yield incorrect results (Soong and Dowdee, 1974). Matis and Wehrly (1979) demonstrated this further using the Jensen's inequality

$$h(\mathbb{E}(X)) \leq \mathbb{E}(h(X)) \tag{3.33}$$

where $h(\cdot)$ is a convex function, such as the recursive use of the Markov chain (3.3) to obtain our forecast. Therefore, the *expected forecast* is the sample average of the simulated forecasts, obtained using (3.3) where $\mathbf{P}$ at each of the forecasting steps is a realization from some generating process. The output of the expected forecast computed via simulation is different to one obtained using $\mathbf{P}$ as evident from (3.33).

3.5.1 Parametric

The Gaussian assumption is satisfied only when the observations do appear normally distributed, but the bounds imposed by (3.5) makes the normal distribution a poor choice. This is supported by the data where many zeros can be observed for some $\mathbf{P}_{i,j}$ and $\mathbb{E}(\mathbf{P}_{i,j})$ is close to zero or when $\text{Var}(\mathbf{P}_{i,j})$ is large. Therefore, a truncated univariate normal (TUVN)

$$X \sim \mathcal{N}(\mu, \sigma^2, \mu^-, \mu^+) \tag{3.34}$$

is a better representation where the support of the distribution is defined by the lower and upper bound μ^-, μ^+. Furthermore, the Gaussian assumption allows easy extension into the multivariate form that allows dependency between the random variables.

3.5.1.1 Truncated Multivariate Normal

Let $\mathbf{P}_{-i}$ represent the vector of the i^{th} row without the i^{th} column of the matrix $\mathbf{P}$ and $\boldsymbol{X} = \mathbf{P}_{-i}$ is our random variable. The correlation structure of all the rates going out of state i can be modeled by a Truncated Multivariate Normal (TMVN)

(3.35) of d dimension, where the realization $\mathbf{x}$ is subject to the linear inequality $\mathbf{lb} \leq \mathbf{Rx} \leq \mathbf{ub}$. The linear inequality extends the box constraints of (3.34) as $\mathbf{lb}$ is a p dimension vector and $\mathbf{R}$ is a $[p \times d]$ matrix. It allows us to enforce the vector constraint as well as the element wise constraints in (3.5).

$$\boldsymbol{X} \sim \mathcal{N}_d(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}; \mathbf{R}, \mathbf{lb}, \mathbf{ub}) \tag{3.35}$$

The inequalities are defined by $\mathbf{R} = [\mathbf{I}_{k-1} \; ; \; \mathbf{e}_{k-1}]^\top$, $\mathbf{lb} = [\mathbf{e}_{k-1} \; ; \; 0]$ and $\mathbf{ub} = [\mathbf{e}_{k-1} \; ; \; 1]$, where $\mathbf{e}_c$ is a c length column vector of 1's. Even though efficient sampling procedures have been developed, see Yu and Tian (2011) and references therein, parameters in (3.35) cannot be estimated in closed form due to the difficulty induced by the linear equality when evaluating the normalizing constant $\int_{\mathbf{lb} \leq \mathbf{Rx} \leq \mathbf{ub}} \phi(x)dx$. The estimation problem is exaggerated for our dataset because there are forums with ≈ 50 data points over 10 dimensions (one dimension for a role).

Another formulation which simplifies the estimation is to assume that only the diagonals of $\mathbf{P}$ follows a TMVN with only element wise constraints $0 \leq \text{diag}(\mathbf{P}) \leq 1$, such that we only model the total flow out of the compartments, and the proportions for each of the flow. Modeling the proportions can be done via a Dirichlet distribution or alternative methods such as sum of log–normals (Gelman et al., 1996) or a Logistic-normal (Aitchison and Shen, 1980). We use the Dirichlet assumption for simplicity and the ease of ML estimation. The stochastic representation is then

$$(1 - \mathbf{P}_{i,i}^{-1}(t))\mathbf{P}_{-i}(t) \sim \mathcal{D}(k-1, \boldsymbol{\alpha}),$$

such that the generating process for the non–diagonals realizations are

$$\text{diag}(\mathbf{P}) \sim \mathcal{N}_k(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}; \boldsymbol{\mu}^-, \boldsymbol{\mu}^+) \tag{3.36}$$

$$\mathbf{c}_i \sim \mathcal{D}(k-1, \hat{\boldsymbol{\alpha}}_i) \tag{3.37}$$

$$\mathbf{P}_{-i} = (1 - \mathbf{P}_{i,i}) \times \mathbf{c}_i \;\; \text{for } i = (1, \ldots, k). \tag{3.38}$$

This sacrifices the ability to model the correlation between migration rates going out of a compartment, but allows modeling of the correlation on the total outgoing rate between the compartments. Again, the difficulty of estimating the parameters of TMVN remains even though the linear inequality does not exist due to the curse of dimensionality. Here, we propose to simplify the multivariate distribution into their univariate marginals TUVN where their correlation structure is determined

by a copula such that the number of parameters (without closed form solution) reduces from $O(k^2)$ to $O(k)$.

3.5.1.2 Truncated Univariate Normal With Copula

Given a vector of random variables $\boldsymbol{X} = (X_1, \ldots, X_d)$ with joint distribution F and the marginal of variable i as $F_i(x_i) = \Pr(X_i \leq x_i)$. Then a copula C is a function that maps $\mathbf{I}_d$ to $\mathbf{I}$ where the copula of F is one that satisfies

$$F(\mathbf{x}) = C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)). \tag{3.39}$$

In essence, a copula is a function that constructs the dependence between the marginals using the fact that all probability distribution satisfy $F_i(X_i) = U_i \sim \mathcal{U}[0,1]\ \forall i$, and the copula is unique when all F_i are continuous. This means that given any set of marginals, we have a joint distribution with the correlation properties of the chosen copula. A more comprehensive overview of copula can be found in Nelson (1999).

The Gaussian copula represents the case when the correlation structure is being induced by a normal distribution and it can be written as

$$C_{Ga}(\boldsymbol{u}) = \Phi_{\boldsymbol{\Omega}}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_d)), \tag{3.40}$$

where Φ is the c.d.f. of the standard normal and $\Phi_{\boldsymbol{\Omega}}$ is the c.d.f. of a multivariate normal with correlation matrix $\boldsymbol{\Omega}$ and mean 0. This is a popular choice due the simplicity of sampling from a multivariate normal even in high dimension. Substitute the LHS of (3.40) with the RHS of (3.39)

$$F(\mathbf{x}) = \Phi_{\boldsymbol{\Omega}}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_d))$$

then recognizing that $Z \sim \mathcal{N}(0, \boldsymbol{\Omega}) = \Phi_{\boldsymbol{\Omega}}^{-1}(U)$ with $U \sim \mathcal{U}[0,1]$ and $F(\mathbf{x}) = u$, we have

$$\boldsymbol{z} = (\Phi^{-1}(F_1(x_1)), \Phi^{-1}(F_d(x_2)), \ldots, \Phi^{-1}(F_d(x_d))),$$

which implies that the random vector $\boldsymbol{X}$ from our joint distribution is a transformation on the realization of $\boldsymbol{Z}$:

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_d) = (F_1^{-1}(\Phi(Z_1)), F_2^{-1}(\Phi(Z_2)), \ldots, F_1^{-1}(\Phi(Z_d)))$$

This is also known as NORTA (NORmal To Anything) (Cario and Nelson, 1997) and it only requires the knowledge of the marginals and it's inverse F_i^{-1} in addition to the correlations between pairs of random variables. The common strategy is to obtain the Spearman rank correlation of $\boldsymbol{X}$ and construct an appropriate $\boldsymbol{\Omega}$ for (3.40) (Cario and Nelson, 1997; Ghosh and Henderson, 2003; Avramidis et al., 2009; Channouf and L'Ecuyer, 2009) depending on the type of marginals. Ghosh and Henderson (2003) gave an example where rank correlation was superior to product–moment correlation, further justifications and examples can be found in Embrechts et al. (2002) and therein.

Let the Spearman rank correlation matrix be ρ_S and the element $\rho_S(i,j)$ be the correlation between two variables X_i, X_j, an exact transformation (3.41) was derived by Kruskal (1958) between rank and product–moment correlation when $F(X)$ is continuous such that $\boldsymbol{\Omega}$ (3.40) yields the desired correlation for $\boldsymbol{X}$:

$$\boldsymbol{\Omega}_{i,j} = 2\sin\left(\frac{\rho_S(i,j)}{6}\right). \tag{3.41}$$

But $\hat{\boldsymbol{\Omega}}$ obtained directly using (3.41) is not guaranteed to be positive semi–definite (Ghosh and Henderson, 2003) for a feasible correlation matrix even when all the marginals are continuous (Ghosh and Henderson, 2002). So an approximation is required by minimizing $d(\hat{\boldsymbol{\Omega}}, \boldsymbol{\Omega})$ where $d(\cdot)$ is some distance function.

Replace TMVN with the copula formulation for diag($\mathbf{P}$), parameters estimation can be performed univariately (Cohen, 1950; Halperin, 1952) and obtaining the Spearman rank correlation is trivial. Inversion of TUVN is the same as a standard normal with an adjusted normalizing constant, which is approximated with a high precision on modern computers. But when the estimated parameters correspond to a scenario where the majority of the density of a p.d.f. is outside the boundaries $[0,1]$ of (3.5), the inversion provides a poor representation of the distribution and a sampled based inversion should be used by generating J sample from $X \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2; \mu^-, \mu^+)$ (using for example, an algorithm by Damien and Walker (2001)) and finding the Ju^{th} value of the sorted realizations for the marginal given uniform realization $\boldsymbol{u}$ from the copula.

3.5.1.3 Binomial Formulation

A Markov chain interpretation also implies that the model is stochastic by nature. Simulating a discrete time Markov chain given the estimated transition matrix $\hat{\mathbf{P}}$

can be done on a state to state basis, by first finding the number of people leaving a state through a binomial distribution

$$w_i \sim \mathcal{B}(m_i, 1 - \hat{\mathbf{P}}_{i,i}) \tag{3.42}$$

then allocate them into its possible destination using a multinomial distribution

$$\mathbf{v}_{-i} \sim \mathcal{M}_{k-1}(w_i, \hat{\mathbf{P}}_{-i}), \tag{3.43}$$

where $\hat{\mathbf{P}}_{-i}$ represents the i^{th} row without the i^{th} element of $\hat{\mathbf{P}}$ and $\mathbf{v}_{-i}$ is the flow vector out of compartment i.

Forecasts generated using this formulation results in non–negative integers for all roles and all forecasting time steps. This has a more natural interpretation when compared to the forecasts generated using the truncated normal distribution (non–negative reals). As mentioned previously in Section 3.2.3, the expected forecasts are not integers while other common statistics (of a distribution) are.

3.5.2 Mixture of Historical Transition Matrices

The binomial formulation (Section 3.5.1.3) is a parametric model that uses the same $\hat{\mathbf{P}}_{i,i}$ across all time. The hypothesis that transition probabilities are constant or not can be verified using the contingency table test by Anderson and Goodman (1957). Applying it to our data reveals that most of the elements in $\mathbf{P}$ are not constant.

We extend the two weighted methods in Section 3.4 and introduce stochasticity via the corresponding weighted empirical distribution. Given a weight vector $\mathbf{w}$, a sample from the empirical distribution can be expressed as a mixture with the following stochastic representation

$$\mathbf{P}^* = \sum_{t=1}^{N-1} \mathbf{z}_t \mathbf{P}_\dagger(t), \quad \boldsymbol{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N-1}), \quad \boldsymbol{Z} \sim \mathcal{M}_{N-1}(1, \mathbf{w}). \tag{3.44}$$

Estimation under the linear formulation (Section 3.3.1) for both the weight vector $\hat{\mathbf{w}}$ and penalty $\hat{\alpha}$ are the same as the deterministic version (Section 3.4). This is because we only need the expectation for a single step.

For the non–linear formulation (Section 3.3.2), the MSE curve over α can also be multimodal for the stochastic version like the deterministic setup (Figure B.2).

Given that there is only a single parameter, bounded by 0 and 1, it is feasible to evaluate and obtain $\hat{\alpha}$ using say 100 equally spaced points, we denote this method as **SP**. The forecasts generated using the expected transition matrix given $\hat{\alpha}$ can be significantly different from the expected forecast obtained via Monte Carlo as explained previously. Figure B.3 shows the difference in MSE over α between the two, where the curves become similar as $\alpha \to 0$ and is identical when $\alpha = 0$ as both of them reduces to **DL**.

For the weights of (3.24), denoted as **SW**, we employ a direct search with a generating set search algorithm (Lewis et al., 2007) using the MATLAB function **patternsearch**. The initial value is obtained by finding the set of weights that gives the lowest objective value out of a set of samples, which includes the vertices as well as 100 random realizations sampled uniformly from the $N-2$ simplex (Rubin, 1981). Both the forecasts used in penalized and weighted estimation are generated by taking the sample average of the Monte Carlo simulated forecast, of 10^4 iterations, which approximates the expected value of the forecast. An approximation is used here because computing all the realization requires too much time.

Assuming that the $N-1$ observed $\mathbf{P}_{\dagger}(t)$ are all unique, then there exists $(N-1)^q$ possible realization at the q^{th} forecasting step. Although the empirical distribution approach is time consuming in the estimation, it is significantly faster than the Binomial formulation when generating the confidence interval for the forecast. This is because the Binomial formulation requires a sample from both (3.42) and (3.43) for each of the $k-1$ roles at each time step. On the other hand, the empirical approach only need a $\mathbf{P}_{\dagger}(s)$, where s is a random integer from $\{1, 2, \ldots, N-1\}$.

3.6 Results

We compare all the models mentioned previously, the estimated matrices **PEL** and **PENL** in Section 3.3 as well as the penalty **DP** and weighed approach **DW** in Section 3.4. The results of both the **I** and **W** formulation are investigated, using the actual observations of y_{Join}, y_{Return} for the forecast. All of these methods can use either all N observation $t_0 = 1, q = N-1$ or the last 10 observations with $t_0 = N-10, q = 10$ in estimation.

We have also considered the idea of using $t_0 = N-10$ for **DW** even though the solution of (3.26) may not be unique due to rank deficiency. When this happens,

Method	**W**	**I**
DL	149	149
DE	26	40

(A) Forum 353

Method	**W**	**I**
DL	90	90
DE	17	18

(B) Forum 256

Method	**W**	**I**
DL	218	227
DE	51	75

(C) Forum 264

		Forum 353		Forum 256		Forum 264	
		t_0		t_0		t_0	
Method	Formulation	1	$N-10$	1	$N-10$	1	$N-10$
PEL	**W**	49	34	27	20	93	65
	I	32	117	21	19	122	68
PENL	**W**	49	39	45	22	80	80
	I	46	35	36	21	113	72
DPL	**W**	27	25	18	18	51	48
	I	35	26	15	17	58	52
DPNL	**W**	27	25	16	17	51	47
	I	37	26	16	16	63	53
DWL	**W**	25	27	18	19	49	50
	I	29	30	18	19	54	56
DWNL	**W**	25	28	21	20	52	56
	I	30	29	24	19	57	61

(D) MSE under different estimated methods

TABLE 3.2: Summary of error over 90 consecutive weeks for all methods, each with a 10 week ahead forecast for three different forums. Both formulations of the inactive users were used with initial observation in the estimation either at $\mathbf{m}(1)$ or $\mathbf{m}(N-10)$, where N is the total number of observed mass. Numbers highlighted in green is the lowest MSE achieved between the methods under the same number of observations.

we remove one of the repeated columns through a QR decomposition. The deterministic models (Section 3.3) will be examined first before moving on to the stochastic models (Section 3.5).

3.6.1 Deterministic Models

The number of observations N changes as time progresses, i.e. for the 90 predictions in consecutive week, N ranges from 20 to 110. The two baseline models, **DE** and **DL**, only have one variant where all the observations were used. Summary results in terms of MSE can be seen in Table 3.2, which are averages over the 90 consecutive weeks where a 10 steps ahead forecast was made at each of the 90 weeks. The baselines of **DE** and **DL** for forum 353, 256 and 264 are in Table 3.2a, 3.2b, 3.2c respectively, with the forecasts in Table 3.2d for all three. As seen in Table 3.2, all the proposed methods were either similar or worse than **DE** in the

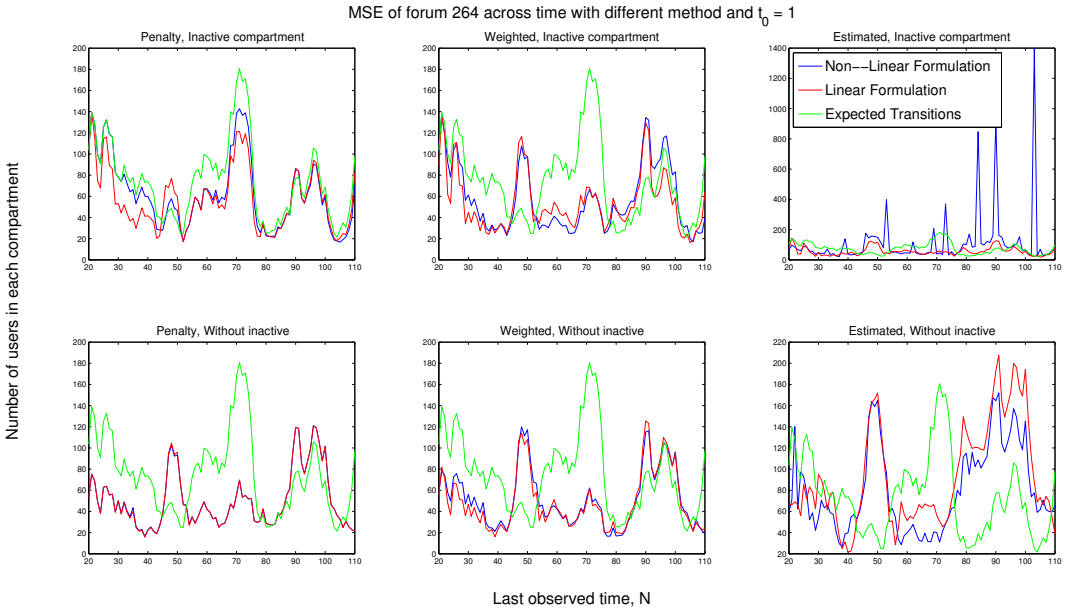


FIGURE 3.7: MSE of different methods across time for forum 264.

W setting. Strictly speaking, comparison between **I** and **W** is unfair as the latter depends on y_{Return}, which will be unobserved during the forecast. This is an issue which will be revisited in Chapter 4 and Chapter 5.

On the other hand, most of the methods performed better than **DE** under the **I** formulation and even outperformed the **W** formulation, contrary to **DE**. In particular, **DP** and **DW** generally has better performance when $t_0 = N - 10$. This is because a community goes through changes as time progresses and using the full observations ignores the local dynamic. When the full set of observations were used for **DP**, the set of $\hat{\alpha}$'s obtained (in both **I** and **W**) were all very close or exactly 1 most of the time under the non–linear formulation. Hence, it failed to exploit the recent trend and ended up with an almost identical forecast to **DE**.

Furthermore, $\hat{\alpha}$ under the linear and non–linear formulation when $t_0 = 1$ was almost identical across all the 90 weeks attempted. This is due to the restriction on the set of feasible solutions and $\left\|\hat{\mathbf{P}}_{\dagger}^{L} - \hat{\mathbf{P}}_{\dagger}^{NL}\right\|^2$, the matrix norm on the difference between the matrices under the linear and non–linear formulation, increases as the model complexity increases. The MSE across time in Figure 3.7 shows the difference between the linear and non–linear formulation, and the two diverge with increasing model complexity, i.e. **DP** is a restricted version of **DW** because a set of weights that correspond to any α value can always be found. Similarly, the parameter space of **DW** is only a subset of **PEL**/**PENL**.

Method	**W**	**I**
DL	5882	1225
DE	7613	1317

(A) All N

Method	**W**	**I**
DL	80	206
DE	74	204

(B) N up to 90

Method	**W**	**I**
DL	5020	4628
DE	32743	25220

(C) N from 91

		All N		N up to 100		N from 101	
		t_0		t_0		t_0	
Method	Formulation	$t_0 = 1$	$N - 10$	$t_0 = 1$	$N - 10$	$t_0 = 1$	$N - 10$
PEL	**W**	2290	1186	128	90	9498	4838
	I	2009	1113	83	80	8429	4556
PENL	**W**	7340	1124	181	88	31205	4581
	I	6703	1434	110	90	28679	5913
DPL	**W**	2010	1189	75	63	8850	4942
	I	3066	1615	74	66	13040	6779
DPNL	**W**	5130	1744	80	64	21967	7345
	I	7208	2284	74	67	30990	9764
DWL	**W**	2135	967	74	70	9005	3956
	I	2604	1308	79	74	11025	5421
DWNL	**W**	4272	1536	92	81	18207	6387
	I	6050	1817	97	84	25886	7593

(D) MSE under different estimated methods

TABLE 3.3: Forum 50, with additional information after splitting the time period at 90. Summary of error over 90 consecutive weeks for all methods, each with a 10 week ahead forecast for three different forums. Both formulation of the inactive users were used with initial observation t_0 in the estimation at either $\mathbf{m}(1)$ or $\mathbf{m}(N-10)$, where N is the total number of observed mass at the current week.

For a forum that has a sudden change in behavior, such as forum 50 (Figure A.4), the predictability is low after the change (at approximately $N = 100$) as seen in Table 3.3. When setting a cut off point at $N = 100$, the difference in predictability of the two periods is evident in Table 3.3. The most successful forecasts were produced by **DL** for the latter period. This shows that none of the other methods adapted fast enough to the change of dynamics in the community. All the methods did adapt to the local structure when $t_0 = N - 10$ as it discarded older observations before the shift in community dynamic occurred. An example is to look at the changes of $\hat{\alpha}$ when using fewer observations, say by using $t_0 = N - 5$ instead of $t_0 = N - 10$, which then shows that $\hat{\alpha}$ for the former is always lower for time periods $N > 100$.

Figure 3.8 shows information on the number of new users joining and returning for all the compartments apart from the inactive users, which instead shows the number of inactive users and the total migration rate going out $(1 - \mathbf{P}_{0,0})$. Not

only was there a sudden surge of new users as well a return of inactive users, they also came through different compartments, i.e. changes in $\gamma_{Join}, \gamma_{Return}$. Even though the number of in–samples steps determine how sensitive the model is to recent trend, the effect of this will not be investigated further and the estimation of α is tuned by using the same number of steps as the desired out–of–sample forecasts which is fixed at 10.

In such scenario, it will be of major benefit if one can predict the change in dynamics. Unfortunately, such prediction usually require deep insights to the system and take into account many factors. The detection of shifts (in dynamics) will also help in discarding irrelevant observations but it is out of the scope of this thesis. We refer interested readers to the *change point analysis* literature; such as Rabiner (1989), Chib (1998) and Poor and Hadjiliadis (2008).

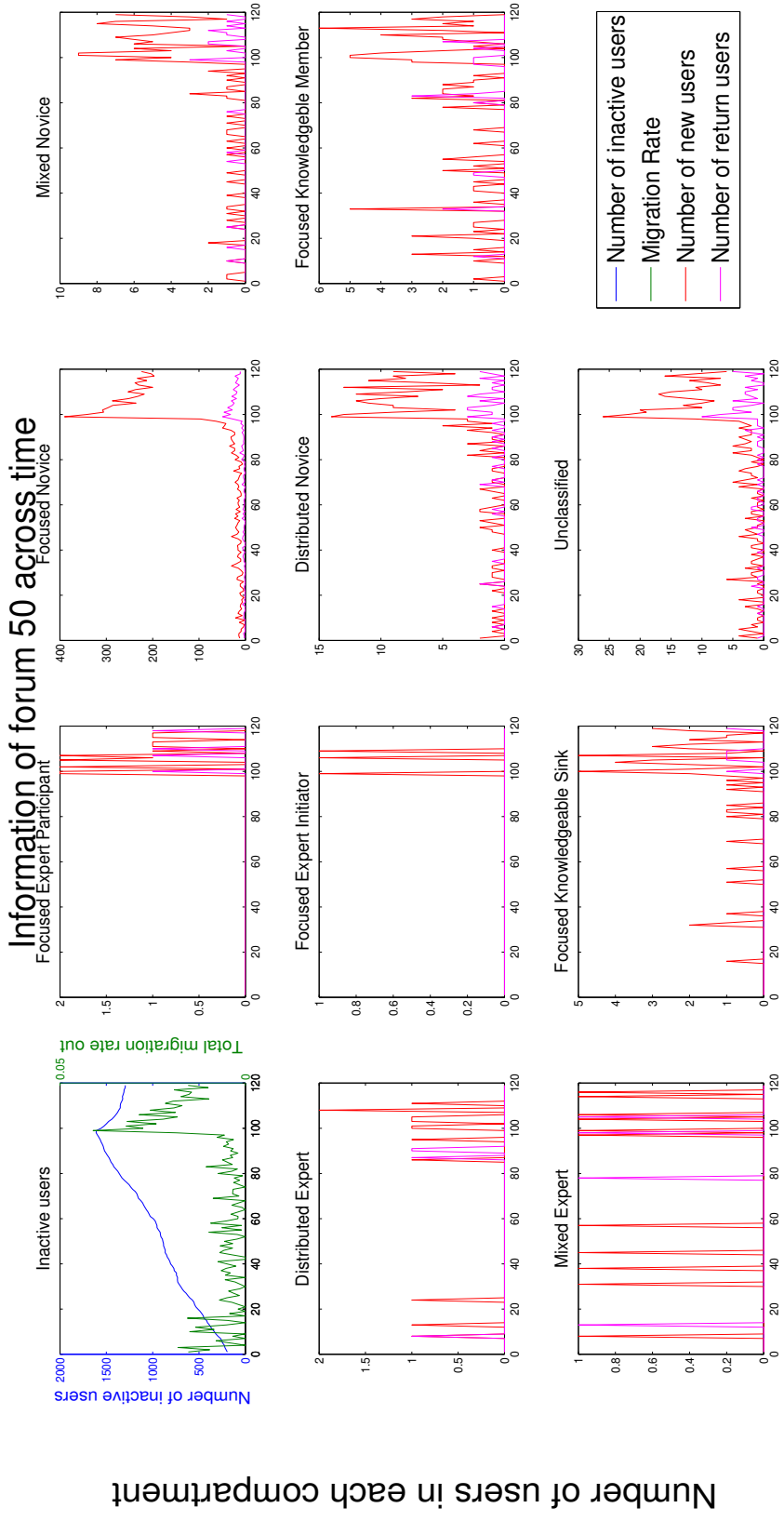


FIGURE 3.8: The number of users joining and returning for all the compartments apart from the inactive compartment. The inactive compartment shows the number of inactive users and the migration rate going out through time.

		Forum 353		Forum 256		Forum 264	
Method	Formulation	MSE	Coverage	MSE	Coverage	MSE	Coverage
SB	**W**	26.74	95	17.77	95	50.72	95
	I	40.48	94	17.20	95	74.56	94
SG	**W**	26.74	91	17.70	94	324.59	90
	I	40.51	91	17.45	95	347.96	90
SP	**W**	29.24	89	21.95	86	58.79	85
	I	29.73	88	16.77	92	65.20	85
SW	**W**	29.90	93	18.11	95	150.32	76
	I	31.34	92	20.15	93	73.65	72

TABLE 3.4: Summary of error over 90 consecutive weeks for the three stochastic methods as described in Section 3.5, each with a 10 weeks ahead forecast for three different forums.

3.6.2 Stochastic Models

The performances of the three different stochastic methods described in Section 3.5 will be explained in detail in this section. The Gaussian, Binomial and penalized methods described in Section 3.5.1.2, Section 3.5.1.3 and Section 3.5.2 are denoted as **SG**, **SB** and **SP** respectively.

In addition to MSE, the quality of the forecast is also measured by the coverage, defined as the percentage of out–of–sample observations that falls within the 95% pointwise confidence interval of the mean forecast, constructed via a Monte Carlo using J iterations, over all user roles and forecasted time steps.

$$\text{Coverage} = (qk^{-1}) \sum_{i=1}^{q} \sum_{g=1}^{k-1} \mathbf{1} \left\{ m^*_{g,0.025J}(t_0+i) \leq m_g(t_0+i) \leq m^*_{g,0.975J}(t_0+i) \right\} \tag{3.45}$$

Results were generated using the same set of observations points from the deterministic formulation, see Table 3.4. As mentioned previously, **SB** is a simulation version of a discrete time Markov chain that also provides the confidence interval. Hence, the almost exact MSE of **SB** in Table 3.4 and baseline of Table 3.2. The coverage of **SB** obtained is approximately 95% across all three forums, and none of them fell lower than 70 throughout all 90 weeks.

For **SG**, the results were similar for two forums but Forum 264 also demonstrated that this formulation can yield extremely poor results. This is because even though the expected value of the estimated TUVN (3.34) should be similar to $\text{diag}(\mathbb{E}(\mathbf{P}))$, it can be dramatically different when the observed transitions are either clustered

near the boundaries or spread evenly across the support $[0, 1]$. In fact, only two weeks produced very poor forecast in Forum 264 while the rest performed very similarly to **SG** like the other two forums.

Although **SG** has a lower coverage than **SB** for the three forums in Table 3.4, this is not always true because the variance of an TUVN is dependent on both the mean and variance while the variance of a Binomial is $p(1-p)$. Although the introduction of correlation did not change the sample mean of the resulting forecasts, it did lead to a different set of confidence intervals that are narrower in general.

Similarly for **SW**, forecasts for forum 264 were poor in certain weeks, different from those of **SG**. The coverage for **SW** were also the lowest out of the proposed methods for forum 264, due to the fact that the optimal weight vector $\hat{\mathbf{w}}$ had the majority of the weights placed on only a few observations. Note that $\hat{\mathbf{w}}$ is a minima that is not necessary global while $\hat{\alpha}$ is.

Both the MSE and coverage differ significantly for **SP** when compared to the other methods. The coverage was inferior for both the **W** and **I** formulation, whereas the MSE was only better when under the **I** formulation. Average coverage was close to 90% in most cases, but a more detailed analysis revealed that it has huge variability, anywhere from 10% to 100%. The amount of coverage follows closely with $\hat{\alpha}$, given that as $\alpha \to 0$, $\hat{\mathbf{P}} \to \mathbf{P}(t_0)$ and the sample variance $\text{Var}(\mathbf{P}) \to 0$.

We also considered the idea of using $\hat{\alpha}^L$, the penalty estimated under the linear formulation (3.31), to generate our Monte Carlo forecast but the results (in both MSE and coverage) were worse nearly all the time. This is simply due to the fact that the two sets of $\hat{\alpha}$ were significantly different. Similar to the deterministic models, it fails to provide adequate forecasts when the forum undergoes sudden change.

Chapter 4

Auxiliary Problem

4.1 Introduction

The models and results in the previous chapter (Chapter 3) assumed that we have the observations for both the number of users joining (y_{Join}) and returning (y_{Return}) to a community from inactivity, but in reality such information is not available. Now, we turn our attention to making such predictions for both of these unknowns.

As both y_{Join} or y_{Return} are non–negative integer sequential observations, obtained at equally spaced time points, they can be interpreted as a Poisson process, similar to say, the number of calls a telephone center receives in some fixed time period. Serial correlation cannot be seen at the start of the time series but tends to appear with an increasing number of observations in most forums. Neither of them seem to possess any (obvious) seasonal pattern, but $\text{Corr}(y_{Join}, y_{Return})$ is usually statistically significant.

Unfortunately, low value observations (< 5 or even 0) are not uncommon so classic time series analysis based on the Gaussian assumption seems inappropriate. Also note that if a series of observations are generated from a Poisson process, they are independent of each other by definition. The focus of this chapter is to introduce models that predict $\mathbf{Y} = [y_{Join}\ y_{Return}]$ either univariately or simultaneously as a multivariate response and is organized as follows.

First, we provide some generic and basic information of the Poisson family models that sets the scene for the rest of the chapter. An overview of count data prediction is presented in Section 4.2 before moving on to describe the univariate Poisson

regression in Section 4.3. Section 4.4 then details the estimation procedures for the simple Poisson regression, and the more popular version with regularization. The case of an overdispersed Poisson, when the equal mean–variance assumption of the Poisson fails to hold, is discussed in detail in Section 4.5. Both the univariate case (Section 4.5.2) and the extension to the multivariate version (Section 4.5.4) will be covered, where correlation between our response is modeled in addition to the overdispersion. A simple state space model with a latent autoregressive process, namely the AR(1), will be introduced in Section 4.5.3. Finally, the performance of all the models discussed will be compared in Section 4.6.

4.2 Regression of Count Data

Count data occurs naturally where the observations belong to the set of non–negative integers. The most common approach to model and predict such observations is to use regression. A regression model is based on the assumption that y is the conditional mean $\mathbb{E}(y \mid X = x)$ of a bivariate distribution $F(Y, X)$ given $\mu(x), \sigma^2(x)$ – functions on x for the mean and variance. The aim is to find a model that relates the observation Y to $\mu(x)$. There have been many models developed to model count data in both the (static) regression and the (dynamic) time series setting, a summary can be found in Cameron and Trivedi (1998) and only a brief introduction is provided here.

The Poisson distribution $y \sim \mathcal{P}(\lambda)$ is the most common and obvious choice to model count data as its support is defined by $\mathbb{N}$. Regression using the Poisson distribution assumes that $\log(\lambda) = \mathbf{x}^\top \boldsymbol{\beta}$ with regression coefficients $\boldsymbol{\beta}$, a more in depth introduction will be covered in Section 4.3. As the counts get large, the Poisson can be approximated by the normal distribution $y \sim \mathcal{N}(\lambda, \lambda)$ and the standard linear regression may be appropriate. When the counts are small, simply extensions by changing the basic assumptions of the Poisson regression like the quasi–Poisson have been used, and will be covered in Section 4.5 together with the Negative–Binomial model. In the case of excessive zeros, zero inflated Poisson or the hurdle model provides ways to model the zeros separately to the Poisson.

For counts that are all larger than 4, Bartlett (1936) proposed to take the square root of the observations so that they appear normal after the transformation. Anscombe (1948) later derived $\sqrt{(Y + 3/8)}$ using Taylor expansion under the aim of stabilizing variance and Brown et al. (2009) used $\sqrt{(Y + 1/4)}$ under the argument of minimizing bias. These are also known as “root–unroot” method and

are useful because of the normality assumption on the transformed data and have been used to tackle inhomogeneous Poisson process (Shen and Huang, 2008; Brown et al., 2009). But care should be taken when regressing $\sqrt{y}$ on x because the squaring the prediction is not a monotonic transformation, i.e. the prediction $\sqrt{\hat{y}} = -1$ is bigger than any prediction $-1 < \sqrt{\hat{y}} < 1$. An alternative transformation is to take the natural logarithm of Y and is especially useful when data is skewed. Due to the fact that the log of 0 is not defined, some small value is usually added to the observations before taking the log. Note that the predictions given by the direct inverse of these transformations $h(\cdot)$ using the estimated regression coefficient $\hat{\boldsymbol{\beta}}$ are biased because

$$\mathbb{E}(h(\mathbf{x}^\top \hat{\boldsymbol{\beta}})) \neq \mathbb{E}(h(\mathbf{x}^\top \hat{\boldsymbol{\beta}} + \varepsilon)).$$

This can be demonstrated by the Lognormal case, $X \sim \log \mathcal{N}(\mu, \sigma^2)$, where $\mathbb{E}(X) = \exp(\mu + \sigma^2/2)$ contains an additional term (variance contribution) in the exponent compare to the direct inverse transform of $h(\cdot) = \exp(\cdot)$. Appropriate techniques like using the Lognormal regression directly or the smearing estimate of Duan (1983) should be used when considering transformation techniques.

Alternative count modeling, also based on Poisson process are available. For example, "renewal equation" used in epidemiology (Fraser, 2007) assumes that the next new observation comes from a Poisson distribution with the mean a weighted average of past observations. The model carries an important idea in disease transmission in that the number of new cases are dependent on the current number of infected individuals minus those recovered. This idea is similar to the classic time series models based on the Gaussian assumption but is not considered as a time series model on discrete observations. Rather, time series models on count data are usually understood in the literate (Jung et al., 2006) to be one of the two; parameter–driven and observation–driven.

Parameter–driven models fall into the category of state–space models (Durbin and Koopman, 2001) which also includes classic time series theories when observations are of the Gaussian nature (Hamilton, 1994; Brockwell and Davis, 2009). For non–Gaussian observations, a temporal dependency is constructed using a latent autoregressive process that is independent of the observations, before transforming into the range of the observations. In contrast, observation–driven models, as the name suggest, use past observations to propagate the error forward. An INteger AutoRegressive (INAR) model that uses the last p observations is defined as (Jin-Guan and Yuan, 1991)

$$X_t = \psi_1 \circ X_{t-1} + \psi_2 \circ X_{t-2} + \ldots + \psi_p \circ X_{t-p} + \varepsilon_t, \tag{4.1}$$

where $\psi_i \circ X_{t-i}$ is the binomial thinning operator such that ψ_i and X_{t-i} defines the probability and the observation of a binomial distribution respectively. The error ε_t is a non–negative integer valued realization governed by some probability distribution. When ε_t (4.1) comes from a Poisson distribution then it is known as the Poisson auto–regression (Al-Osh and Alzaid, 1987; McCabe et al., 2011).

Unfortunately, the time series approaches we just described carry interpretations that do not appear to fit our data. The renewal equation assumes that the number of new users arose due to the influence of "recent new users" while observations from INAR have a proportion of "past new users" remaining. Such interpretation seems counterintuitive given our data and definition of new users. Therefore, we will not pursuit them further in this thesis.

4.3 Univariate Poisson Introduction

Regressions that are based on the exponential family with density of the form (4.2) are call *Generalized Linear Model* (GLM) which includes the linear regression. It contains θ the parameter of interest, $a(\phi)$ the dispersion factor and $\mathbf{y}$ the observations with normalizing constant $c(y, \phi)$. Then differentiating (4.2) shows that $b'(\theta) = \mathbb{E}(y) = \mu$ and in fact $b(\theta)$ is the cumulant that described the moment of the distribution.

$$f(y_i; \theta, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y, \phi)\right\} \tag{4.2}$$

This formulation connects the linear predictor η to the response through a link function $g(y) = \eta$, so that regression in the linear predictor $\eta = \mathbf{x}^\top\boldsymbol{\beta}$ can be transformed to the range of the response under the inverse link $\mu = g^{-1}(\eta)$. The dispersion ϕ is a measure of the model hypothesis and estimated using

$$\hat{\phi} = (n-p)^{-1}\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{\mathrm{Var}(\hat{y}_i)}, \tag{4.3}$$

where n and p are the number of observations and covariates respectively. The Poisson distribution is a member of the exponential family defined by a single parameter λ with p.d.f.

$$f(y; \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$$

and the log–likelihood function

$$\mathcal{L}(\lambda; y) \propto y \log(\lambda) - \lambda.$$

Matching the Poisson p.d.f. to (4.2), we get $\theta = \log(\lambda)$ and $b(\theta) = \lambda$ with $\phi = 1$. Hence, $g(\cdot) = \log(\cdot)$ is known as the log link function and the linear predictor is $\log(\lambda) = \mathbf{x}^\top \boldsymbol{\beta}$. Fitting the model and estimation of $\boldsymbol{\beta}$ will be covered in Section 4.4. Given the estimated regression coefficient $\hat{\boldsymbol{\beta}}$, the most straight forward way to quantify the fit of a GLM is the *Deviance*

$$D(\mathbf{y}; \hat{\mathbf{y}}) = 2\left(\mathcal{L}(\mathbf{y}; \mathbf{y}) - \mathcal{L}(\mathbf{y}; \hat{\mathbf{y}})\right), \tag{4.4}$$

two times the difference between the log–likelihood of the full and estimated model. The scalar of 2 makes (4.4) equivalent to MSE for a normal model, so the deviance is distributed as χ^2. For other members of the exponential family, deviance is asymptotically χ^2. This allows model comparison between nested models using likelihood ratio tests (McCullagh and Nelder, 1989) based on the χ^2. Deviance has also been used as a direct model comparison in Bayesian statistics (Spiegelhalter et al., 2002) where the deviance (4.4) only consists of the estimated model. It is especially useful when comparing model estimated using simulation based methods, where model selection is done by comparing the Deviance Information Criterion (DIC), commonly defined as

$$\text{DIC} = D(\bar{\theta}) + 2pD, \tag{4.5}$$

where

$$pD = \bar{D} - D(\bar{\theta}) \tag{4.6}$$

with $\bar{D}, \bar{\theta}$ being the sample average of the deviance and the parameters from the simulation. Like other covariance type penalty (such as AIC/BIC), the model with a lower DIC is the favorable model and it penalizes overparameterization through the effective number of parameters pD.

Goodness of fit on any model cannot be based on the likelihood alone. Information on the residuals, such as the Durbin–Watson test (Durbin and Watson, 1971) on autocorrelation and Cook's distance (Cook, 1979) to detect outliers, provides important assessment on the model fit. But unlike the normal model, the raw residuals $e_i = y_i - \hat{y}$ should not be used because the residuals should be normally distributed. Furthermore, there is no strict definition on the source of error in other

GLM and there exist many definitions of residuals. The most common adjustment is the Pearson residuals

$$e_{Pearson} = \frac{y_i - \hat{y}_i}{\sqrt{\mathrm{Var}(\hat{y}_i)}},$$

which is the raw over the standard deviation like the central limit theorem. Other definition of residual like the deviance residual

$$e_{Deviance} = \mathrm{sgn}(y_i - \hat{y}_i)\sqrt{D(y_i; \hat{y}_i)}$$

where $\mathrm{sgn}(\cdot)$ is the sign function, and the linear predictor residual

$$e_{LP} = \frac{g(y_i) - g(\hat{y}_i)}{\sqrt{g'(\hat{\mu}_i)^2 \,\mathrm{Var}(\hat{y}_i)}}$$

are also used to represent the different ways errors enter into GLM. Pearson residuals is the most popular choice because it is required to estimate dispersion, where $\hat{\phi}$ (4.3) is summed over the Pearson residuals squared divided by the model degree of freedom. Therefore, it is readily available and is routinely used for model diagnosis, especially for models that uses a theoretical dispersion such as the Binomial and Poisson.

The Poisson distribution has a mean–variance relationship, i.e. $\mathbb{E}(y) = \mathrm{Var}(y) = \lambda$, hence the theoretical value of $\phi = 1$ and deviation from that show signs of a poor fit. As $\hat{\phi}$ scales with $n - p$, the model degree of freedom, a better model in terms of deviance can yield higher dispersion.

4.4 Estimation Procedures

The basic linear regression $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ assumes that the response $\mathbf{y}$ has a linear relationship with the linear predictor $\mathbf{X}\boldsymbol{\beta}$ and the error ε, where $\mathbb{E}(\varepsilon) = 0, \mathrm{Var}(\varepsilon) = \sigma^2$. Finding the regression coefficients to the linear system is usually expressed using the normal equation

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

This is because it contains the Hessian $(\mathbf{X}^\top \mathbf{X})$ of the linear model which contributes to the variance $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}$, where $\hat{\sigma}^2$ is estimated after finding $\hat{\boldsymbol{\beta}}$. This is also known as ordinary least squares (OLS) and is a special case of the generalized least squares (GLS) which assumes $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \Omega$. The

regression coefficients of GLS is obtained via

$$\arg\min_{\boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\top}\Omega^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).$$

and the solution can be shown (Hayashi, 2000, sec. 1.6) to have

$$\mathbb{E}(\hat{\boldsymbol{\beta}})=(\mathbf{X}^{\top}\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}^{\top}\Omega^{-1}\mathbf{y},\ \mathrm{Var}(\hat{\boldsymbol{\beta}})=(\mathbf{X}^{\top}\Omega^{-1}\mathbf{X})^{-1}. \tag{4.7}$$

When Ω is a diagonal matrix, then it is also known as weighted least squares (WLS) and both WLS/GLS are fundamental building blocks in the estimation of GLMs'. We start off with the maximum likelihood estimation (MLE) in Section 4.4.1 before moving on to the Bayesian formulation that incorporate prior information in Section 4.4.2. We also make the connection between the Bayesian formulation and the so called *shrinkage estimators* through the Maximum–a–posteriori estimation (MAP). Finally, we present simulation based estimation, namely Monte Carlo Markov Chain (MCMC) which will be used in the latter part of this chapter when the likelihood becomes intractable.

4.4.1 Maximum Likelihood Estimate (MLE)

The set of regression coefficients $\boldsymbol{\beta}$ in the linear predictor η can be found by maximizing the log–likelihood function (4.8) or equivalently, minimizing the negative log–likelihood function.

$$\mathcal{L}(\theta;y_i,\phi)\propto\frac{y_i\theta_i-b(\theta_i)}{a(\phi)} \tag{4.8}$$

Standard procedure to obtain $\hat{\boldsymbol{\beta}}$ is to use the Iterative Reweighted Least Square (IRLS) algorithm (4.9), derived by using a Taylor's expansion on the first derivative of the log–likelihood (Hardin and Hilbe, 2007, sec. 3.3). Convergence check can be performed on either the log–likelihood $\left|\mathcal{L}^{(t+1)}-\mathcal{L}^{(t)}\right|<\epsilon$ or in the regression coefficients $\left|\beta_p^{(t+1)}-\beta_p^{(t)}\right|<\epsilon\ \forall\ p$ for some tolerance value ϵ. As the name suggest, $\boldsymbol{\beta}$ is estimated by performing a series of weighted least squares (WLS)

$$(\mathbf{z}^{(t)}-\mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{W}^{(t)}(\mathbf{z}^{(t)}-\mathbf{X}\boldsymbol{\beta}) \tag{4.9}$$

in the linearized version of the log–likelihood to obtain the updated $\boldsymbol{\beta}^{(t+1)}$. Both $\mathbf{z}^{(t)}, \mathbf{W}^{(t)}$ are in fact functions of $\boldsymbol{\beta}^{(t)}$ here, defined as

$$z_i = (y_i - \mu_i)g'(\mu_i) + \mathbf{x}_i^\top \boldsymbol{\beta} - offset_i \tag{4.10}$$

$$w_{i,i} = (a(\phi)\,\mathrm{Var}(y_i)g'(\mu_i)^2)^{-1}. \tag{4.11}$$

The variable $offset$ in (4.9) can be thought to have a regression coefficient of 1 that provides the ability to incorporate prior knowledge into the fitting procedure. In the Poisson case, $offset$ is usually used to model the rate of a process because $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + offset_i$ becomes $e^{\mathbf{x}_i^\top \boldsymbol{\beta}} e^{offset_i}$, so the mean parameter $e^{\mathbf{x}i^\top \boldsymbol{\beta}} = \lambda_i$ describes the rate y_i / e^{offset_i}.

As (4.9) is a WLS, it is immediate that $\boldsymbol{\beta}$ at each iteration including $\hat{\boldsymbol{\beta}}^{\mathrm{MLE}}$, is distributed according to

$$\boldsymbol{\beta}^{(t+1)} \sim \mathcal{N}(\mathbf{c}(\boldsymbol{\beta}^{(t)}), \mathbf{H}^{-1}(\boldsymbol{\beta}^{(t)})), \qquad \begin{aligned} \mathbf{c}(\boldsymbol{\beta}) &= \mathbf{H}^{-1}(\boldsymbol{\beta})\mathbf{X}^\top \mathbf{W}(\boldsymbol{\beta})\mathbf{z}(\boldsymbol{\beta}) \\ \mathbf{H}(\boldsymbol{\beta}) &= \mathbf{X}^\top \mathbf{W}(\boldsymbol{\beta})\mathbf{X} \end{aligned} \tag{4.12}$$

where $\mathbf{H}(\boldsymbol{\beta})$ is the expected Fisher information (4.13), usually denoted as $\mathbf{I}(\boldsymbol{\beta})$ or more generally as $\mathbf{I}(\theta)$. We use $\mathbf{H}(\boldsymbol{\beta})$ because it is also the expected Hessian, variance of the log–likelihood gradient.

$$I(\boldsymbol{\theta}) = \mathbb{E}\left(\left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} \right)^2 \,\middle|\, \theta \right) \tag{4.13}$$

An alternative is to use Newton–Raphson to estimate our parameters, which uses the observed Hessian (4.14). Then the two formulations (4.13) and (4.14), are in fact equivalent under a canonical link function which means that $\mathbf{c}(\boldsymbol{\beta}), \mathbf{H}(\boldsymbol{\beta})$ can be obtained directly using the gradient and Hessian of the log–likelihood (4.8) instead of evaluating (4.9, 4.10, 4.11)

$$I(\boldsymbol{\theta}) = -\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \,\middle|\, \boldsymbol{\theta} \tag{4.14}$$

Obviously, IRLS is a generic algorithm and it can be used to solved other types of regression problem in addition to GLM. For example, the Least Absolute Deviation regression, where the objective is to minimize $|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|$, using the appropriate $\mathbf{W}, \mathbf{z}$ (Vanderbei, 1998, chap. 12).

4.4.2 Maximum–a–posteriori (MAP)

The Bayesian formulation is based on Baye's Rule (4.15) where a prior $f(\theta)$ is placed on the parameter of interest such that the posterior $f(\theta \mid y)$ is proportional to the likelihood $f(y \mid \theta)$ times the prior. The parameters in the prior distribution are called *hyperparameters* which can either be fixed or estimated.

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{f(y)} \tag{4.15}$$

The posterior can be obtained analytically for certain models i.e. in the case of a linear regression. In the GLM setting, the MAP estimate is the mode of the posterior and it can be obtained by extending the IRLS algorithm for those priors with a first and second derivative. For example, if a Gaussian prior is placed on $\boldsymbol{\beta}$ with hyperparameters $\mathbf{b}_0, \mathbf{B}_0$ as the mean and variance, the posterior is

$$f(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \mathbf{b}_0, \mathbf{B}_0) \propto f(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X}) f(\boldsymbol{\beta} \mid \mathbf{b}_0, \mathbf{B}_0). \tag{4.16}$$

Because we can write the likelihood as a WLS (4.9), which can be interpreted as a normal distribution with covariance matrix $\mathbf{W}^{-1}$, (4.16) becomes

$$f(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \mathbf{b}_0, \mathbf{B}_0) \propto \phi(\mathbf{z} \mid \mathbf{X}\boldsymbol{\beta}, \mathbf{W}^{-1}) \phi(\boldsymbol{\beta} \mid \mathbf{b}_0, \mathbf{B}_0)$$

where ϕ is the normal distribution density function. Then take the natural logarithm and ignore the covariance term (as they vanish in the derivative)

$$\mathcal{L}(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \mathbf{b}_0, \mathbf{B}_0) \propto -\frac{1}{2} \left(\frac{(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})}{\mathbf{W}^{-1}} + \frac{(\boldsymbol{\beta} - \mathbf{b}_0)^\top (\boldsymbol{\beta} - \mathbf{b}_0)}{\mathbf{B}_0} \right) \tag{4.17}$$

shows the addition of a prior to (4.8) is equivalent to adding pseudo observations $\mathbf{b}_0$ with weights $\mathbf{B}_0^{-1}$. Therefore, we can also write (4.17) in the form of (4.9)

$$\|\tilde{\mathbf{W}}^{1/2}(\tilde{\mathbf{z}} - \tilde{\mathbf{X}}\boldsymbol{\beta})\|^2, \ \tilde{\mathbf{z}} = [\mathbf{z} \ ; \ \mathbf{b}_0] \, , \ \tilde{\mathbf{X}} = [\mathbf{X} \ ; \ \mathbf{I}_p] \, , \ \tilde{\mathbf{W}} = \left[\mathbf{W} \ ; \ \mathbf{B}_0^{-1}\right]$$

which is a GLS as the weight matrix is no longer diagonal unless $\mathbf{B}_0$ is a diagonal matrix. Therefore, finding the solution with the addition of a Gaussian prior is no harder than a standard GLM. The regression coefficients are now distributed as (4.18) (West et al., 1985; Gamerman, 1997) and it approaches the MLE solution

as both $\mathbf{b}_0, \mathbf{B}_0 \to 0$.

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{c}(\boldsymbol{\beta}), \mathbf{H}^{-1}(\boldsymbol{\beta})), \qquad \begin{aligned} \mathbf{c}(\boldsymbol{\beta}) &= \mathbf{H}^{-1}(\boldsymbol{\beta})(\mathbf{X}^\top \mathbf{W}(\boldsymbol{\beta})\mathbf{z}(\boldsymbol{\beta}) + {\mathbf{B}_0}^{-1}\mathbf{b}_0) \\ \mathbf{H}(\boldsymbol{\beta}) &= \mathbf{X}^\top \mathbf{W}(\boldsymbol{\beta})\mathbf{X} + {\mathbf{B}_0}^{-1} \end{aligned} \tag{4.18}$$

Both the hyperparameters do not have to be fixed, but estimated from the data by iteratively going through the estimation of $\mathbf{b}_0, \mathbf{B}_0$ given $\boldsymbol{\beta}$ (Lindley and Smith, 1972). When the prior mean is assumed to be zero and the covariance is a scalar, it can be shown that the estimation of (4.17) is equivalent to placing a restriction on the squared norm of the regression coefficient, one of the popular shrinkage methods.

4.4.3 Regularization Methods

Regularization methods aim to solve an ill–posed problem by favoring certain solution. This idea in regression was first introduced in the form of *ridge regression* by Hoerl and Kennard (1970) to find the least squares solution via

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y},$$

where λ is some pre–determined value. The first benefit of this is to make the matrix $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})$ diagonal dominant to ensure that it is invertible. The second is to improve prediction as Stein (1956) and later James and Stein (1961) showed that an unbiased estimator does not necessarily achieve the lowest MSE for the normal distribution. This can be seen from the Bias–Variance decomposition where $\text{MSE} = \text{Variance} + \text{Bias}^2$ (Narsky and Porter, 2013, sec. 5.6), and the addition of λ is an attempt to lower the MSE by introducing bias into the estimation. The third is to prevent overfitting because the effective degree of freedom increases as λ increases (Hastie et al., 2009).

Ridge regression is in fact the same as (4.17) when the prior is the normal distribution given that $\mathbf{z}(\boldsymbol{\beta}) = \mathbf{y}$ and $W(\boldsymbol{\beta}) = \sigma^{-2}$. Let $\mathbf{B}_0 = \tau^2 \mathbf{I}$ and $\mathbf{b}_0 = 0$, then we can see that $\lambda = \sigma^2/\tau^2$ (Lindley and Smith, 1972). In the GLM setting, the ridge estimator can be seen from a penalization point of view by letting $\mathbf{b}_0 = 0, \lambda = \mathbf{B}_0^{-1}$, such that our objective function becomes

$$\arg\min_{\boldsymbol{\beta}} -\mathcal{L}(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}, \phi) + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2. \tag{4.19}$$

Writing (4.19) in the equivalent form (4.20) shows that $\|\hat{\boldsymbol{\beta}}^{\text{Ridge}}\|^2 \leq \|\hat{\boldsymbol{\beta}}^{\text{MLE}}\|^2$ and they are only equal when $t \geq \sum_{j=1}^{p}(\hat{\boldsymbol{\beta}}_j^{\text{MLE}})^2$. Hence, ridge regression is also known as a shrinkage method.

$$\begin{aligned} \arg\min_{\boldsymbol{\beta}} \quad & -\mathcal{L}(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}, \phi) \\ \text{s.t.} \quad & \textstyle\sum_{j=1}^{p} \beta_j^2 \leq t \end{aligned} \tag{4.20}$$

The same connection can also be made for another popular shrinkage estimator, the Lasso (Tibshirani, 1996), where the penalty is applied to the L_1 norm of the parameters (4.21).

$$\arg\min_{\boldsymbol{\beta}} -\mathcal{L}(\boldsymbol{\beta}; y, \phi) + \lambda\|\boldsymbol{\beta}\|_1 \tag{4.21}$$

This is a popular method because it also acts as a variable selection method by forcing components to zero (Hastie et al., 2009). The L_1 penalty is equivalent to placing a Laplace prior (4.22) on $\boldsymbol{\beta}$, which can be seen from (4.21) by using the same argument as the ridge. But the Lasso estimator cannot be incorporated into the estimation of GLM as easily as the ridge, because $\|\boldsymbol{\beta}\|_1$ is not differentiable everywhere. Estimating (4.21) is a much harder optimization problem and many methods have been proposed throughout the years, such as Tibshirani (1996); Osborne et al. (2000); Efron et al. (2004); Lee et al. (2006); Park and Hastie (2007); Goeman (2010); Friedman et al. (2010), in an attempt to not only estimate a single λ, but for the whole path of λ ranging from 0 to some large value where all $\boldsymbol{\beta}$'s are zero. Selecting the appropriate λ is then quantified through some model selection criteria such as Mallow's C_p (Efron et al., 2004) or cross–validation (Park and Hastie, 2007) to prevent overfitting.

$$f(x; \lambda) = \frac{\lambda}{2} e^{-\lambda|x|} \tag{4.22}$$

Although (4.18) only holds for a normal distribution prior, it is useful in MCMC estimation. This is because random samples from other common priors, such as the t–distribution and Laplace distribution, have a normal mixture representation (Andrews and Mallows, 1974). Park and Casella (2008) used the mixture representation to build up a hierarchical structure for the Laplace prior and performed their estimations using MCMC. It is important to note that shrinkage is not applied to the intercept and usually on a standardized design matrix to ease the interpretation of the regression coefficients as the magnitude becomes comparable.

4.4.4 Monte Carlo Markov Chain (MCMC)

Estimation of the parameter can also be thought of as making inference on the posterior distribution, where the point estimate of MAP represents the mode of the posterior. Let $f(\theta)$ be an unnormalized density function such that $p(\theta) = f(\theta)/\int f(\theta)d\theta$ is a probability density function. The main advantage of MCMC is that it draws samples from $f(\theta)$ directly and therefore is able to make inference on $f(\theta)$ even when the normalizing constant is unknown. Details of using MCMC and the convergence theory behind it can be found in (Gelman et al., 2003; Robert and Casella, 2005) so the following discussion will be brief.

The idea behind MCMC is to construct a Markov chain on the stationary distribution $\pi(\cdot)$ that is the same as the posterior of the parameter. Each move on the Markov chain corresponds to generating a new sample, and the moves must be reversible, i.e. satisfying the detail balance condition

$$\pi(\theta^*)q(\theta^*, \theta) = \pi(\theta)q(\theta, \theta^*),$$

where $q(\theta, \theta^*)$ is the transition density from θ to θ^*. Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) is a general algorithm that satisfies the reversibility condition (Chib and Greenberg, 1995) by proposing a move $\theta \to \theta^*$ with density $q(\theta^* \mid \theta)$ and accepting the move with probability

$$\alpha(\theta, \theta^*) = \min\left\{\frac{p(\theta^*)q(\theta \mid \theta^*)}{p(\theta)q(\theta^* \mid \theta)}, 1\right\}. \tag{4.23}$$

A popular choice is to perform a random walk by using a symmetric distribution such as the normal distribution with some covariance matrix $\mathbf{V}$ and the proposal generated by $\theta^* \sim \mathcal{N}(\theta, \mathbf{V})$. This means that $q(\theta^* \mid \theta) = q(\theta \mid \theta^*)$ and they cancel in (4.23), so only the evaluation of the likelihood is required. Selecting a suitable covariance matrix $\mathbf{V}$ that explores the posterior well is a difficult task and usually requires a lot of tuning, either manually or automatically through some adaptive schemes such as those described in Roberts and Rosenthal (2009).

Gibbs sampling is a special case of the Metropolis–Hastings where the acceptance probability is 1. It exploits the fact that a conditional distribution is proportional to its joint distribution (without normalizing constant), and obtaining samples from certain distributions are relatively straight forward. Therefore, it is suffice to sample from each and everyone of the conditionals iteratively, see Casella and George (1992) for a more in depth discussion.

When there are multiple parameters of interest, $\boldsymbol{\theta} = (X_1, \ldots, X_p)$ say, they can be performed simultaneously on the full joint distribution $f(\boldsymbol{\theta})$ by making Metropolis–Hastings move or a Gibbs sampling step if available. Although a Gibbs sampling step may not exist for the full joint distribution, it may exist in the conditional distributions. Let $\mathcal{F}$ be a set that contains all indexes for the parameters of interest $\boldsymbol{\theta}$, $\mathcal{A}$ a subset of $\mathcal{F}$ with complement $\mathcal{A}^c$ such that $\mathcal{A} \cup \mathcal{A}^c = \mathcal{F}$. Then there may exist a Gibbs sampling step for the conditional $f(\boldsymbol{X}_{\mathcal{A}} \mid \boldsymbol{X}_{A^c})$, where $\mathcal{A}$ may consist of multiple or a single element of $\boldsymbol{\theta}$.

The predictions can be generated within the MCMC scheme, which accounts for both the inferred variance of the parameters and the model uncertainty. Let the parameter of interest be $\boldsymbol{\theta}$ and denote y_{pred} as the prediction of y, then the posterior predictive distribution is

$$f(\mathbf{y}_{\text{pred}} \mid \mathbf{y}) = \int f(\mathbf{y}_{\text{pred}} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}, \tag{4.24}$$

i.e. the integral of the likelihood function of $\mathbf{y}_{\text{pred}}$ with respect to the posterior distribution of the parameters $\boldsymbol{\theta}$. When making out–of–sample predictions, i.e. y_+ is the unobserved response with observed covariates $\mathbf{x}_+$ and $\hat{y}_+$ be the prediction on our true value y_+, then (4.24) can be written as

$$f(y_+ \mid \mathbf{x}_+) = \int f(y_+ \mid \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid \mathbf{x}_+) d\boldsymbol{\theta}, \tag{4.25}$$

where the again samples from (4.25) provides a summary of the distribution. Obtaining samples in regression for a member of the exponential family only requires a few additional steps during the estimation stage, namely, at each iteration $t = 1, 2, \ldots, J$

1. Given samples $\boldsymbol{\beta}^{(t)}$ and observation $\mathbf{x}_+$
2. Compute $\eta_+^{(t)} = \mathbf{x}_+^\top \boldsymbol{\beta}^{(t)}$
3. Sample from $y_+^{(t)} \sim f(g^{-1}(\eta_+^{(t)}))$

where $\mu = g^{-1}(\eta)$ is the prediction and $f(\cdot)$ the density of the distribution.

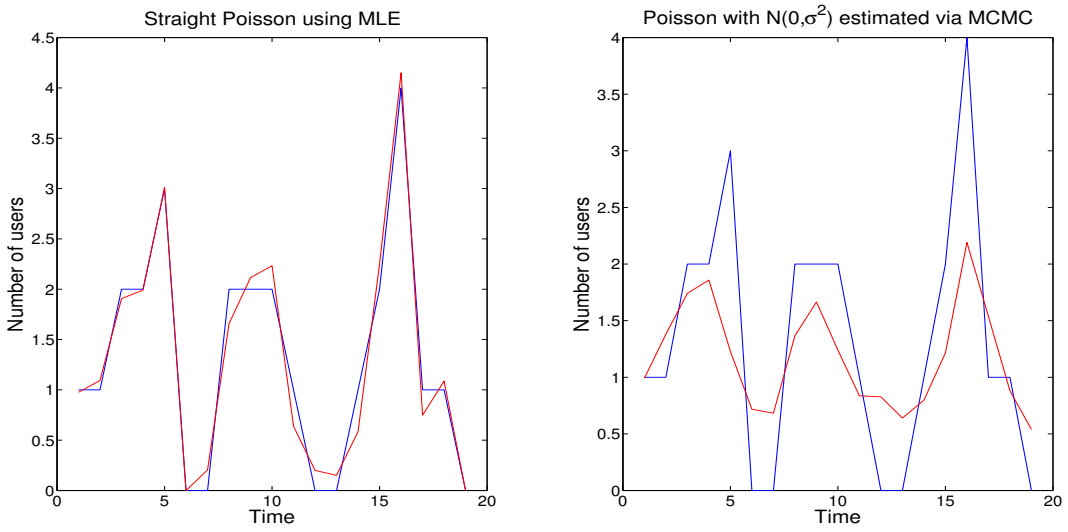


FIGURE 4.1: Observed (blue) and predictions (red) by the fitted model of the first 19 observations, using both the lagged and time dependent covariates

4.4.5 MCMC Estimation For Poisson Regression

As mentioned previously, regularization methods prevent overfitting by restricting the contributions of the regression coefficients. We begin by briefly demonstrating the benefit before moving on to describe the MCMC setup of a Poisson regression.

A total of 19 observations, the number of users returning to forum 353, was fitted using 15 variables and the MLE version produced a near perfect fit, see left panel of Figure 4.1. A standard t–test revealed that none of the $\boldsymbol{\beta}$'s estimated under MLE were significant even though three of the regression coefficients had an absolute value larger than 30. When estimated under a Gaussian prior on the regression coefficients with the variance parameter estimated via MCMC, only the trend/seasonality remains. To make it more comparable, we find the MAP estimates by using the point estimate $\lambda = \hat{\sigma}^2$, the sample average of the MCMC simulation. The AIC for MAP is 57 (lower than MLE) while the significant reduction in the effective number of parameters as it decreased from 16 to 6.2. Now, we turn to the MCMC formulation and the sampling schemes of the parameter simulations.

For the Poisson regression, Albert (1992) tackled it using a quasi–likelihood formulation and used Gibbs sampling at the mode of $\boldsymbol{\beta}$. Gamerman (1997) used (4.18) to perform a series of Metropolis–Hastings steps. Frühwirth-Schnatter and Wagner (2006) used an approximate Gibbs sampling procedure through the use of a finite normal mixture to represent $\log(\lambda)$, which is linear in terms of $\boldsymbol{\beta}$. Chib et al. (1998) used a t–distribution proposal with the variance evaluated at the

mode of the conditional distribution at each iteration, while Martin et el. 2011 performed a random walk using a scaled covariance matrix at MAP.

For a standard Poisson regression, the only parameter of interest is $\boldsymbol{\beta}$. From a Bayesian perspective and the arguments in Section 4.4.2, we will also infer the variance parameter by assuming $\mathbf{b}_0 = 0, \mathbf{B}_0 = \sigma^2\mathbf{I}$. It can also be written as

$$\boldsymbol{\beta}_j \sim \mathcal{N}(0, \sigma^2)\ \forall j, \tag{4.26}$$

and either one will be used depending on which provides the more compact expression. As mentioned in Section 4.4.3, this formulation is akin to the ridge estimator but unlike the MAP estimation, we can use the conjugate prior

$$\sigma^{-2} \sim \mathcal{G}a(a_0, b_0)$$

to infer the distribution of σ^2 given some hyperparameters a_0, b_0. It is usually performed using diffuse hyperparameters $a_0 = b_0 = \epsilon$ where ϵ is something small like 0.001 such that the gamma distribution has an expectation of 1 and a large variance. But this may be a poor choice and alternatives such as the lognormal (Barnard et al., 2000) or uniform distribution as a prior has been used in the literature (Gelman, 2006). A non–parametric estimate of σ^2 can also be done using an empirical Bayes technique (Casella, 2001) that only requires the evaluation of the conditional expectation. Despite the alternatives, we use a diffuse conjugate prior as it has the advantage of preserving a closed form posterior with a simple Gibbs sampling step, which amounts to sampling

$$(\sigma^{-2})^{(t+1)} \sim \mathcal{G}a(a, b) \tag{4.27}$$

where $a = a_0 + p/2$ and $b = b_0 + \|\boldsymbol{\beta}^{(t)}\|^2/2$. The same gamma prior is applicable for the Lasso using the formulation of Park and Casella (2008) as

$$\boldsymbol{\beta}_j \sim \mathcal{N}(0, \tau_j^2), \quad \tau^2 \sim \mathcal{E}xp(\lambda^2/2), \quad \lambda^2 \sim \mathcal{G}a(a_0, b_0). \tag{4.28}$$

To sample the regression coefficients, we use the sampling plan in Zeger and Karim (1991); Gamerman (1997), a Metropolis–Hastings steps based on (4.18). This is because the posterior of $\boldsymbol{\beta}$

$$f(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \sigma^2) \propto f_{\mathcal{P}}(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2) f_{\mathcal{N}}(\boldsymbol{\beta} \mid \sigma^2), \tag{4.29}$$

does not have any close form solution for a Poisson density. But because the

Poisson uses a canonical link function, one iteration of (4.9) is simply a Newton step, where the gradient and Hessian are (C.5) and (C.6) respectively. Therefore, we can evaluate $\mathbf{c}(\boldsymbol{\beta}), \mathbf{H}(\boldsymbol{\beta})$ directly given $\boldsymbol{\beta}^{(t)}$ and $(\sigma^2)^{(t)}$. The proposal $\boldsymbol{\beta}^*$ is then drawn from from (4.18) with transition probability $\mathcal{N}(\boldsymbol{\beta}^* \mid \mathbf{c}(\boldsymbol{\beta}^{(t)}), \mathbf{H}^{-1}(\boldsymbol{\beta}^{(t)}))$. The acceptance probability (4.23) at each iteration given the proposal is

$$\min\left\{\frac{\mathcal{N}(\boldsymbol{\beta}^{(t)} \mid \mathbf{c}(\boldsymbol{\beta}^*), \mathbf{H}^{-1}(\boldsymbol{\beta}^*))\mathcal{N}(\boldsymbol{\beta}^* \mid \mathbf{b}_0, \mathbf{B}_0)\prod_{i=1}^n f_{\mathcal{P}}(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}^*)}{\mathcal{N}(\boldsymbol{\beta}^* \mid \mathbf{c}(\boldsymbol{\beta}^{(t)}), \mathbf{H}^{-1}(\boldsymbol{\beta}^{(t)}))\mathcal{N}(\boldsymbol{\beta}^{(t)} \mid \mathbf{b}_0, \mathbf{B}_0)\prod_{i=1}^n f_{\mathcal{P}}(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}^{(t)})}, 1\right\}$$

with reverse transition probability computed using the proposal $\boldsymbol{\beta}^*$. This is an unconventional choice as a multivariate t–distribution (MVT) (Chib et al., 1998) is more commonly used as the proposing distribution,

$$\boldsymbol{\beta}^* \sim MVT(\boldsymbol{\mu}, \mathbf{V}, v_0) \tag{4.30}$$

where $\boldsymbol{\mu}$ is the mean, $\mathbf{V}_0$ the variance and v_0 some pre–determined degree of freedom, say 5 or 10. Then there are two common ways to find $\boldsymbol{\mu}$, both making use of the information at the mode. Let the regression coefficients and variance at the mode be $\boldsymbol{\beta}^M$ and $\mathbf{V}^M$ respectively, both can be found by a few Newton steps on the conditional (4.29). The first is to let $\boldsymbol{\mu} = \boldsymbol{\beta}^M$ which requires the calculation of the transition probability $q(\cdot, \cdot)$. The second scheme omits the transition probability via a reflect in the proposal with $\boldsymbol{\mu} = \boldsymbol{\beta}^M - (\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^M)$ (Chib and Greenberg, 1995; Chib et al., 1998) as it becomes a random walk, i.e. the forward transition is

$$\begin{aligned} q(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(t)}) &= MVT(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^M - (\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^M), \mathbf{V}^M, v_0) \\ &= MVT(\boldsymbol{\beta}^* + \boldsymbol{\beta}^{(t)} - 2\boldsymbol{\beta}^M \mid 0, \mathbf{V}^M, v_0) \end{aligned}$$

which is the same as the backward transition $q(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^*)$.

We demonstrate the efficiency of the three sampling schemes mentioned previously, IRLS – which samples from (4.18) and MVTM/MVTR that samples from (4.30) at the mode and under reflection respectively. Define the inefficient factor as $\text{Ineff} = 1 + \sum_{k=1}^{\infty} \rho(k)$ (Kass et al., 1998), where $\rho(k)$ is the autocorrelation at lag k. Taking the number of iterations, G, over the inefficient factor yields *Effective Sample Size* (ESS), which is $\text{ESS} = G/\text{Ineff}$ and is a common measure for the efficiency in generating samples. The aim is to achieve as little autocorrelation in the sample as possible, with $\rho(k) = 0$ for all k (i.e. $\text{Ineff} = 1$) indicating that we have i.i.d. realizations.

Samples of $\boldsymbol{\beta}$ for number of users joining forum 353 using the first 100 observations were generated using all 3 sampling schemes described above, the results can be

Prior	Method	Ineff (min/max)	Time	Ineff × time
No Prior	IRLS	3.10 (1.90/4.21)	23.06	71
	MVTM	4.07 (2.84/6.61)	27.60	112
	MVTR	27.84 (13.84/53.49)	23.61	657
Normal	IRLS	2.39 (1.96/3.25)	24.21	57
	MVTM	3.04 (1.91/4.77)	30.95	94
	MVTR	24.08 (14.34/37.52)	26.87	646
Laplace	IRLS	3.78 (1.81/6.98)	24.78	94
	MVTM	4.06 (2.43/9.54)	32.10	130
	MVTR	25.26 (14.56/44.05)	26.18	661

TABLE 4.1: Forum 353 with first 100 observations. Demonstrating the difference in inefficient factor (Ineff) and the time taken (seconds) between the 3 methods for regression coefficients for 1×10^5 iteration

seen in Table 4.1. The average Ineff between all p regression coefficient were reported, as well as the maximum and minimum for the three schemes mentioned above. The total time taken, for 1×10^4 iterations as well as the total inefficiency (denominator of ESS) can also be seen. It was found that for our data, sampling using IRLS is about as efficient as MVTM but MVTR performed extremely badly.

When taking into account the time spent, IRLS had better performance as it is a quicker procedure. This is because finding the mode requires a few extra Newton steps, average of about 5 regardless of prior, compare to the two used in IRLS. Taking these few extra Newton steps are relatively expensive, as it scales with the number of observations and covariates. Autocorrelation plots in Figure C.1 for the Gaussian prior and Figure C.2 for the Laplace prior again demonstrates that MVTR was clearly inferior to the other two schemes.

4.5 Overdispersed Poisson

A common problem with count data is that the equal mean–variance relationship of the Poisson is not satisfied. Overdispersion is the case when the estimated dispersion is higher than the theoretical value of 1. Conversely, underdispersion occurs when it is under 1. We place our focus in the models that account for overdispersion because underdispersion usually occurs when there are excess amounts of zero, a scenario that rarely occurs in our data. When the observed underdispersion is not due to zeros, a simply quasi–Poisson can be applied when the counts are not large and is usually a sign of overfitting by the model or the data being highly predictable.

A regression based approach (Cameron and Trivedi, 1990) can be used to test for both under and overdispersion. Let α be a latent variable acting on $h(\hat{y})$ (usually equal to $\hat{y}$ or $\hat{y}^2$, the quasi–Poisson and negative binomial model respectively) that accounts for the extra variance, we test the hypothesis

$$\begin{aligned} H_0 &: \mathrm{Var}(y) = \hat{y} \\ H_1 &: \mathrm{Var}(y) = \hat{y} + \alpha h(\hat{y}) \end{aligned} \tag{4.31}$$

The alternative model H_1 can also be written as $(y-\lambda)^2 = y - \alpha h(\hat{y}) + \varepsilon$ where $\log(\lambda) = \mathbf{x}^\top \hat{\boldsymbol{\beta}}$ is simply the MLE prediction. This is a least squares with α as regression coefficient without an intercept where a standard t–test applies. The alternative model (4.31) also represents the general case of heterogeneity (4.32) where the extra variability u_i of y_i is unaccounted for by our covariates $\mathbf{x}_i$.

$$Y \sim \mathcal{P}(\tilde{\lambda}), \quad \tilde{\lambda} = \lambda U \tag{4.32}$$

4.5.1 Quasi–Poisson and Negative Binomial

The most simple case of overdispersion (4.32) is when $u = (1+\alpha)$ which correspond to $h(\hat{y}) = \hat{y}$ in (4.31). This is referred to as the quasi–Poisson because the variance differs by a factor of α relative to the standard Poisson. Some authors (Cameron and Trivedi, 1998) also refer them as NB1 model as α is acting on $\hat{y}$, whereas NB2 models refer to the case $\alpha\hat{y}^2$. The most well known case of the NB2 model is the negative binomial, also known as the Poisson–Gamma (PG) mixture (Cameron and Trivedi, 1998, sec. 4.2.2) where the rate λ is distributed according to a Gamma distribution.

With the introduction of covariates, the Poisson–Gamma is usually expressed in the form of (4.33), where the Gamma distribution $X \sim \mathcal{G}a(a,b)$ is parametrized with shape a and rate b such that $\mathbb{E}(X) = ab^{-1}, \mathrm{Var}(X) = ab^{-2}$.

$$\begin{aligned} y_i &\sim \mathcal{P}(\lambda_i u_i) \\ \log(\lambda_i) &= \mathbf{x}_i^\top \boldsymbol{\beta} \\ u_i &\sim \mathcal{G}a(\gamma,\gamma) \end{aligned} \tag{4.33}$$

MLE estimation for both the NB1 and NB2 only require a change in the variance function to accommodate the extra variance. Standard estimation procedures for Poisson, described in Section 4.4.1, applies but with an extra parameter α. It is

obtained initially through the estimated dispersion (4.3) as $\alpha = \hat{\phi}^{-1}$, then updates as $\alpha^{(t+1)} = \alpha^{(t)}\hat{\phi}^{(t)}$ at each iteration, until $\hat{\phi} = 1$ or within some suitable tolerance when converged.

In MCMC, the posterior simulation includes the latent variables/missing data as they are also generated at each iteration. This is known as data augmentation as the introduction of "observed" missing data lead to a complete likelihood. For a historical account, see Meng and van Dyk (1997) who also described its connection with the EM–algorithm, the MLE estimation through the generation of the missing data.

Our full set of parameters in the negative binomial model therefore becomes $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{u}, \gamma)$, where $\mathbf{u}$ are the missing values.

$$f(\boldsymbol{\beta}, \gamma, \mathbf{u} \mid y) \propto f_{\mathcal{P}}(y \mid \boldsymbol{\beta}, \mathbf{u}) f_{\mathcal{G}a}(\mathbf{u} \mid \gamma) f_{\mathcal{N}}(\boldsymbol{\beta} \mid \mathbf{b}_0, \mathbf{B}_0) f(\gamma) \tag{4.34}$$

Sampling $\boldsymbol{\beta}$ under a Gaussian prior is the same as a standard Poisson regression using the IRLS formulation as described in Section 4.4.5 as the regression coefficients are orthogonal to the other parameters, $\boldsymbol{\beta} \perp \mathbf{u}, \gamma$. The only difference is the additional conditioning on $\mathbf{u}$ which means that λ_i becomes $\tilde{\lambda}_i = \lambda_i u_i$ for both the gradient and the Hessian.

There is no natural (conjugate) prior for γ. Therefore we use a uniform prior $\mathcal{U}(0, ub_0)$ with ub_0 something large say 1000. Since γ is the reciprocal of α (4.31), a value of 100 corresponds to virtually no dispersion. Sampling the posterior of γ

$$f(\gamma \mid \mathbf{u}) \propto f_{\mathcal{G}a}(\mathbf{u} \mid \gamma) f_{\mathcal{U}}(\gamma \mid 0, ub_0) \tag{4.35}$$

has to resort to either a random walk or slice sampling (Neal, 2003) as γ exists in both parameters of the gamma distribution. Given that it is univariate and bounded by $0, ub_0$, slice sampling appears to be a natural option because tuning the sampler is not a necessity as it only changes the time spent for each iteration and not the convergence.

Elements updates can be performed for the vector $\mathbf{u}$ given that $u_i \perp u_j$. As the gamma distribution is the conjugate prior to the mean parameter λ of a Poisson distribution, we have a Gibbs sampling step as the conditional posterior of $\mathbf{u}$

$$f(\mathbf{u} \mid y, \boldsymbol{\beta}, \gamma) \propto f_{\mathcal{P}}(y \mid \boldsymbol{\beta}, \mathbf{u}) f_{\mathcal{G}a}(\mathbf{u} \mid \gamma) \tag{4.36}$$

is a compound of the Poisson and Gamma distribution. Let $y \sim \mathcal{P}(\lambda u)$ and $u \sim \mathcal{G}a(a, b)$, then dropping the constants in (4.36) yields

$$\begin{aligned} f(u \mid y, a, b, \lambda) &\propto \frac{(\lambda u)^y e^{-\lambda u}}{y!} \frac{b^a (u)^{a-1} e^{-bu}}{\Gamma(a)} \\ &\propto u^{y+a-1} e^{-(b+\lambda)u}, \end{aligned}$$

which means that the next iteration of the dispersions can be generated by

$$u_i^{(t+1)} \sim \mathcal{G}a(y_i + \gamma^{(t)}, \lambda_i^{(t)} + \gamma^{(t)}) \quad \forall\, i. \tag{4.37}$$

given that $a = b = \gamma$.

4.5.2 Poisson–Lognormal (PLN)

Although the extra variance in the form $\alpha\hat{y}^2$ is usually associated with the Poisson–Gamma model, it is in fact a general result, see Appendix C.2. Therefore, the heterogeneity can be of another form such as the log–normal distribution $U \sim \log \mathcal{N}(0, \sigma^2)$. Because the covariates are linked with λ under a log function, this has a very intuitive interpretation as

$$\begin{aligned} y_i &\sim \mathcal{P}(\tilde{\lambda}_i) \\ \log(\tilde{\lambda}_i) &= \mathbf{x}_i^\top \boldsymbol{\beta} + v_i \\ v_i &\sim \mathcal{N}(0, \sigma^2). \end{aligned} \tag{4.38}$$

Using the moments derived by Bulmer (1974)

$$k^{\text{th}}\text{moment} = \exp\left\{ k\mu + \frac{k^2}{2}\sigma^2 \right\}, \tag{4.39}$$

the expectation and variance for the marginals of Y is readily available (by plugging it into (C.3) and (C.4))

$$\begin{aligned} \mathbb{E}(Y \mid X) &= \lambda e^{\frac{\sigma^2}{2}} \\ \text{Var}(Y \mid X) &= \lambda e^{\frac{\sigma^2}{2}} + \lambda^2 e^{\sigma^2}(e^{\sigma^2} - 1). \end{aligned}$$

So a Poisson–Lognormal is only equivalent to a standard Poisson when the Lognormal distribution is degenerate, i.e. $\sigma^2 = 0$, and it can only account for overdispersion like the negative binomial. Using a Gaussian assumption increases the flexibility of the model such as including an autoregressive structure (Section 4.5.3) or

extends to the multivariate case (Section 4.5.4). But this flexibility comes at a cost because using a Gaussian assumption means that there is no close form estimation for (4.38) and its extension. Therefore simulation based technique like simulated moments (Gouriéroux and Monfort, 1997), or more complicated techniques such as importance sampling or MCMC are usually used.

When only the first two moments are specified, the variance of the dispersion $\mathbf{u}$ can be estimated using a quasi–likelihood (McCullagh, 1983) which can also incorporate a first order serial correlation, i.e. $v_t = \rho v_{t-1}$ (Zeger and Qaqish, 1988). Special cases arise when the linear predictor of an exponential family is of a Gaussian nature; a random effect model such as (4.38) or more generally a linear mixed model (McCulloch et al., 2008) or a state space model when autocorrelation exist (Durbin and Koopman, 2001).

The estimation process for the Poisson–Lognormal is similar to those described in Section 4.5.1 for the Poisson–Gamma and the same procedure can be used for $\boldsymbol{\beta}$. The variance parameter σ^2 that governs the dispersion $\mathbf{v}$ can be sampled easily if a conjugate prior is used. Alternatively, we can assume that the overdispersion is of reasonable magnitude and place a uniform prior with some upper bound ub_0 on the standard deviation

$$\sigma \sim \mathcal{U}(0, ub_0). \tag{4.40}$$

An upper bound of say 2 is suitable while preserving a proper posterior. This is because $\sigma = 2$ corresponds to an overdispersion of $\alpha \approx 3000$ in the y^2 term (4.31), while $\alpha = 4.67$ when $\sigma = 1$. So an upper bound of $ub_0 = 2$ is relatively large, given that $\sigma > 1$ brings the suitability of the model into question. Obtaining a sample from (4.40) is simply $(\sigma^2)^* \sim s^2/\chi^2(n-1)$ where $s^2 = \sum_{i=1}^{n} v_i^2$ subject to satisfying $(\sigma^2)^* \leq ub_0$ imposed by the prior bounds (Appendix C.3.2).

The same sampling scheme for each element of $\mathbf{v}$ is the same as $\boldsymbol{\beta}$ as $v_i \perp v_j$ for $i \neq j$. We use the IRLS sampling scheme where the parameters corresponding proposal for $\mathbf{v}_i$ (4.41) is found by swapping $\mathbf{v}$ and $\boldsymbol{\beta}$ in (4.18), i.e. let $offset_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ and use (4.10, 4.11). Again, this can also be achieved by differentiating the log–likelihood directly and obtain the gradient and Hessian (C.7, C.8).

$$v^{(t+1)} \sim \mathcal{N}(\mathbf{c}(v^{(t)}), \mathbf{H}^{-1}(v^{(t)})), \qquad \begin{aligned} \mathbf{c}(v_i^{(t)}) &= \mathbf{H}^{-1}(v_i^{(t)})\mathbf{W}(v_i^{(t)})\mathbf{z}_i(v_i^{(t)}) \\ \mathbf{H}(v_i^{(t)}) &= \mathbf{W}(v_i^{(t)}) + (\boldsymbol{\Sigma}^{(t)})^{-1} \end{aligned} \tag{4.41}$$

4.5.3 Poisson–Lognormal With Autoregressive (PLNAR)

As mentioned previously, the use of a Gaussian assumption in the latent noise $\mathbf{v}$ allows it to extend the standard Poisson by introducing correlation in $\mathbf{v}$. One of them is serial correlation such as the AR(1) when $\mathbf{y}$ is a time series. To make it explicit that our observations come from a time series, we write (4.38) with T number of observations as

$$y_t \sim \mathcal{P}(\tilde{\lambda}_t), \quad \log(\tilde{\lambda}_t) = \mathbf{x}_t^\top \boldsymbol{\beta} + v_t, \quad t = 1, 2, \ldots, T \tag{4.42}$$

where the dispersions are

$$v_t \mid v_{<t} \sim \mathcal{N}(\phi v_{t-1}, \sigma^2) \;\text{ for }\; t > 1 \tag{4.43}$$

and the first one being

$$v_1 \sim \mathcal{N}\left(0, (1-\phi^2)^{-1}\sigma^2\right). \tag{4.44}$$

A benefit of including an autoregressive term is that future predictions are also driven by the dispersion. It is commonly used with time varying covariates to make multi step ahead predictions in a time series.

The serial dependence increases the complexity of the estimation and requires more advanced simulation based methods such as a Monte Carlo EM (Chan and Ledolter, 1995), importance sampling (Jung et al., 2006) or MCMC (Yu and Meng, 2011). Although the autoregressive process can be extended to an arbitrary, of p order, only the estimation of an AR(1) is of interest here as explained later in Section 4.6.

Sampling of the dispersion is similar to the Poisson–Lognormal, but now the conditional posterior is

$$f(v_t \mid v_{-t}, \phi, \sigma^2, \boldsymbol{\beta}, y_t) \propto f_{\mathcal{P}}(y_t \mid v_t, \phi, \sigma^2, \boldsymbol{\beta}) f_{\mathcal{N}}(v_t \mid v_{-t}, \phi, \sigma^2), \tag{4.45}$$

where v_{-t} is the vector of $\mathbf{v}$ without the element v_t. The conditionals are not the same as (4.43, 4.44) because the conditional of v_{t+1} and v_{t+2} is dependent on v_t and v_{t+1} respectively. So a p order autoregressive model (4.45) requires conditioning on both sides that involves terms up to v_{t-p}, v_{t+p}. Although the conditionals of the dispersions are now different to the Poisson–Lognormal, the marginals of the dispersion are $v_t \sim \mathcal{N}(0, \delta^2)$ where $\delta^2 = (1-\phi^2)^{-1}\sigma^2$ and is equal for all t. The corresponding full joint distribution of $\mathbf{v}$ is a multivariate normal with zero

mean and covariance matrix with diagonals as δ^2 and autocovariance in the off–diagonals. The conditionals can then be obtained from the full joint distribution (see Appendix C.4) through a double sided conditioning up to the p^{th} order, which is 1 in this case. For $t = 2, \ldots, T-1$, the conditional prior of the dispersions are

$$v_t \sim \mathcal{N}\left(\frac{\phi(v_{t-1} + v_{t+1})}{1+\phi^2}, \frac{\sigma^2}{1+\phi^2}\right) \tag{4.46}$$

where the first and last dispersion are

$$v_1 \mid v_2 \sim \mathcal{N}(\phi v_2, \sigma^2), \quad v_T \mid v_{T-1} \sim \mathcal{N}(\phi v_{T-1}, \sigma^2), \tag{4.47}$$

and they represent the $f_{\mathcal{N}}(\cdot)$ contribution in (4.45) where the Poisson part $f_{\mathcal{P}}(\cdot)$ is simply

$$f_{\mathcal{P}}(y_t \mid v_t, \phi, \sigma^2, \boldsymbol{\beta}) \propto -\exp\left\{x_t^\top \boldsymbol{\beta} + v_t\right\} + y_t v_t.$$

Updates are performed by making single moves through v_1 to v_T, but this can equally be done in the reverse order by going from v_T to v_1. Although they can also be performed using the IRLS update (4.41), there is no speed advantage using the IRLS over a t–distribution proposal. This is due to the fact that the v_t's are only single dimension variables, unlike $\boldsymbol{\beta}$ and $\mathbf{v}$ in the PLN model (4.38) which updates as a block.

Both the autocorrelation and the variance parameter only depend on $\mathbf{v}$. So the variance σ^2 can be sampled using the same uniform prior with an upper bound like the Poisson–Lognormal. As $\phi \in (-1, 1)$ is a necessary condition for an AR(1) process to be stationary, a flat (proper) prior is a suitable (and commonly used) choice instead of a normal distribution. We sample the variance using a Gibbs step

$$(\sum_{t=2}^{T}(v_t - \phi v_{t-1})^2 + (1-\phi^2)^{-1}v_1^2)\sigma^2 \sim \text{inv–}\chi^2(T-1).$$

while using a Metropolis–Hastings step for ϕ by first generating ϕ^* via

$$\phi^* \sim \mathcal{N}(\hat{\phi}, \hat{s}^{-1}\sigma^2)\mathbf{1}\left\{\phi^* \in (-1,1)\right\}, \quad \hat{\phi} = \hat{s}^{-1}\sum_{t=2}^{T} v_t v_{t-1}, \quad \hat{s} = \sum_{t=1}^{T-1} v_t^2,$$

and accept with probability

$$\min\left\{\frac{\mathcal{N}(\mathbf{v}_1 \mid 0, (1-(\phi^*)^2)^{-1}\sigma^2)}{\mathcal{N}(\mathbf{v}_1 \mid 0, (1-(\phi^{(t)})^2)^{-1}\sigma^2)}, 1\right\}.$$

For the details and the origin of these equations, see Appendix C.4.2.

4.5.4 Poisson–Multivariate Lognormal (PMLN)

When the dispersion takes the form of a normal distribution, it can be easily extended to the multivariate case. A multivariate Poisson response of k dimension (Aitchison and Ho, 1989) is written as

$$\begin{aligned} \mathbf{Y}_{i,j} &\sim \mathcal{P}(\tilde{\lambda}_{i,j}) \\ \log(\tilde{\lambda}_{i,j}) &= \mathbf{x}_i^\top \boldsymbol{\beta}_j + v_{i,j} \\ \boldsymbol{v}_i &\sim \mathcal{N}_k(0, \boldsymbol{\Sigma}) \; \forall \; i. \end{aligned} \tag{4.48}$$

This was used by Chib and Winkelmann (2001) to model health care utilization and airline incidents, who also investigated the scenario when the dispersion $\mathbf{v}_i$ follows a multivariate t distribution. We simplify the notation here by assuming that the same design matrix $\mathbf{X}$ is used for all k response. Let $\sigma_{i,j} = \boldsymbol{\Sigma}_{i,j}$, the correlation between the response i, j is

$$\text{Corr}(Y_i, Y_j) = \mathbb{E}(Y_i)\mathbb{E}(Y_j)(\exp\{\sigma_{i,j}\} - 1). \tag{4.49}$$

Alternatively, modeling count data with multiple responses can also be formulated as a multivariate Poisson (Johnson et al., 1997), where the correlation structure is built using an additional independent Poisson. For the simplest bivariate case, $(Y_1, Y_2) \sim BiP(\lambda_1, \lambda_2, \lambda_0)$, it has a latent variable representation with three independent Poisson $Z_i \sim \mathcal{P}(\lambda_i), i = 0, 1, 2$,

$$\begin{aligned} Y_1 &= Z_1 + Z_0 \\ Y_2 &= Z_2 + Z_0 \end{aligned}$$

such that $Z_0 = \text{Cov}(Y_1, Y_2)$. This allows the introduction of covariates to the latent variable Z_0 that may reveal information regarding the correlation between Y_1, Y_2. Jung and Winkelmann (1993) performed maximum likelihood estimation on a bivariate Poisson using Newton's method. Tsionas (1999) implemented the Bayesian version using MCMC, the MLE equivalent via the EM–algorithm (Karlis and Ntzoufras, 2003) have also been used. A more flexible covariance structure was implemented by Karlis (2003) through the introduction of additional latent Poisson into the model. But the problem is the restriction that the Poisson random variable $Z_0 \geq 0$ by definition, which means that it lacks the ability to model

negative correlation unlike (4.49). Similarly, the same is true for the multivariate Poisson–Gamma mixture where it can only model a positive correlation (Nelson, 1985; Schmidt and Rodriguez, 2011).

The correlation contribution from (4.49) does not only depend on the standard deviations. It exhibits a negative correlation when either $\sigma_{i,j} < 0$ or $\mathbb{E}(Y_i)\mathbb{E}(Y_j) < 0$, but because $\sigma_{i,j}$ depends on the size of σ_i and σ_j, this model relies on the data being overdispersed as well as modeling $\boldsymbol{\Sigma}$ accurately. The conjugate prior for the covariance of a normal distribution is the inverse–Wishart distribution, but is well known to have problems; the impact on the correlation due to the strength of the variance (Barnard et al., 2000), and the bias that it introduce (O'Malley and Zaslavsky, 2008). Therefore we use the separation strategy of (Barnard et al., 2000) and model the covariance as

$$\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\sigma})\mathbf{R}\operatorname{diag}(\boldsymbol{\sigma})$$

by assuming independence between standard deviation and correlation with separable independent prior $f(\boldsymbol{\sigma}, \mathbf{R}) = f(\boldsymbol{\sigma})f(\mathbf{R})$. Unfortunately, $\boldsymbol{\sigma}$ cannot be sampled using the schemes mentioned previously in Section 4.4.5 when the same prior $\boldsymbol{\sigma}_j \sim \mathcal{U}(0, ub_0)\ \forall j$ is used. This is because the conditional

$$f(\boldsymbol{\sigma}, \mathbf{R} \mid \mathbf{v}) \propto f_{\mathcal{N}}(\mathbf{v} \mid \boldsymbol{\sigma}, \mathbf{R})f(\boldsymbol{\sigma})f(\mathbf{R}) \tag{4.50}$$

consist of a multivariate normal and changing an element of $\boldsymbol{\sigma}$ also changes $\boldsymbol{\Sigma}$. We turn to a Metropolis–Hastings update using a uniform random walk, $\sigma_j^* = \sigma_j^{(t)} + \delta s, S \sim \mathcal{U}(-0.5, 0.5)$ with δ being a scaling factor that is adjusted during the burn–in period. For the (non diagonal) elements of $\mathbf{R}$, we employ a proper joint uniform prior $f(\mathbf{R}) \propto 1$ and sample using slice sampling. This was preferred instead of a random walk given that the bounds of the conditional distribution (which is narrower than $\mathbf{R}_{i,j} \in [-1, 1]$, the natural bound) can be obtained analytically (Barnard et al., 2000, sec. 5).

Alternatively, using the knowledge that a k–dimension multivariate normal random variable is generated using a deterministic transformation on k uncorrelated standard normal realizations (Devroye, 1986), we can write the dispersions as

$$\mathbf{v}_i = \operatorname{diag}(\boldsymbol{\sigma})\mathbf{R}^{1/2}\boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}_k(0, \mathbf{I}). \tag{4.51}$$

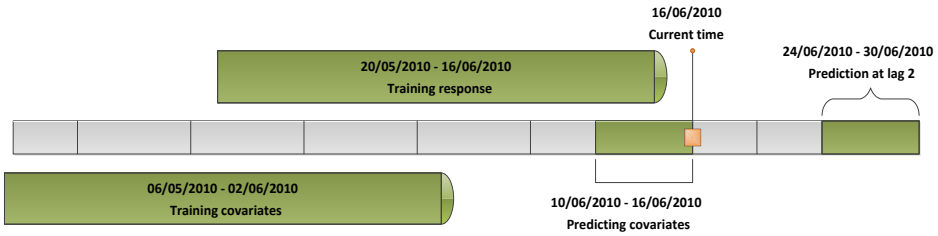


FIGURE 4.2: Example regression for a lag of 2 using lagged exogenous regressors

The joint posterior of $\boldsymbol{\sigma}, \mathbf{R}$ can now be written as

$$f(\boldsymbol{\sigma}, \mathbf{R} \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\xi}) \propto \prod_{i=1}^{n} \prod_{j=1}^{k} f_{\mathcal{P}}(y_{i,j} \mid \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}_j + v_{i,j}\right\}) f(\boldsymbol{\sigma}, \mathbf{R}), \tag{4.52}$$

which consist of the Poisson likelihood without the Gaussian component. These two formulations, (4.52) and (4.50), are referred to as "ancillary augmentation" and "sufficient augmentation" respectively by Yu and Meng (2011) because of the role the missing values $\mathbf{v}$ play in the posterior of $\boldsymbol{\sigma}, \mathbf{R}$. Experimentation (Figure C.3 for example) shows that the parameterization of (4.52) is much more efficient in our model for both $\boldsymbol{\sigma}$ and $\mathbf{R}$.

4.6 Results

We compare the models described previously by comparing their DIC over time as well as their predictive performances. To make predictions for future y_{Join}, y_{Return} more than one week ahead using static regressions, we make use of lagged regressors. We construct a design matrix $\mathbf{X}$ for the q forecasting steps at time t_0 using observations of those in Table 4.2 at $t_0 - j$ where $j = 1, \ldots, q$. So we assume that the number of new users joining at time $t_0 + 1$ is dependent on the covariates observed at time t_0 upon discretization. To produce the j^{th} step ahead prediction given the current observation at t_0, we use the lagged covariates j steps behind t_0 for the model estimation stage and use the current observations $\mathbf{x}_{t_0}$ for prediction, i.e. $y_{t_0+q} = \mathbf{x}_{t_0}^\top \boldsymbol{\beta}$. Therefore $\mathbf{x}_0$ is used for the prediction of all q steps. This is demonstrated in Figure 4.2 where we use covariates that are 2 lags behind the current observed time t_0, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t_0-2})$ against $\mathbf{y} = (y_{1+2}, y_{2+2}, \ldots, y_{t_0})$ to estimate our model, the prediction for the number of users joining and returning at $t_0 + 2$.

The variables in Table 4.2a were selected because they all seem to be naturally linked to how many users may join a community: the number of active users reflect

# of active users	
# of active threads	
# of messages	
# of new threads created	
# of points awarded	

(A) Lagged Covariates

$\cos(2\pi t/(52/2))$	$\sin(2\pi t/(52/2))$
$\cos(2\pi t/(52/4))$	$\sin(2\pi t/(52/4))$
$\cos(2\pi t/(52/8))$	$\sin(2\pi t/(52/8))$
$\cos(2\pi t/(52/13))$	$\sin(2\pi t/(52/13))$
$\cos(2\pi t/(52/26))$	t

(B) Time Covariates

TABLE 4.2: Variables thought to be useful in predicting the number of users moving in and out of communities

the size of the community and a larger community tends to attract people (non users), the number of active threads and messages posted shows interest level, and new threads represent fresh information which other people (non users) may find useful. Alternatively, we can use a set of covariates that depends on time (Table 4.2b). Let $t = (1, 2, \ldots, T)$ be a vector with length equal to T number of observations, then our time covariates consist of t and trigonometry functions $\cos(2\pi t/(52/i)), \sin(2\pi t/(52/i))$ for a set of suitable i, say $26, 13, 8, 4, 2$ that enable us to model possible seasonality in the data over both the long and short term. Obviously, $\sin(2\pi t/2)$ is not suitable because it is equal to 0 for all t and it was not used.

4.6.1 Quality of Fitted Models

First, we compare the DIC between methods and the two proposed sets of covariates. Three different versions of the prior were tested: a fixed prior (4.26) with $\sigma^2 = 100$, the variance σ^2 inferred via (4.27) is denoted as L_2 and the Laplace prior (4.22) with λ estimated via (4.28) denoted as L_1. We denote the straight Poisson, Poisson–Lognormal and Poisson–Gamma as P, PLN and PG respectively. The combined DIC of y_{Join} and y_{Return} was also calculated and will be used to compare against the PMLN model later. Before that, we look at the performance of the two sets of covariates, see Table 4.3.

The results were generated over the same 90 week period as seen in Chapter 3 and the average DIC per observation of forum 353 can be seen in Table 4.3. Even though none of the y_{Join} and y_{Return}'s were deemed to be overdispersed when using the lagged covariates, 76 and 74 out of the 90 weeks had $\hat{\phi} > 1$ for y_{Join}, y_{Return} respectively. This is not the case for the time covariates where 31 achieved a p–value lower than 0.05 for y_{Join} for the hypotheses test for overdispersion described in Section 4.5 with 79/90 weeks had $\hat{\phi} > 1$. For y_{Return}, 54 weeks had $\hat{\phi} > 1$ with none not them being statistically significant at the 0.05 level

		Lagged Covariate		Time Covariate		Combined	
Method	Prior	y_{Join}	y_{Return}	y_{Join}	y_{Return}	y_{Join}	y_{Return}
P	Fixed	5.9331	3.7805	6.1818	3.7127	6.1275	3.7943
	L_2	5.9316	3.7699	6.0993	3.6183	6.0241	3.7290
	L_1	5.9296	3.7734	6.0967	3.6260	6.0007	3.7156
PLN	Fixed	5.9427	3.7783	6.1089	3.7381	6.1259	3.8213
	L_2	5.9406	3.7664	6.0320	3.6376	6.0141	3.7480
	L_1	5.9380	3.7629	6.0284	3.6401	6.0001	3.7276
PG	Fixed	5.9273	3.7756	6.1177	3.7132	6.1177	3.7983
	L_2	5.9257	3.7662	6.0429	3.6182	6.0122	3.7299
	L_1	5.9234	3.7621	6.0392	3.6215	5.9918	3.7103

TABLE 4.3: The average DIC per observation over 90 weeks for both the number of users joining and leaving for forum 353 using different methods and priors

	Prior		
	Fixed	L_2	L_1
Lagged Covariates	9.7033	9.7017	9.6972
Time Covariates	9.7445	9.5829	8.9005
Both Covariates	9.8073	9.5848	8.5642

TABLE 4.4: The average DIC per observation averaged over 90 weeks using PMLN that model both y_{Join} and y_{Return} simultaneously with different prior for forum 353

We can see from Table 4.3 that the lagged covariates performed better than the time covariates for y_{Join} and the exact opposite for y_{Return}. Both the regularization methods had lower DIC than a fixed prior, as expected, with L_1 version dominating L_2 in most of the cases. Although PG appears to perform better than PLN, a more detailed look into the individual weeks revealed that neither model dominated the other and it depends on the level of dispersion. When there is significant overdispersion i.e. tested significant at the 0.05 level, then PLN has lower DIC than PG and showing the flexibility of the Gaussian assumption. But whenever $\hat{\phi}$ is lower than ≈ 1.1 then PG provides the better fit. Also, P with regularization is the best when $\hat{\phi} < 1$ in comparison with the two overdispersed model as expected.

Obviously, we can leverage both sets of covariates in Table 4.2 by combining them together. But it failed to perform any better for both y_{Join} and y_{Return}. Furthermore, the fit for y_{Return} at earlier weeks were nearly perfect (see Figure 4.1), and the overparameterization led to convergence issue for the models without regularization. This was because $\text{Var}(\boldsymbol{\beta})$ was extremely large, so the proposals were all very far away from the mode and the current value. Using both set of covariates tended to improve model fit as the number of observations increased, but only marginally.

	Prior		
	Fixed	L_2	L_1
y_{Join}	5.9301	5.9257	5.9218
y_{Return}	3.7422	3.6479	3.6521

TABLE 4.5: The average DIC per observation over 90 weeks using the PLN–AR(1) model under different prior for forum 353

Modeling of both the responses together was also attempted, again with the three different priors. Separate variance parameters for each response were estimated in the L_1 and L_2 setting. We can see that modeling both $\mathbf{Y} = [y_{Join}\ y_{Return}]$ simultaneously using PMLN (Section 4.5.4) yield a lower DIC when compared combining the two separate models, i.e. $\min(y_{Join}) \approx 5.9$ and $\min(y_{Return}) \approx 3.6$ in Table 4.3 and they added up to 9.5. This was also true in some cases where neither of the responses were overdispersed. More importantly, it was not just the combined DIC that got smaller, but also individually for each response suggesting that the correlation should not be ignored.

This would seem rather obvious when looking at the residuals from a standard Poisson, which were correlated nearly all the time. Furthermore, there was a difference between the dispersion estimated under a model under regularization and one without. Since a regularized model is more conservative in the fit, it usually has a higher dispersion (but not guaranteed to be overdispered) than the one obtained from the MLE. As the dispersions were effectively there to model the remaining error not captured by the covariates, an overdispersed model under regularization provided an opportunity for PLN/PG to improve the fit. In the cases where there were significant correlation between the residuals of the two responses, PMLN was able to exploit this artifact and improved even further.

It is unsurprising then that PLN–AR(1) failed to provide a better fit overall as the residuals did not usually exhibit autocorrelation, even though they did in the actual observations. The introduction of the covariates was able to account for the autocorrelation, especially for y_{Return} where it was obviously there exist some sort of seasonality given the impact of the time covariates. The DIC of the PLN–AR(1) model can be seen in Table 4.5, and when looking at the result carefully, it could be observed that the average deviance was virtually the same as PLN but now pD (the effective number of parameter) has increased. Difference between PLN and PLN–AR(1) decreased as the number of observations increased, because the degree of freedom has also increased and the covariates no longer captured the time effects accurately.

P				PMLN				PLN–AR(1)	
Lagged		Combined		Lagged		Combined		Time	
y_{Join}	y_{Return}	y_{Join}	y_{Return}	y_{Join}	y_{Return}	y_{Join}	y_{Return}	y_{Join}	y_{Return}
304.4	66.5	306.9	69.5	319.0	63.6	318.9	64.2	307.6	54.7

TABLE 4.6: The MSE (top) over a 10 weeks ahead forecast averaged over 80 weeks for both the number of users joining and leaving for forum 353 using different methods and priors

4.6.2 Out–of–Sample Performance

For the predictive ability of the models, we compare the MSE for the out–of-samples observations. The mean prediction (sample averages of the inferred parameters) usually does not give the best performance nor is it necessarily representative of the predictive distribution. Instead, we use the median of the simulated predictions as our point prediction (Barbieri and Berger, 2004), where the full set of simulated samples will be inserted into the compartment model in the next chapter.

The PLN–AR(1) model used the time covariates and made a 10 weeks forecast, while the static regression models were trained separately, one for each of the 10 predictions. Only PMLN was tested because PLN is a special case when correlation do not exist. If we used the same time period as Chapter 3 to perform a 10 weeks ahead prediction, the first week at $N = 20$ using the static regression has only 10 observations at lag 10 with 19 at lag 1. This meant that we had more variables than observations when both the lagged and time covariates were used (16 in total), and the predictions were way off the actual observations so we omitted them from the results. Instead, we started at $N = 30$ and ran through the remaining 80 weeks. All the models were inferred under a Laplace prior.

The averaged results can be seen in Table 4.6, with the lagged and combined covariates shown because the time covariates were not competitive at all for y_{Join}. Although the PMLN provided the better fit, the predictive power was not as impressive as demonstrated in Table 4.6. The addition correlation in PMLN improved the predictions of y_{Return} while the reverse was true for y_{Join}. The PLN–AR(1) surprisingly had a good performance for y_{Return} even though autocorrelation was rarely observed but failed to improve y_{Join}.

The time series and density plot of the 80 consecutive predictions can be seen in Figure 4.3 and Figure 4.4 respectively. Figure 4.3 clearly shows that PLN–AR(1) had the lowest averaged MSE and higher number of lowest MSE for y_{Return}. Using

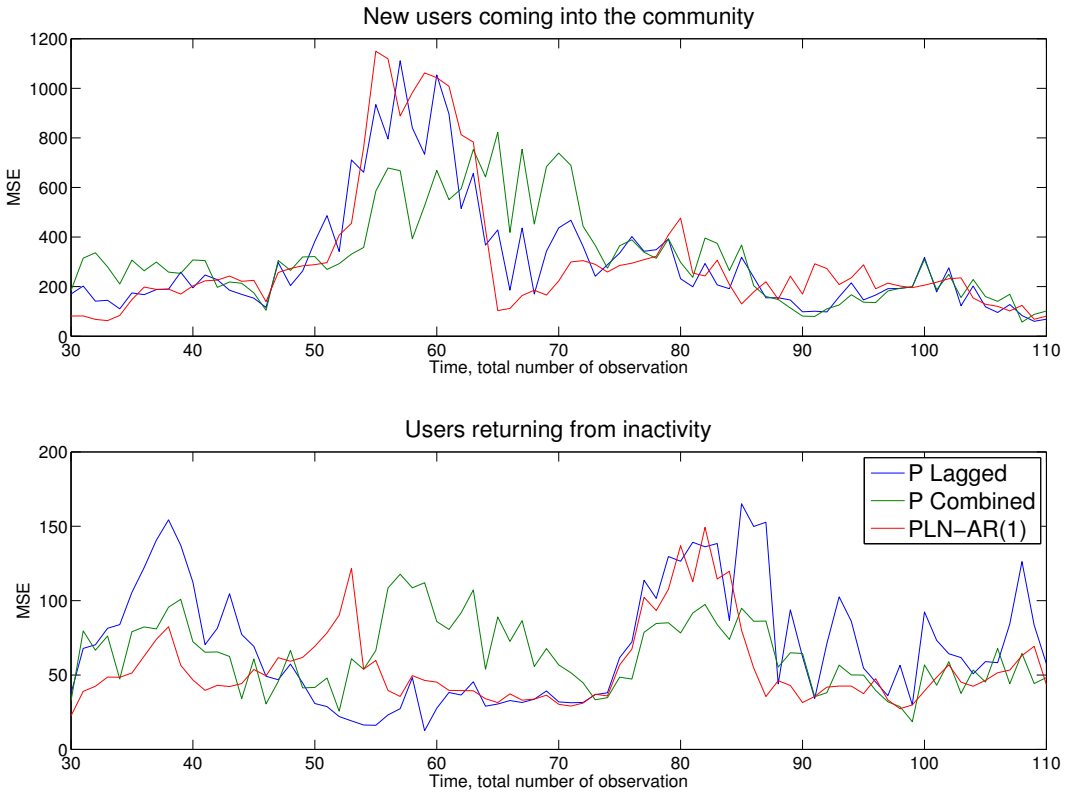


FIGURE 4.3: Time series plot for the out–of–sample MSE for both y_{Join} and y_{Return} over the 80 weeks for forum 353. All three models are under a Laplace prior.

both sets of covariates yielded better prediction apart from a 20 week period starting from week 50. Density plot of the MSE showed that PLN–AR(1) had the lowest variance in addition to the mean. A straight Poisson using the lagged regressors had the lowest MSE in terms of the mode, and the minimum MSE achieved over the 80 forecasts.

On the other hand, there was no clear winner for y_{Join}. The combined set of covariates performed similarly to the lagged covariates for most of the weeks, especially from week 73 onwards where there was virtually no difference. The biggest difference was seen from approximate week 50 to week 73, where the combined set of covariates performed the best from week 50 to week 62, then had the worst performance between the three models. Density plot in Figure 4.4 showed that the lagged covariates had the least variance and the MSE of the modes was smallest of the three. Like y_{Return}, using more covariates again failed to increase the predictive power, where a higher MSE was observed for both the mean and the mode. The almost identical MSE curve at the last quarter of Figure 4.3 suggests that the difference may be due to overfitting, given that the number of observations, hence the degree of freedom, increases as time progresses.

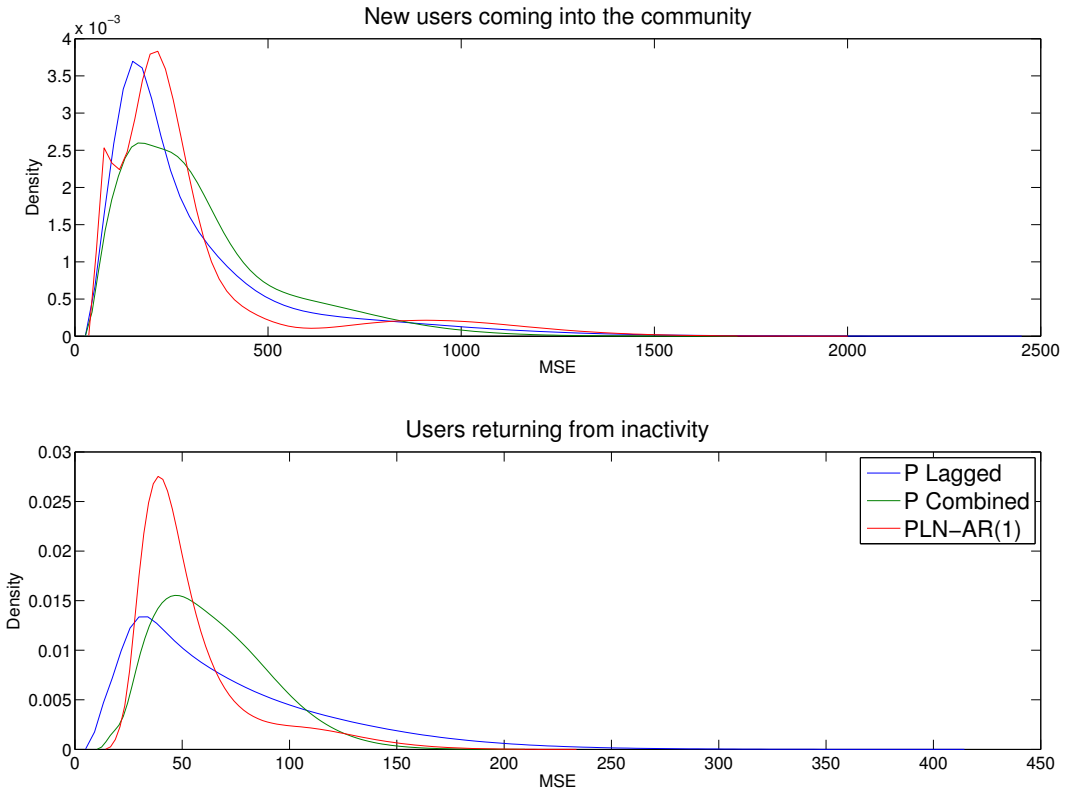


FIGURE 4.4: Density plot for the out–of–sample MSE for both y_{Join} and y_{Return} over the 80 weeks for forum 353. All three models are under a Laplace prior.

4.6.2.1 Summary

Given that there were no obvious autocorrelation structure in the latent variables and the time covariates were not useful in the predictions of y_{Join}, we assumed that the lagged variables will be sufficient for prediction. PLN–AR(1) predicted y_{Return} well where the additional autocorrelation structure increased the predictive power. This could be compared to the migration rates of the compartment model where the total rates going out of the inactive compartment were also autocorrelated. So instead of modeling the autocorrelation in the migration rate, which was problematic because of the constraints, we have switched to modeling the actual user counts.

The PLN–AR(1) model is more flexible and includes the PLN, which is a special case when the autocorrelation coefficient $\phi = 0$. Therefore, the PLN–AR(1) should always be attempted when there exist overdispersion. Introducing the autocorrelation structure effectively overparameterizes the model, and hope that a first order process captures enough useful information if it exists. This approach is similar to classic time series analysis where the order of an AR model is estimated

base on expanding the autocorrelation structure because ϕ_{p+1} has a theoretical value of 0 for an AR(p) model.

Chapter 5

Main Results

5.1 Introduction

In this chapter, we consider the full forecasting problem. That is to say, we make forecast for a finite number of time steps using the compartmental models described in Chapter 3 given the predictions on the number of users joining a community based on the methods in Chapter 4 at the corresponding future time points. The alternative formulation that makes use of the predictions on the number of users returning from inactivity will also be tested and compared against.

From the observations in Section 4.6, predictions of y_{Return} are made either using a straight Poisson or PLN–AR(1) Section 4.5.3, both under a Laplace prior. Whenever there is overdispersion, we use the PLN–AR(1) and a straight Poisson when it is not overdispersed. It is the same for y_{Join} but the PLN model is used instead of PLN–AR(1) because the lagged covariates were better suited. For the deterministic models, we use the median of the predictive distribution (4.25) while drawing a sample at each updating step (3.21) of the forecast for the stochastic models. The realization of a stochastic transition matrix $\mathbf{P}$ is the same as Section 3.6, which is based on one of the models introduced in Section. 3.5.1.2, 3.5.1.3, 3.5.2.

We also introduce another stochastic baseline here, where forecasts are made by drawing from the empirical distribution of the transition matrices uniformly. This is equivalent to **SP** (Section 3.4) when $\alpha = 1$, we denote this baseline as **SE**.

Method	**W**	**I**
DL	177	175
DE	45	73

(A) Forum 353

Method	**W**	**I**
DL	94	95
DE	29	26

(B) Forum 256

Method	**W**	**I**
DL	256	263
DE	72	122

(C) Forum 264

		Forum 353		Forum 256		Forum 264	
		t_0		t_0		t_0	
Method	Formulation	$t_0 = 1$	$N - 10$	$t_0 = 1$	$N - 10$	$t_0 = 1$	$N - 10$
PEL	**W**	60	52	36	28	107	83
	I	53	135	31	25	79	262
PENL	**W**	65	56	52	29	91	100
	I	64	58	45	26	124	99
DPL	**W**	44	52	28	25	74	75
	I	65	51	24	23	95	81
DPNL	**W**	45	50	28	24	72	73
	I	69	51	24	23	104	84
DWL	**W**	43	52	29	26	71	79
	I	55	53	27	26	82	87
DWNL	**W**	44	54	32	27	70	85
	I	54	54	34	25	83	92

(D) MSE under different estimated methods

TABLE 5.1: Average MSE of the out–of–sample forecast error over 90 consecutive weeks for all methods, each with a 10 weeks ahead forecast for three different forums. Both formulations of the inactive users were used with initial observation t_0 in the estimation at either $\mathbf{m}(1)$ or $\mathbf{m}(N-10)$, where N is the total number of observed mass.

5.2 Results

The result of the deterministic formulations can be seen in Table 5.1 and the stochastic formulations in Table 5.2. The baselines outperformed nearly all the proposed models under the **W** formulation, but most of the proposed model performed better than the baseline under the **I** formulation.

When $t_0 = N - 10$, MSE under both the **W** and **I** formulation were very similar through time, an example that shows the MSE through time for **SP** can be seen in Figure D.1. For both the weighted and the penalized method, the quality of forecasts by the **I** formulation had effectively been pulled towards the **W** formulation, such that the error are approximately the same, if not better.

As expected, both the baseline and our forecast have higher MSE than Table 3.2 where the actual observations of y_{Join}, y_{Return} were used. The baseline MSE were ≈ 1.5 times higher when predictions of the external factos were used in (Table 5.1)

		Forum 353		Forum 256		Forum 264	
Method	Formulation	MSE	Coverage	MSE	Coverage	MSE	Coverage
SE	**W**	42.81	95	30.37	94	73.77	94
	I	73.84	93	22.25	98	124.65	93
SB	**W**	43.20	97	26.58	95	75.08	90
	I	72.94	96	26.75	95	122.32	89
SG	**W**	43.36	91	26.53	94	348.48	95
	I	73.04	91	26.04	94	394.24	96
SP	**W**	47.58	93	24.44	90	76.55	92
	I	50.74	90	16.26	98	84.09	91
SW	**W**	55.94	90	25.49	95	182.85	74
	I	64.02	90	27.23	93	120.86	71

TABLE 5.2: Summary of error over 90 consecutive weeks for the three stochastic methods as described in Section 3.5, each with a 10 weeks ahead forecast for three different forums.

place of the actual observations (Table 3.2). But the affect of the external factors depends on the proposed model, i.e. the method of obtaining $\hat{\mathbf{P}}$.

In the deterministic approach, the MSE increased consistently when the predicted external factors are used when $\hat{\mathbf{P}}$ are estimated based on either the weighted or penalized method. The biggest change in MSE was found in **PEL** and **PENL**, with the former even yielded an overall lower MSE such as the case for forum 264 under the **W** formulation.

Modeling the inactive users in a compartment usually results in a lower MSE for the stochastic models (Table 5.2) and not for the deterministic models (Table 5.1). In fact, the MSE by the stochastic model is similar to the deterministic model for all 3 forums even when the forecasts (for each of the roles) are different, i.e. Figure D.3 where the MSE for **DP** and **SP** under the **W** formulation are 104, 116 respectively.

Similar to Section 3.6, **SG** has shown that the forecast can be of poor quality (Forum 264), while the penalized method makes the performance between the two formulations comparable, see Figure D.2. The coverage on the other hand is nearly always higher when using parametric models. We show in Figure 5.5 how the parametric based **SB** compare against the of the empirical models in **SE** and **SP**.

It should be noted that coverage (3.45) is defined to have a target value of exactly 95%. Therefore, any deviation from 95%, even when it is higher, can be considered

Method	Forum 353	Forum 256	Forum 264
W	45	66	50
I	82	73	73

TABLE 5.3: Percentage of times **SP** achieved a lower MSE than **SE** when comparing within each of the formulations, i.e. 45 in forum 353 under **W** indicate that **SP** has a lower MSE for 45% of time

a poor reflection of the real confidence intervals. Figure 5.1, 5.2, 5.3 shows the scatter plot between the MSE and coverage, where the plot for forum 264 have 3 outliers removed, the original can be seen in Appendix D Figure D.4.

In addition to the MSE and coverage, it is also of interest to see the percentage of time a model perform best out of the five proposed, we show this for the three forums in Figure 5.6, 5.7, 5.8. Furthermore, a direct comparison between **SE** and **SP** can be seen in Table 5.3 with the time series plot in Figure 5.4.

The kernel density plot of the raw error, $e = \hat{y} - y$, over all time and the 3 forums for each of the roles can be seen in Figure 5.12, 5.13, 5.14 for the **W** formulation and Figure 5.9, 5.10, 5.11 for the **I** formulation. The mode of both **SE** and **SB** are usually away from 0 regardless of formulation. Differences between the density of **SB** and **SE** are small, while **SP** and **SE** are much greater even though their overall MSE (Table 5.2) are comparable.

In addition to the individual roles, Figure 5.15 shows the raw error between the total number of active users predicted. The top figure, Figure 5.15a, was generated by using the predicted external factors with the bottom figure Figure 5.15b using the forecasts in Section 3.6, i.e. both y_{Join}, y_{Return} are assumed to be observed.

The top figure displays the kernel density when the external factors are predicted while the bottom figure uses the actual observations. Note that the multimodal feature in Figure 5.15a exist in both the **W** and **I** formulation.

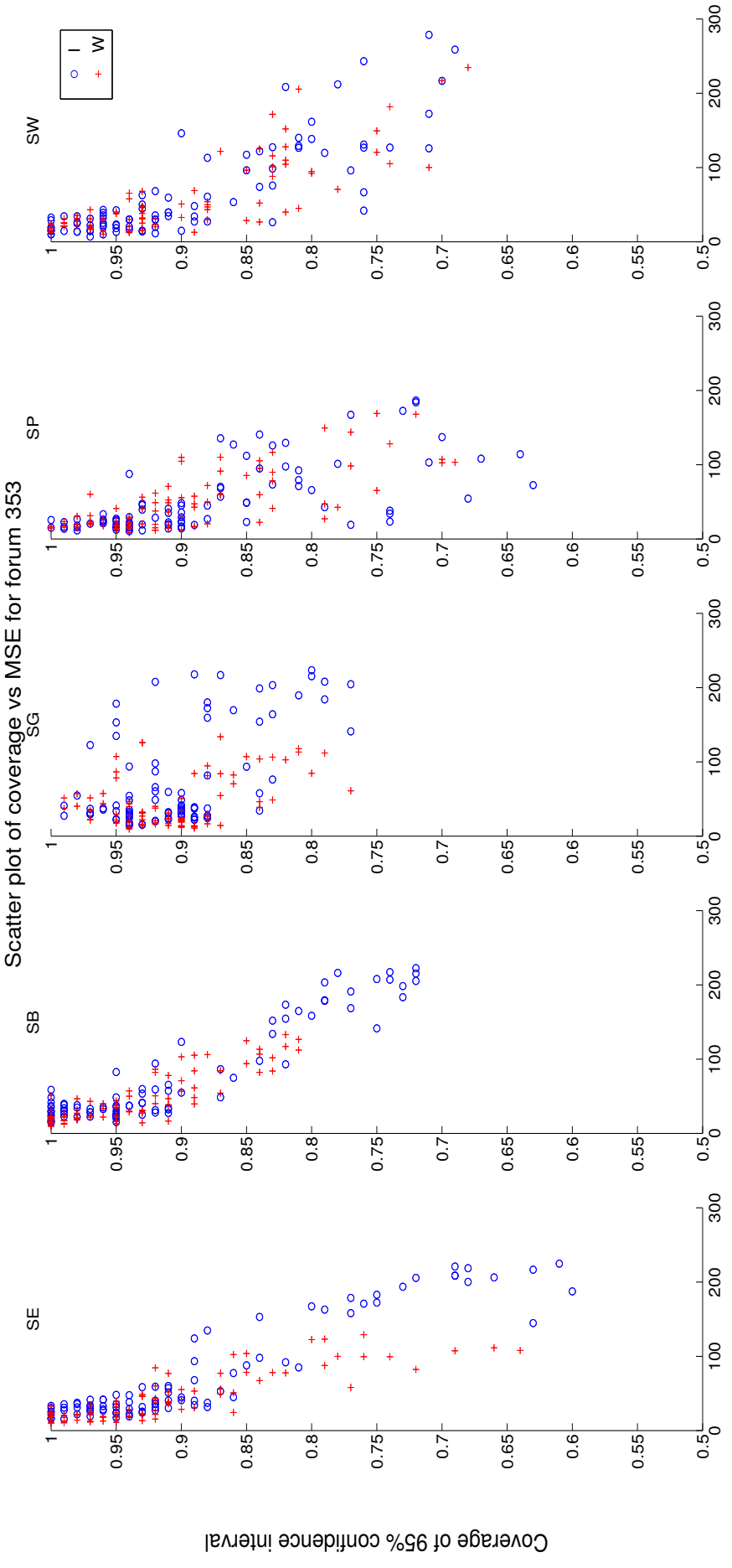


FIGURE 5.1: The coverage against MSE of different methods and formulations over all 90 forecasts for forum 353

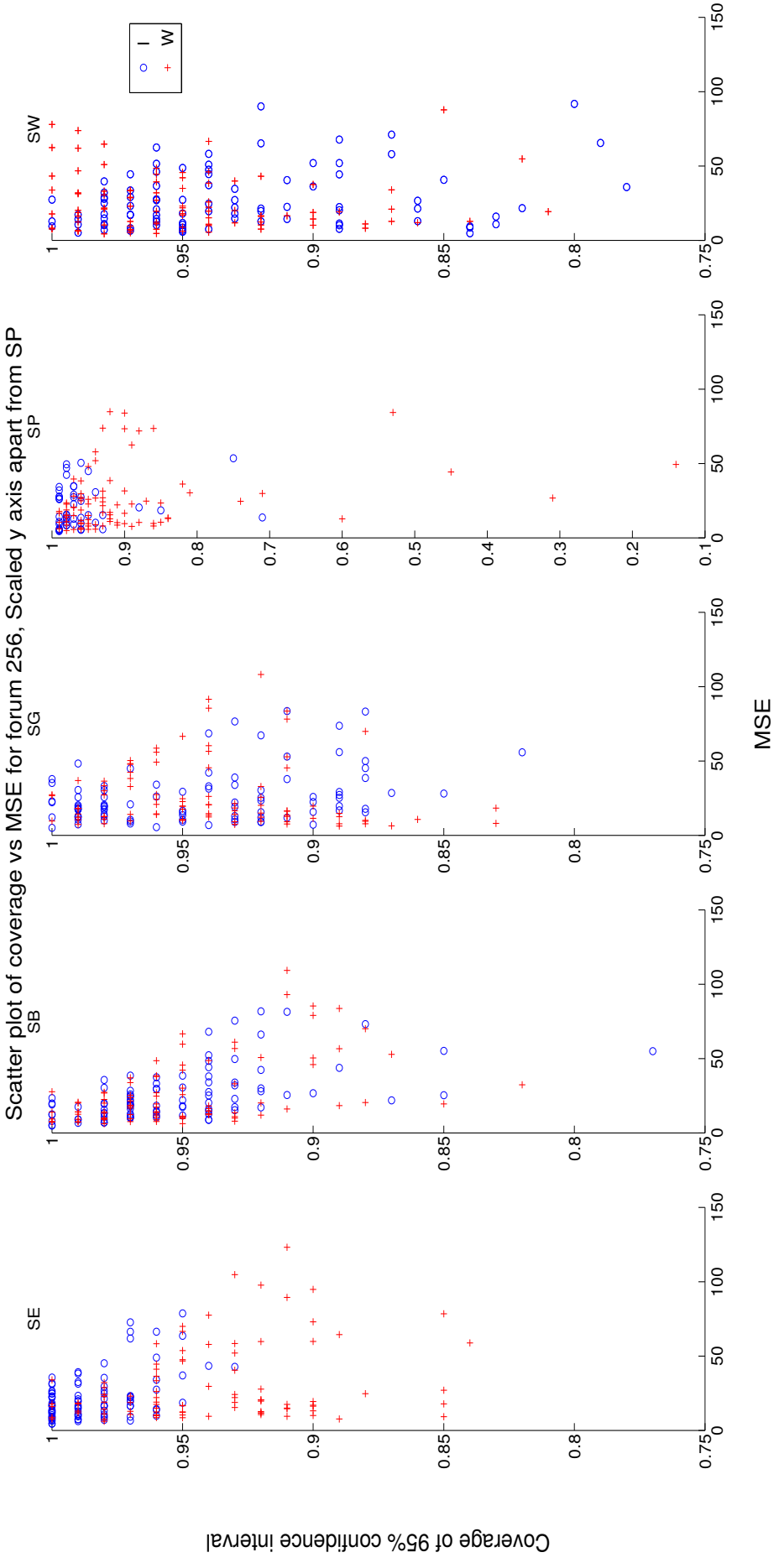


FIGURE 5.2: The coverage against MSE of different methods and formulations over all 90 forecasts for forum 256

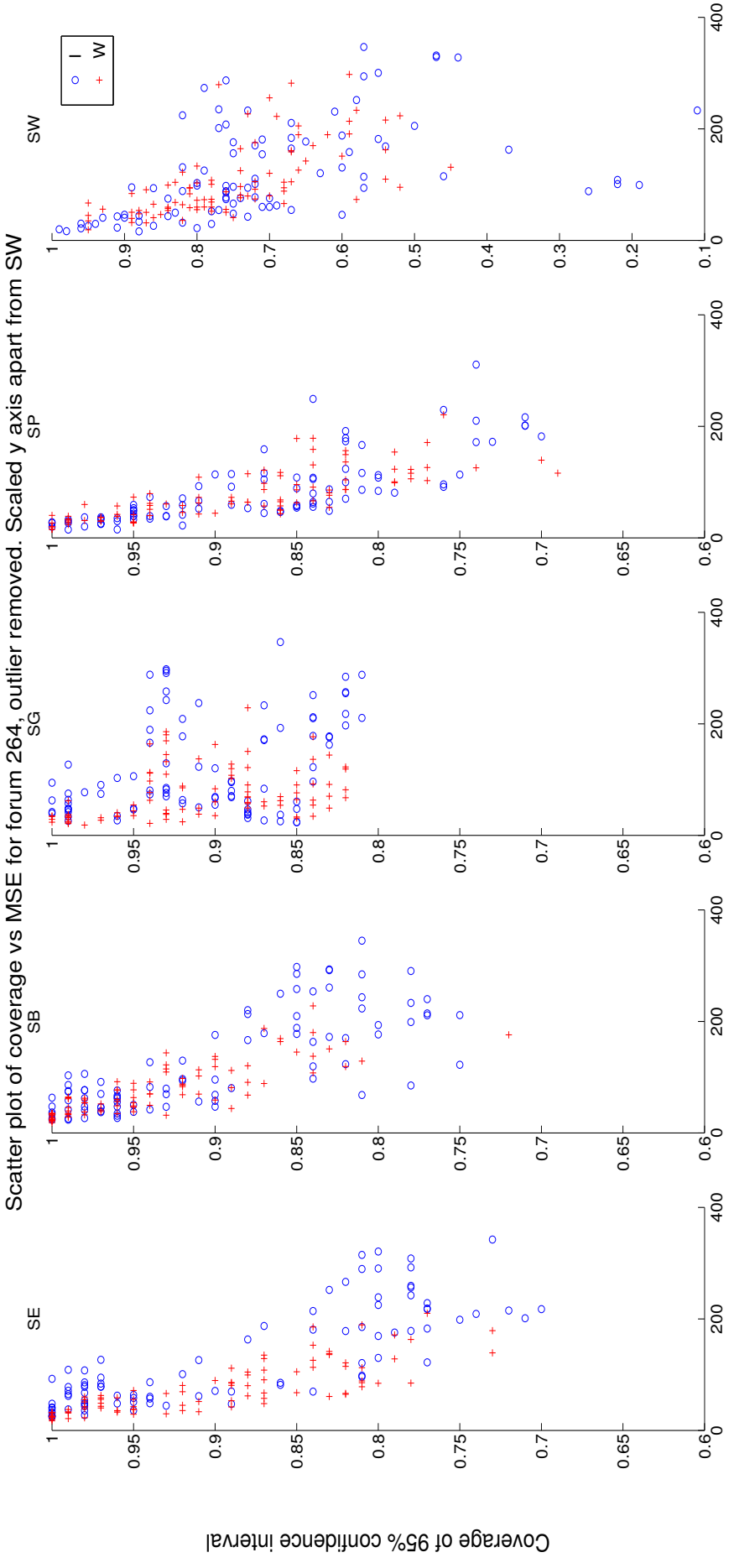


FIGURE 5.3: The coverage against MSE of different methods and formulations over all 90 forecasts for forum 264

5.3 Discussion

Our recommendations are based on the results in the previous section would be to use **SP** when making forecasts based on the observed results. The error generated by **SP** is closest to being symmetric and centered around 0, with a magnitude (of the errors) comparable, if not better than the other stochastic methods. Quality of the forecasts measured by MSE and coverage suggest that the simple empirical approach of **SE** represents the data well when compared to the parametric model of **SB** which requires a lot more computation time. More importantly, the stochastic models provides a set of confidence interval while achieving a similar performance (in terms of MSE) when compared to the deterministic models.

Performance for **SW** is poor for all 3 forums instead of just forum 264 as discussed in Chapter 3. **SP** worked well because the feasible region is constrained between the expected and the last observed transition matrix, with $\alpha = 1$ for the former and $\alpha = 0$ the latter. Both are suitable candidate solutions, instead of **SW** where the optimal set of weight can be placed on observations that minimizes the objective function well but fails to predict accurately. Therefore, we can view **SP** as a safeguard method that prevents overfitting.

However, the coverage by **SP** can be significantly (and usually) lower than **SE** given the penalty α governs not only the mean forecast, but also the width of the confidence interval. Therefore, coverage can be low (< 0.5) as $\hat{\alpha}$ is low enough that the forecast is nearly deterministic as previously mentioned. **SW** suffers from the same problem where neither of the parametric methods do. This does not imply that **SB** guarantees a higher coverage, relative to the empirical methods, even though it is majority of the time as seen in Figure 5.5.

Another benefit is that both the **I** and **W** formulation produced results that were equally good for **SP**. Saving computation time and simplifying the forecast without the prediction of y_{Return}. All other methods on the other hand performs better under the **W** formulation usually have a lower MSE, as evident from Table 5.2, which is expected and consistent with how our models behave.

Recall the model definition from Section 3.2 and note that the total flow going out of the inactive compartment can be interpreted as a regression without intercept (or centered)

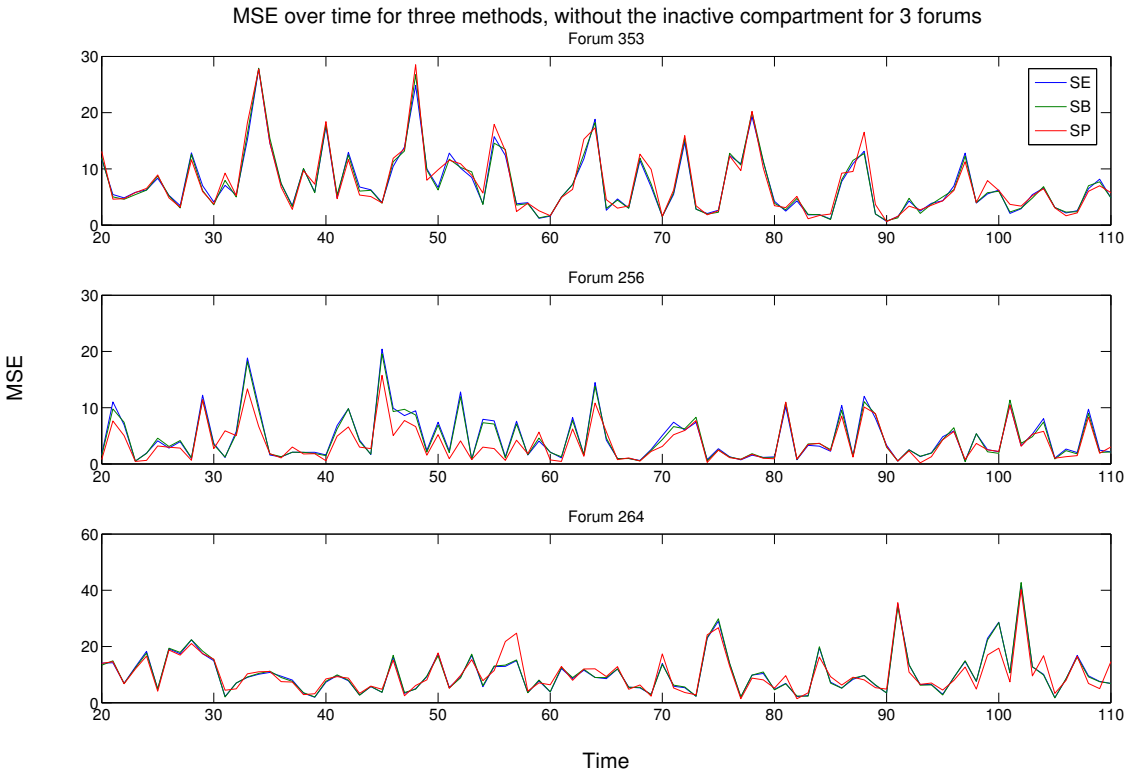
$$\sum_{j=1}^{k} \mathbf{v}_{0,j}(t) = \sum_{j=1}^{k} \mathbf{P}_{0,j}\mathbf{m}_0(t), \tag{5.1}$$

with regression coefficients $\mathbf{P}_{0,j}$'s for all active compartments. Therefore, without the adjustment on the autocorrelation of $\mathbf{P}_{0,j}$'s, the predicted number of users returning is greater than the actual number, with increasing deviation as time progress. This can be observed from the kernel density plots in most of the roles, with most distribution skewed to the right (and greater than zero) for **SE** and **SB**.

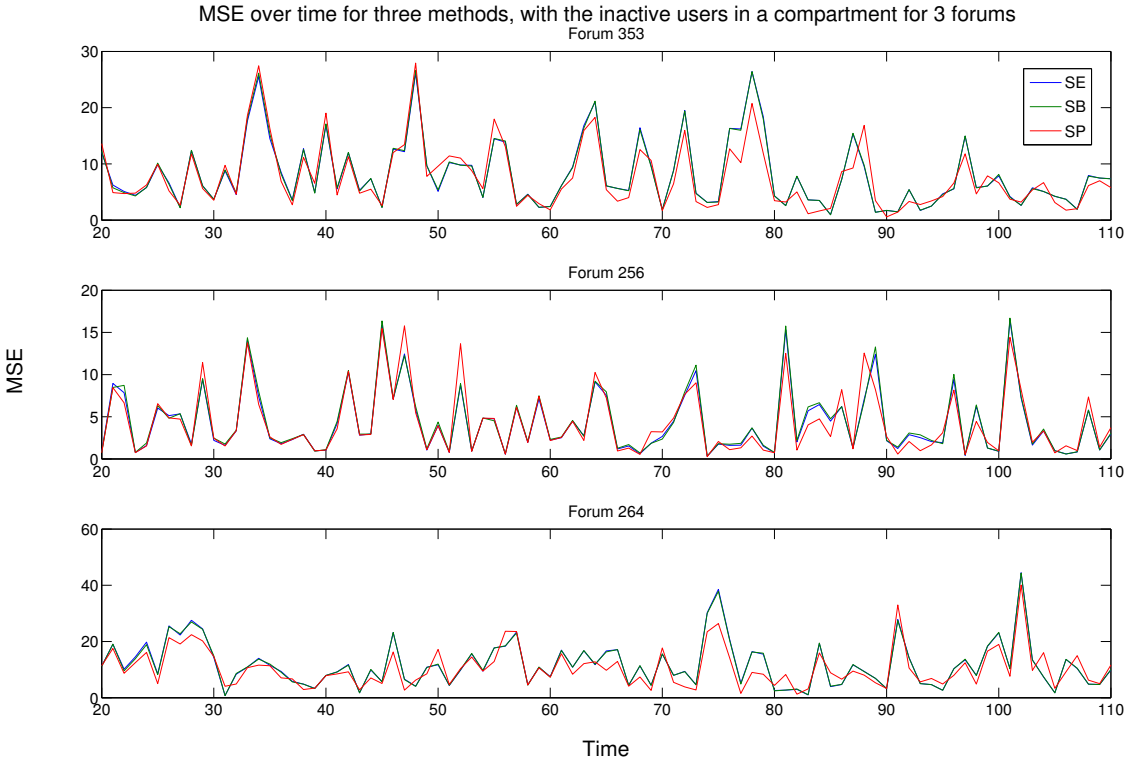
This was further demonstrated in Figure 5.15 which shows the raw error between the predicted and observed total number of active users. The skewness to the right on the kernel density indicate that the models over predict the number of active users much more when under the **I** formulation. The difference between the first and the tenth forecasting step in Figure 5.15b is an illustration of how the over prediction aggregate over time.

Prediction of new users compounded on the over prediction as shown by the difference between Figure 5.15a and Figure 5.15b, with the latter generated using the forecasts in Section 3.6, i.e. both y_{Join}, y_{Return} are assumed to be observed. Given well predicted external factors, it would appear that **SP** have mitigated the over prediction caused by the compartment model. This isolation shows a clear direction in which we can improve our forecasts.

Furthermore, the density curve of the raw error for **SP** has a thicker tail but less skewed than **SE**. It is unsurprising then that the former has a lower error when using the absolute instead of square loss, when averaged over all weeks in each of the 3 forums. Robustness of the **SP** method gives us confidence in saying that it is the best out of all our proposed options.

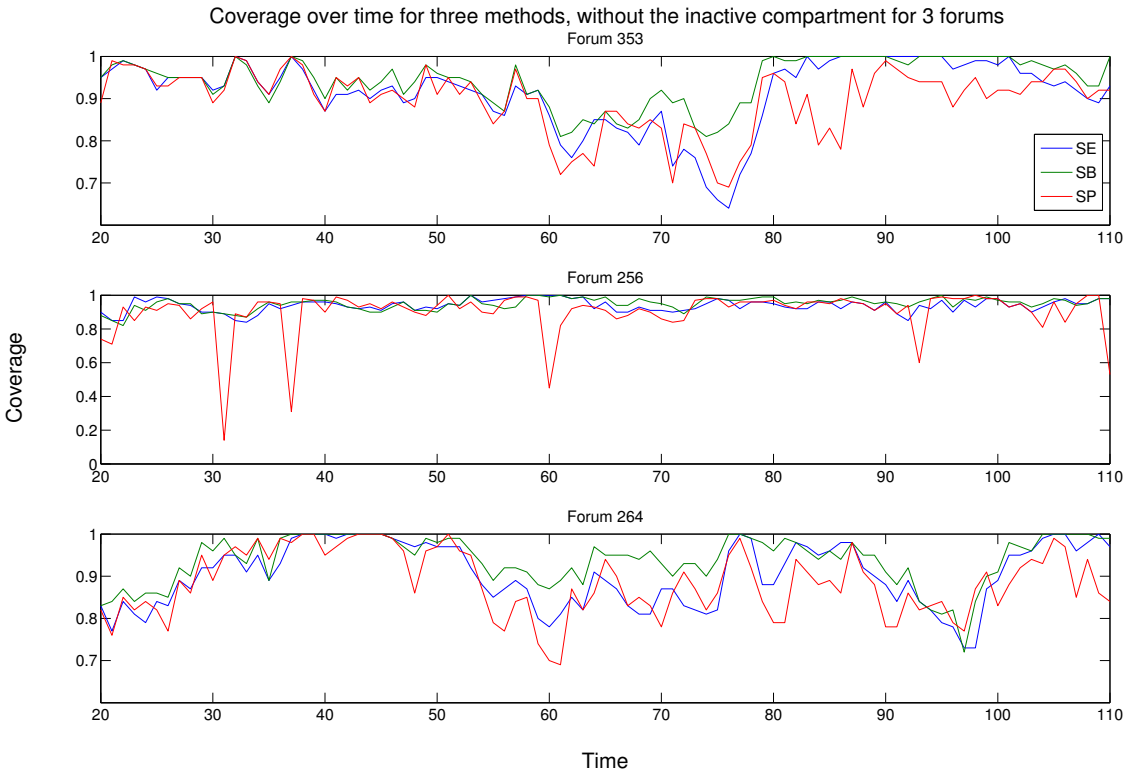


(A) Under the **W** formulation

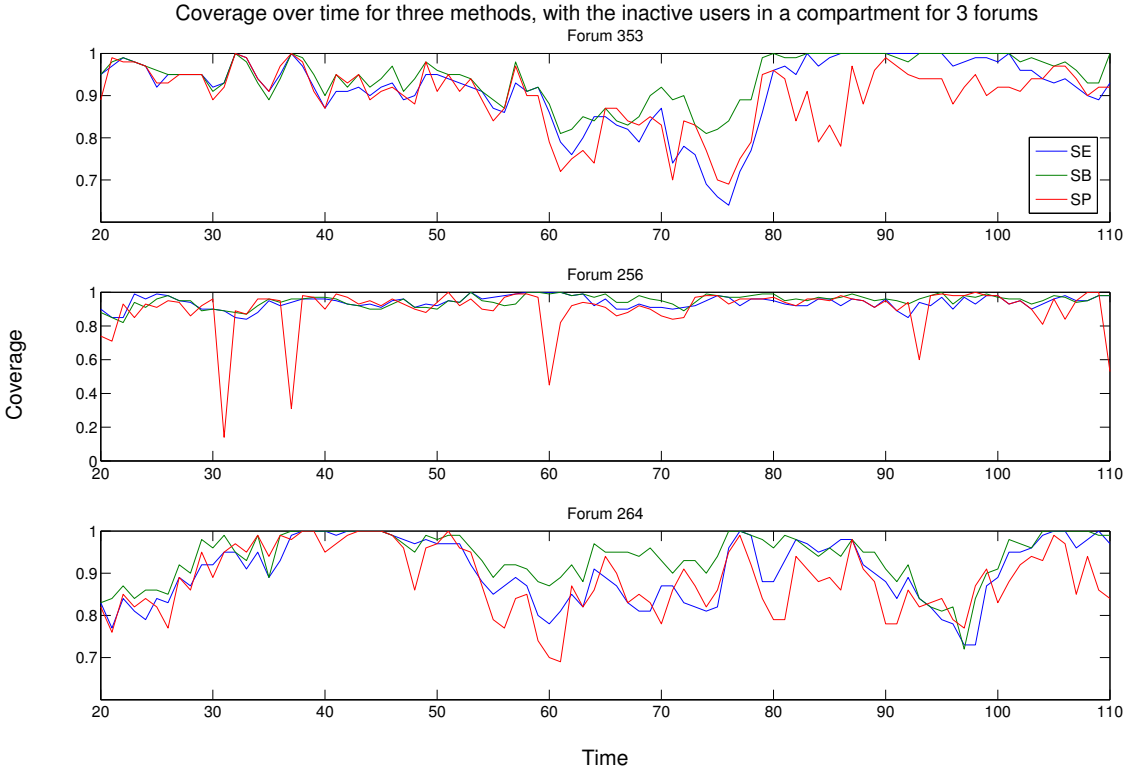


(B) Under the **I** formulation

FIGURE 5.4: The MSE over time between **SB**, **SE** and **SP** for 3 forums for both formulations



(A) Under the **W** formulation



(B) Under the **I** formulation

FIGURE 5.5: The coverage over time between **SB**, **SE** and **SP** for 3 forums for both formulations

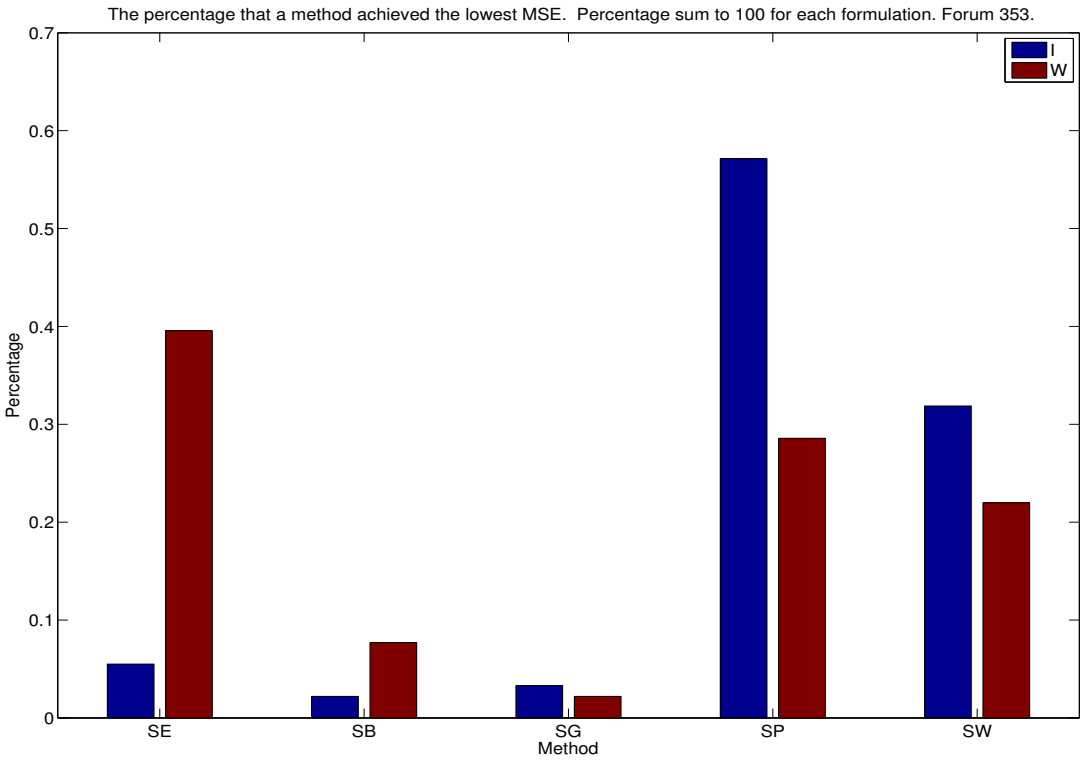


FIGURE 5.6: The fraction of time a method achieved the lowest MSE in forum 353. Each formulation sum to 100.

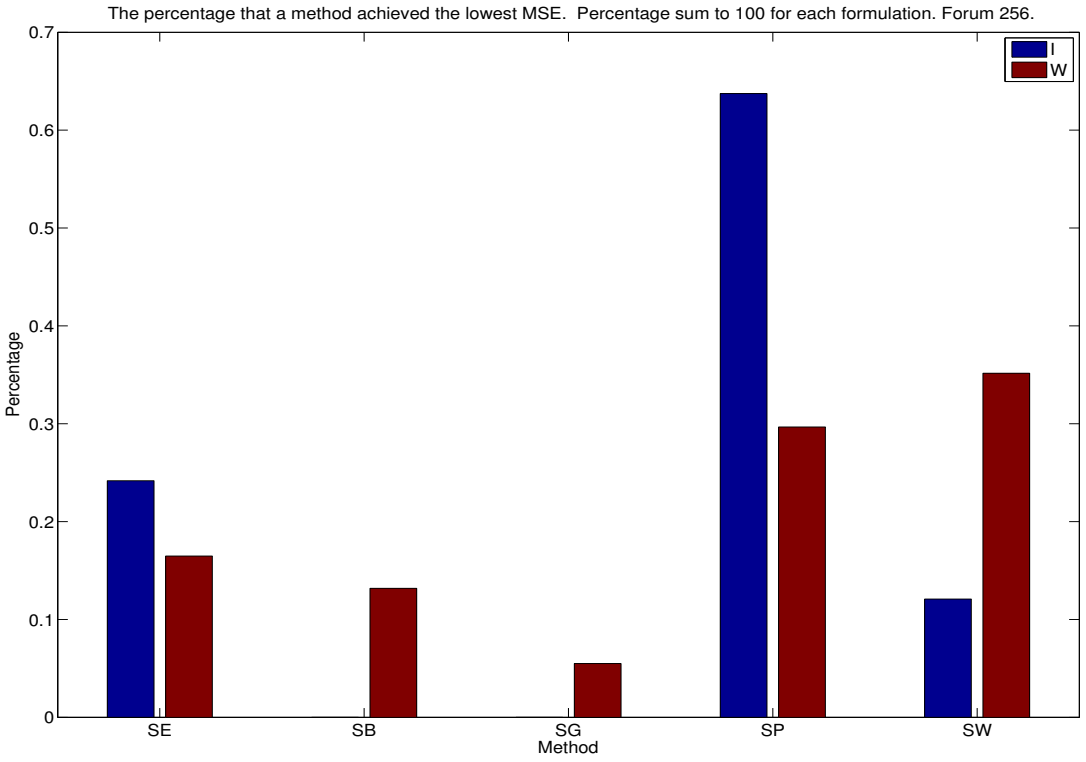


FIGURE 5.7: The fraction of time a method achieved the lowest MSE in forum 256. Each formulation sum to 100.

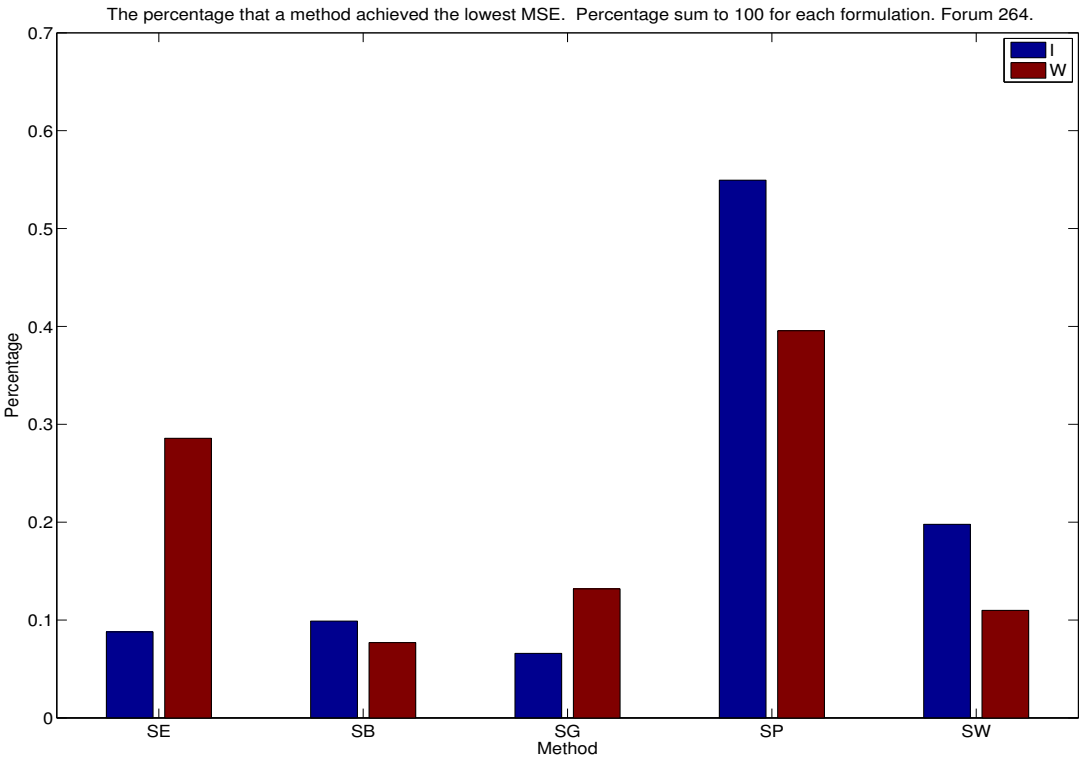


FIGURE 5.8: The fraction of time a method achieved the lowest MSE in forum 264. Each formulation sum to 100.

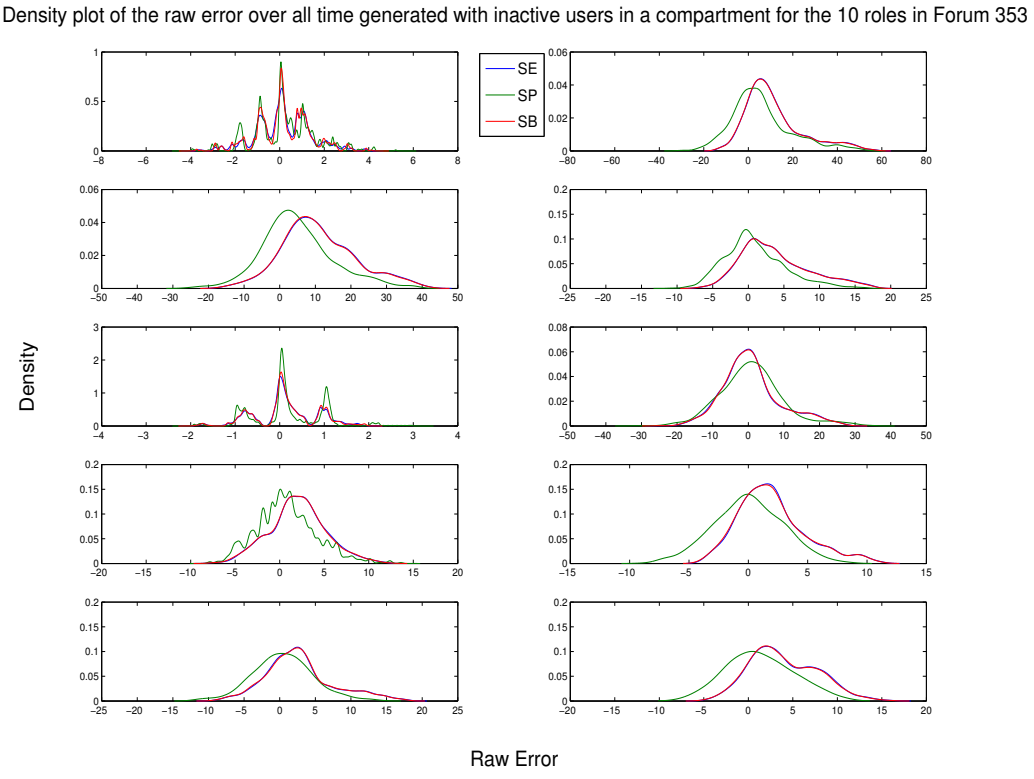


FIGURE 5.9: Kernel density plot of the raw error for each of the roles in forum 353 under the **I** formulation for 3 methods over all time

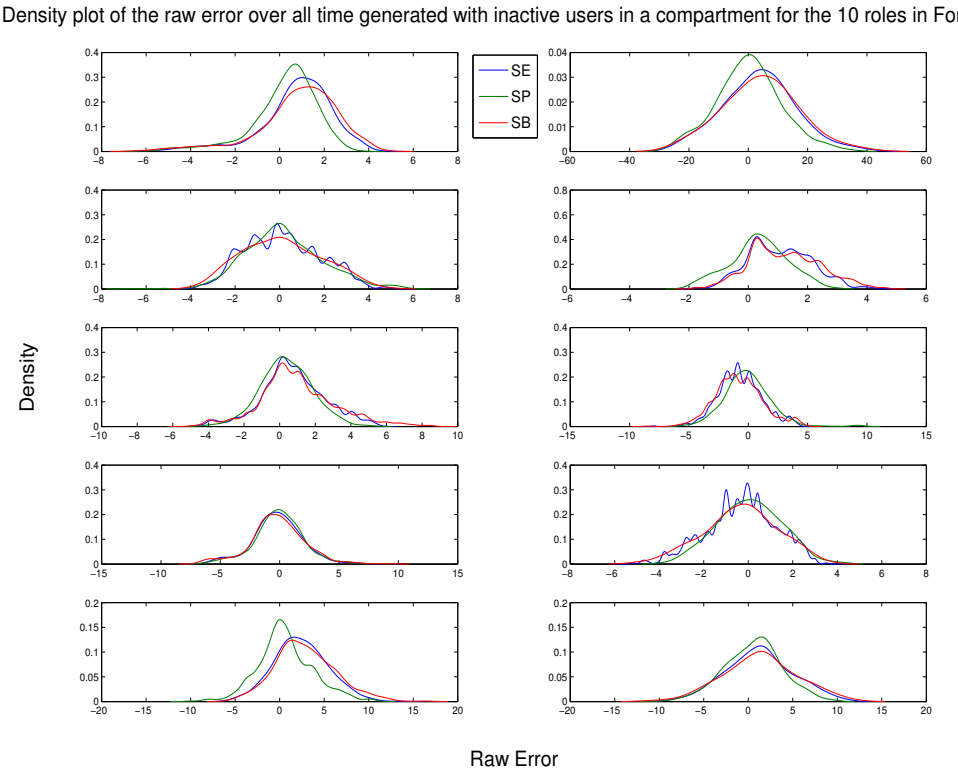


FIGURE 5.10: Kernel density plot of the raw error for each of the roles in forum 256 under the **I** formulation for 3 methods over all time

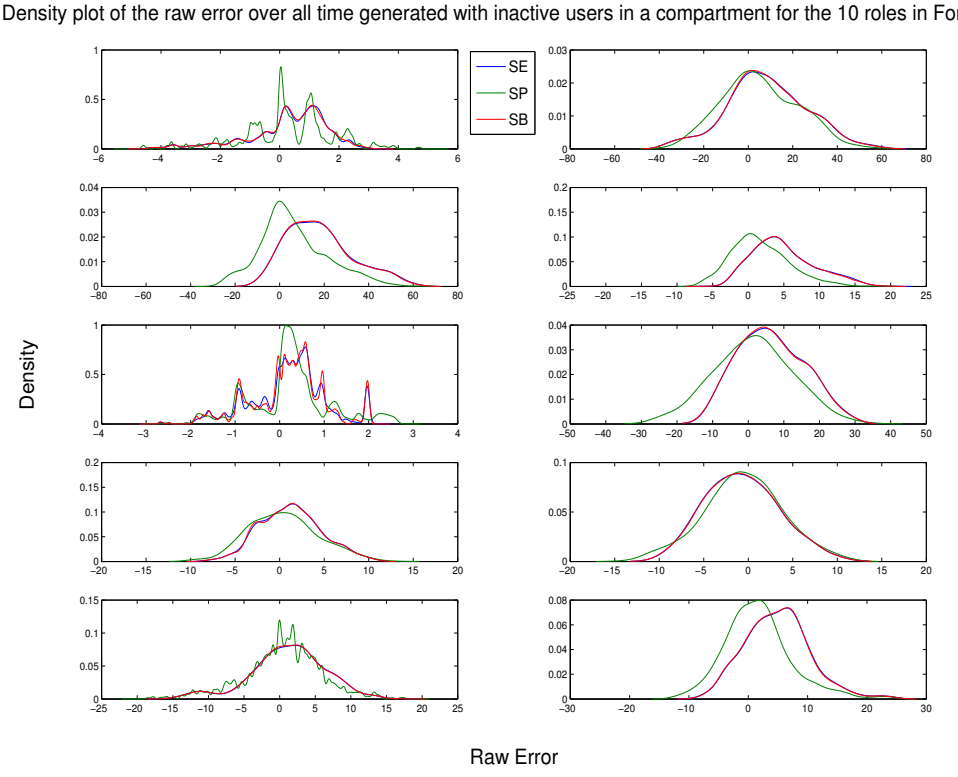


FIGURE 5.11: Kernel density plot of the raw error for each of the roles in forum 264 under the **I** formulation for 3 methods over all time

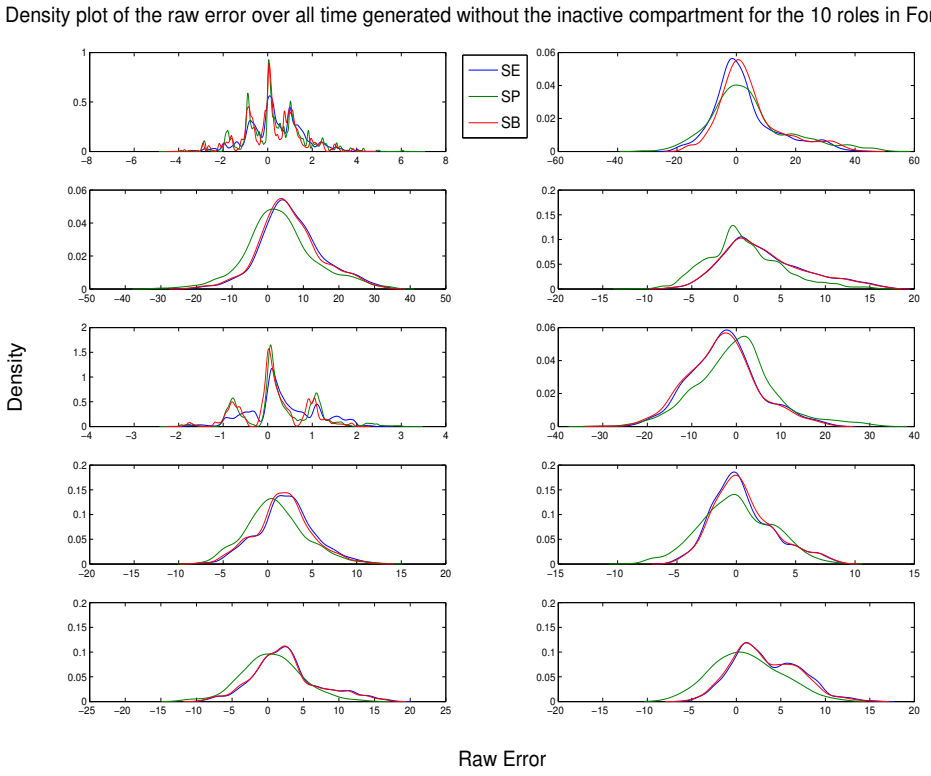


FIGURE 5.12: Kernel density plot of the raw error for each of the roles in forum 353 under the **W** formulation for 3 methods over all time

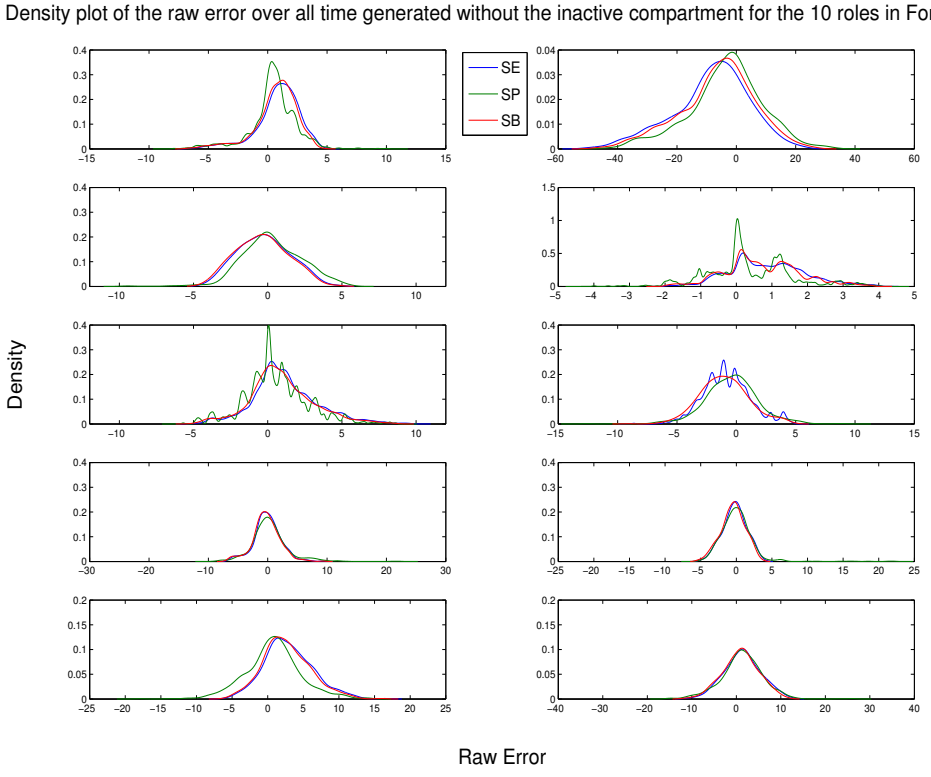


FIGURE 5.13: Kernel density plot of the raw error for each of the roles in forum 256 under the **W** formulation for 3 methods over all time

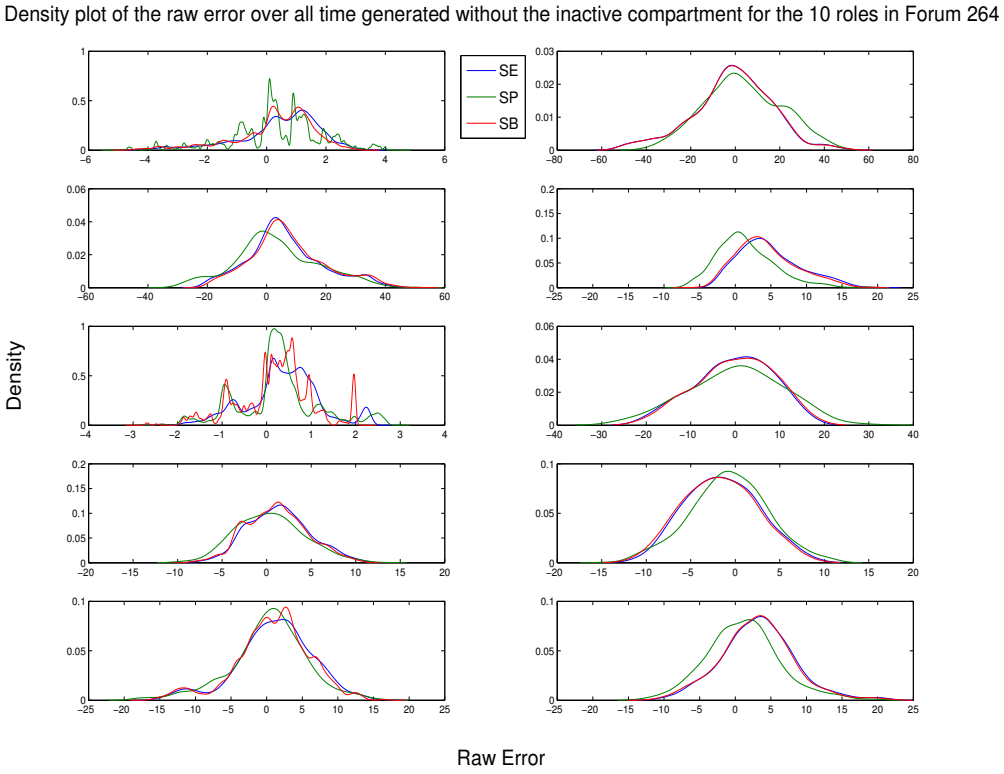
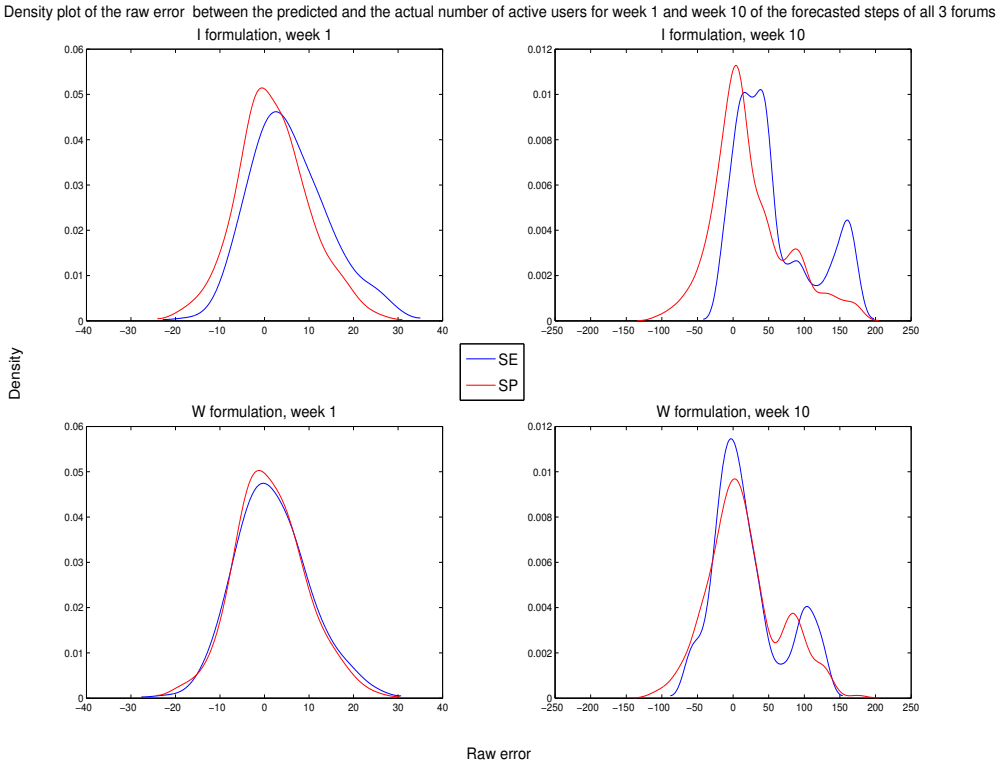
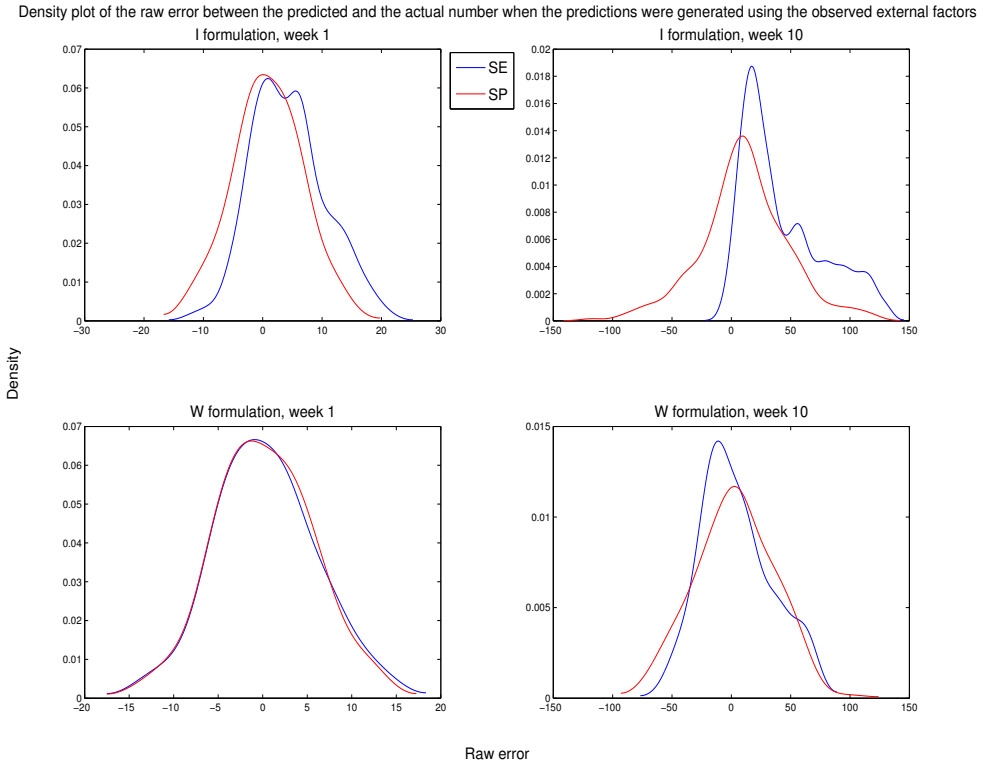


FIGURE 5.14: Kernel density plot of the raw error for each of the roles in forum 264 under the **W** formulation for 3 methods over all time



(A) Forecasts are generated using the predicted $\hat{y}_{Join}, \hat{y}_{Return}$



(B) Forecasts are generated using the observed y_{Join}, y_{Return}

FIGURE 5.15: Kernel density plot of the raw error between the predicted and the actual active users at week 1 (Left panels) and week 10 (Right panels) of the forecast over 3 forums, two methods **SE** and **SP** under both formulations are shown

Chapter 6

Concluding Remarks

6.1 Summary

The motivation for the work of this thesis was based on the increasing impact of online communities, as a thriving community provides knowledge for both the existing and future users. Our models were built based on the perspective that each user in an online community has a certain online social role. We first presented the motivation and method of identifying the different users in Chapter 2, before proceeding to tackle the problem initially defined in Chapter 1; generating role composition forecasts of an online community for some future time point. We separated our problem into two parts; the movement of users between the various roles were investigated in Chapter 3 using a compartment model, then Chapter 4 tackled the problem of predicting the number of new users joining a community and those returning from inactivity using variants of Poisson regression.

Extensive testing for the proposed methods were performed for 3 forums over the span of a two years period. We found that the simple models that make use of observed transition matrices performed best for a deterministic forecast. The overall out–of–sample MSE generated by the stochastic and deterministic models are similar. Furthermore, the stochastic models have the benefit of constructing confidence intervals for each of the roles. The 95% confidence interval around the expected forecast at each of the role contains approximately 95% of the out–of–sample observations, suggesting that the confidence interval is useful. In particular, the penalized method (Section 3.4.2) provided a mean of describing the impact of current trend using a single parameter. The errors obtained from the stochastic penalized method also appeared more homoscedastic than the other proposed

methods. Therefore, we recommend this penalized approach while predicting the total number of users joining and returning when making forecasts.

Predictions of external factors, the number of new users joining the community and users returning from inactivity, was found to be important. The error it contributed was significant as it had doubled the MSE when compared against a forecast generated using the actual observations. When the system is allowed to run for a prolonged period, the role composition undergoes fluctuations when there are incoming new users as demonstrated in Figure 3.3.

6.2 Difficulties and Future Research

The models used for forecasting compose of the two separate problems, investigated and discussed separately in Chapter 3 and Chapter 4. Therefore, we discuss the possible avenues that can be explored for each of them individually. A clear outline of the problems we stumbled upon and a detailed description of the topics that merit further investigation is presented below.

6.2.1 Compartment Model

As mentioned previously in Section 3.5.2, parameter estimation in the empirical based stochastic models are based on approximating the expected forecast via Monte Carlo simulation. Time taken to generate a set of samples, say 10^4, only required one second even with $q = 10$ steps in our forecast, which is independent of the number of observed transitions matrices $N - 1$. The exact expected value on the other hand, is dependent on the number of observations. To see this, let $\mathbf{A}(i) = w_i \mathbf{P}^\top(i)$ and for simplicity assume that $y_{Join}(t) = 0$ for $t = 1, 2, \ldots, N-1$ in the estimation. Then, the expected value of a one step ahead forecast starting at t_0 is

$$\mathbb{E}(\hat{\mathbf{m}}(t_0 + 1)) = \sum_{i=1}^{N-1} \mathbf{A}(i)\mathbf{m}(t_0),$$

and for a q step forecast is

$$\mathbb{E}(\hat{\mathbf{m}}(t_0 + q)) = \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \ldots \sum_{k=1}^{N-1} \sum_{z=1}^{N-1} \mathbf{A}(z)\mathbf{A}(k) \ldots \mathbf{A}(j)\mathbf{A}(i)\mathbf{m}(t_0), \tag{6.1}$$

where the set of indices $\mathcal{F} = \{i, j, \ldots, k, z\}$ contains q elements, that is, a q summation over the q multiplicative weighted transition matrix $\mathbf{A}$. It is of interest to investigate whether there exists an efficient computation for the $(N-1)^q$ summation in (6.1).

Evidently, computing the expected value explicitly includes $\mathbb{E}(\hat{\mathbf{m}}(t_0+q-1))$, the expected forecast at the t_0+q-1, which also exists in the current optimization problem

$$\min \sum_{i=1}^{q} \|\mathbf{m}(t_0+i) - \mathbb{E}(\hat{\mathbf{m}}(t_0+i))\|^2 . \tag{6.2}$$

The optimization problem (6.2) can be thought of in terms of minimizing the squared norm of a constant vector (observed $\mathbf{m}$) minus the random vector (our forecast) that is dependent on the stochastic process. As we are taking the expectation with respect to the forecast, (6.2) is the RHS of

$$\mathbb{E}(\|c-\xi\|^2) \geq \|c - \mathbb{E}(\xi)\|^2, \tag{6.3}$$

which underestimates the error. A challenge will be to tackle the optimization problem that takes the expectation on the squared norm, LHS of (6.3).

Efficient estimation of the parameters in the empirical based stochastic models namely, **SP** and **SW**, should be explored. The current method used for both problems do not take into account the information on the derivative. This is because computing the gradient is nearly as hard as computing the objective function. Making use of the previously simplified objective function (6.1), and differentiate with respect to w_g yields a summation that is one order lower

$$\nabla_{w_g} \mathbb{E}\left(\hat{\mathbf{m}}(t_0+q)\right) = q\mathbf{P}^\top(g) \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \ldots \sum_{k=1}^{N-1} \mathbf{A}(k) \ldots \mathbf{A}(j)\mathbf{A}(i)\mathbf{m}(t_0).$$

This, of course, is proportional to $\mathbb{E}(\hat{\mathbf{m}}(t_0+q-1))$, which brings back issues regarding the efficient computation of the summation terms. Other derivative free methods such as those that uses function approximations to the objective function with respect to the parameters, i.e. radial basis functions Jones et al. (1998); Gutmann (2001); Wild and Shoemaker (2013) may be a good alternative. An efficient implementation in both the interpolation and picking initial sets of points is required because the feasible region is defined in a $(N-2)$–simplex. The common choice of evaluating the corners and the mid point of all edges for a d

dimension problem equates to $d+(d(d-1)/2)$ function evaluation, which is highly inappropriate for large d such as 100 in our problem (Chapter 3).

There also exists a few other interesting adjustments to the model formulation;

- Changing the number of in–sample tuning steps in the non–linear estimation. This will allow more or less adaptation on the current trend by having more or less tuning steps used.
- When estimating the transition matrix directly, additional constraints that restrict the amount of change when compared to some reference matrix $\mathbf{P}^{ref}$, such as the expected transition matrix or latest observed transition matrix, i.e. $\left|\mathbf{P}_{i,j}(t_0) - \mathbf{P}_{i,j}^{ref}\right| \leq \delta$ for some pre–defined value δ, such as a factor of the sample variance.
- Model the autocorrelation on the vector $\mathbf{P}_0$, the set of migration rates going out of the compartments containing inactive users. One particular approach would be to use a latent autoregressive construction as described in Section 4.5.3, but with a Binomial density (3.42) instead of a Poisson density that models the total number of users returning directly.
- Further correlation structure can be imposed onto the migration rates if we do not assume the forums to be independent. This implies that forecasts are produced for more than one forum simultaneously and also the estimation of the correlations.

Alternatively, other formulation for compartmental models can also be used. Notably, the standard deterministic differential equation that is seen commonly with compartmental models (Godfrey, 1983; Brauer and Castillo-Chavez, 2001). Then assume that the observations $\mathbf{m}(t) = \mathbf{y}(t) + \mathbf{e}(t)$ are realizations from some underlying process with error $\mathbf{e}$. Our discrete time Markov chain interpretation is just one of the three common types of stochastic formulation based on differential equations. The other two are continuous time Markov chain and stochastic differential equation (Brauer et al., 2008, chap. 3).

6.2.2 Prediction of External Factors

The predictions on the external factors only used a limited number of covariates, they may be improved by exploring other variables that may be relevant. Further

extension to the PLN–AR(1) model such as a latent vector autoregressive (VAR) on y_{Join}, y_{Return} may further improve the predictive power, i.e. extending (4.48) such that the dispersion with VAR(1) is

$$\mathbf{v}_i \sim \mathcal{N}_k(\mathbf{\Phi}_1 \mathbf{v}_{i-1}, \mathbf{\Sigma}) \text{ for } i \neq 1, \tag{6.4}$$

where $\mathbf{\Phi}_1$ is a $[k \times k]$ matrix that contains the first order autoregressive elements. Evidently, this further increases the model complexity that is already lacking a close form solution. Hence, requiring samplers that are more efficient than those currently used in Chapter 4. A brief survey of the current MCMC literature provides no shortage of ideas, where the higher efficiency sampling can be done using techniques such as: include manifold information of the posterior (Girolami and Calderhead, 2011), adaptive Hamiltonian schemes (Hoffman and Gelman, 2014), alternating between different parameterization (Hobert et al., 2011) and subsampling (Bardenet et al., 2014), etc.

Models outside of the Poisson family were not investigated due to the low value counts observed in the data. This is not a limitation on other datasets where the number of new users joining say, are consistently in the hundreds. Additionally, there is the possibility of predicting the vector of users joining each of the roles directly. Hence, eliminating the estimation of the proportional vector γ (3.8).

Forums with a sudden burst of new users were also ignored, because it was a rare event for the data under study from SAP. If the cause of this surge can be linked to an event, say an advertisement campaign or an event like a conference, then the predictability may be high, especially if the events happen periodically.

Finally, seasonality in the data were not properly explored. A one week discretization was used throughout the thesis, where no obvious seasonality could be observed. Using different time discretization such as month or day of the week may reveal seasonality patterns that our models are able to exploit.

Appendix A

A.1 Data summary for some forums

Forum id	# Total User at $N = 1$	# Total User at $N = 120$
353	588	2339
256	287	1424
264	1018	3445
Forum id	# Active User at $N = 1$	# Active User at $N = 120$
353	542	636
256	264	442
264	974	1009

TABLE A.1: The number of total and active users at two different time point for the 3 forums of interest

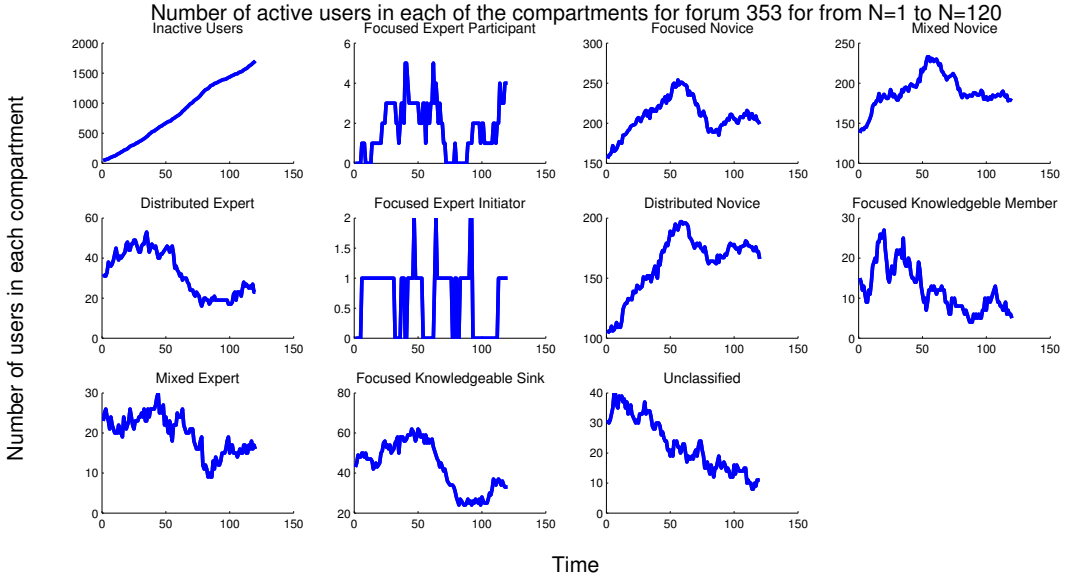


FIGURE A.1: The number of users in each of the compartments through time for forum 353

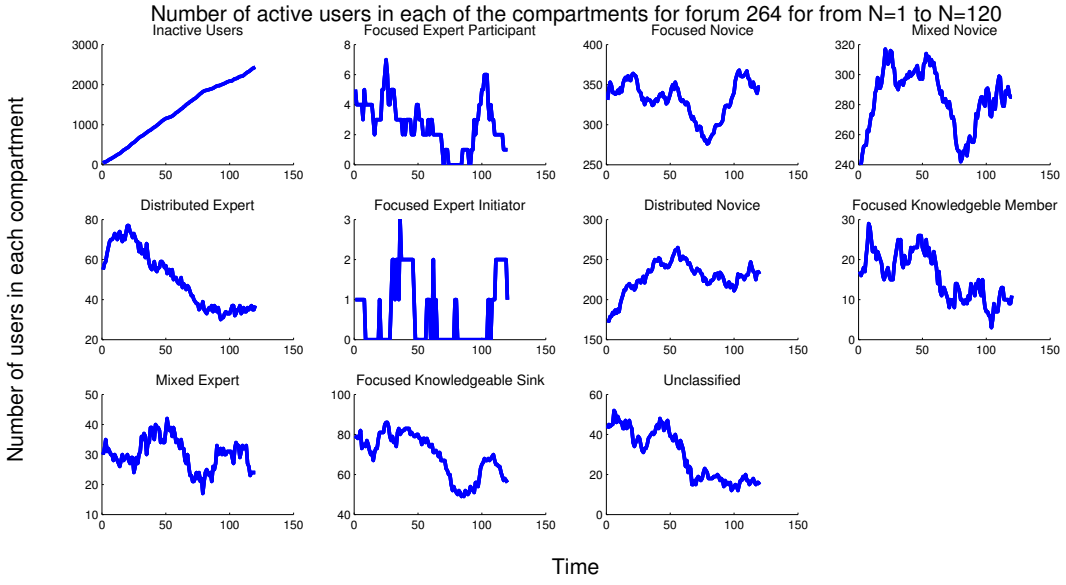


FIGURE A.2: The number of users in each of the compartments through time for forum 264

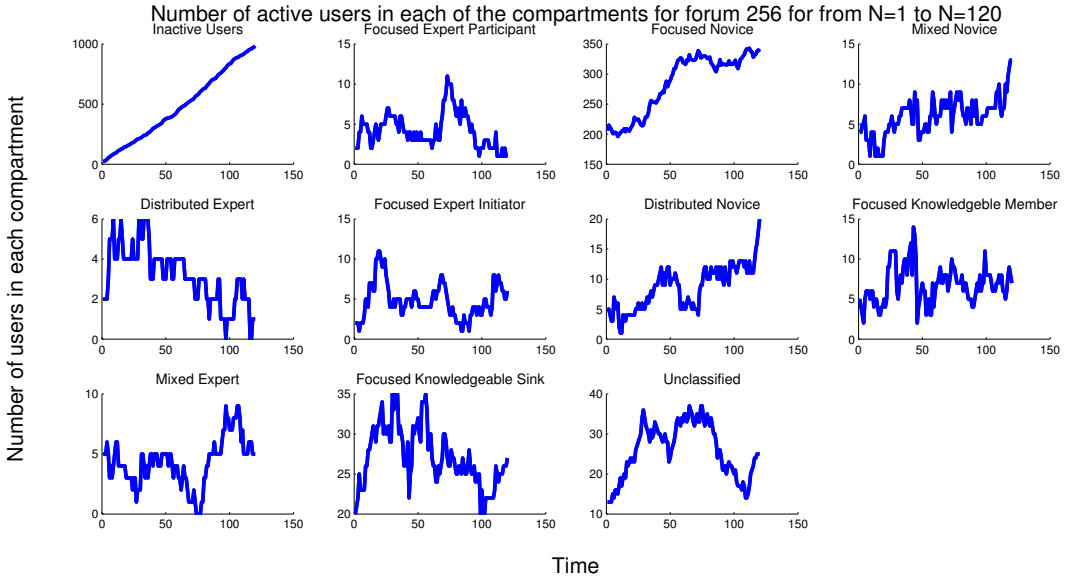


FIGURE A.3: The number of users in each of the compartments through time for forum 256

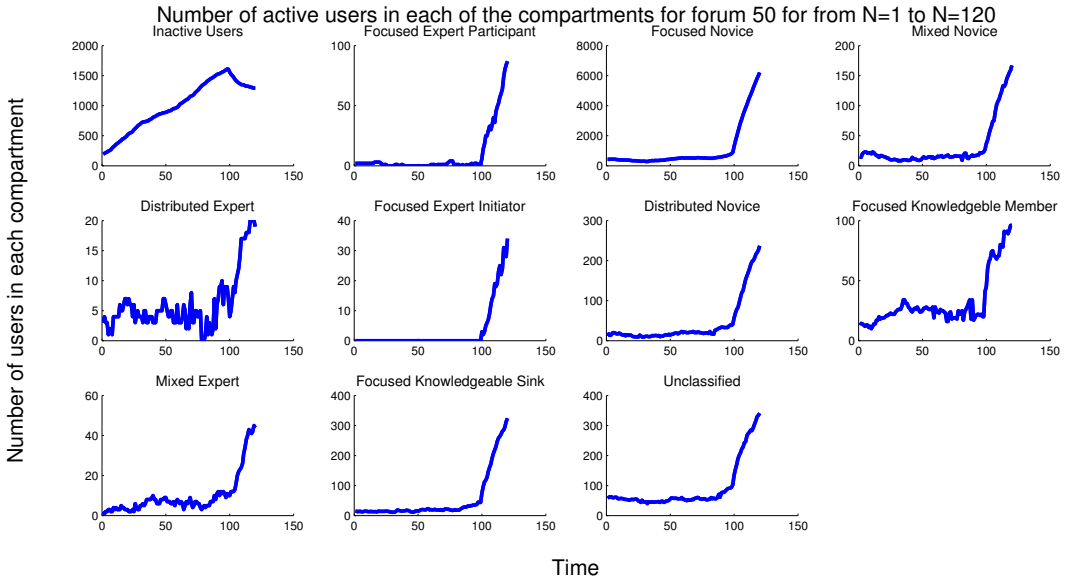


FIGURE A.4: The number of users in each of the compartments through time for forum 50

FIGURE A.5: Network graph for forum 264 with data collected under a 13 week time window ending at time N=120. The bigger the node, the higher the betweenness centrality the node has and an edge with higher weight is darker

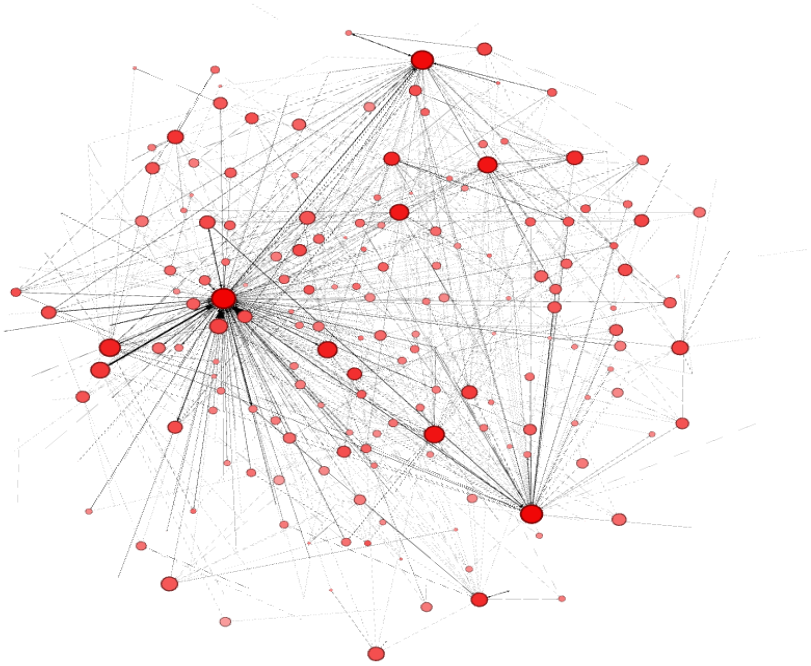


FIGURE A.6: Network graph for forum 353 with data collected under a 13 week time window ending at time N=120. The bigger the node, the higher the betweenness centrality the node has and an edge with higher weight is darker

Forum id	Start date (dd/mm/yy)	Total user	Threads per month	Message per thread (s.d.)	Length per Thread (s.d.)
246	11/01/06	10705	682	4.48	37.59
245	01/19/06	11022	575	4.38	41.41
144	06/30/05	8507	545	4.89	24.07
142	09/20/04	10794	467	4.47	38.11
264	05/25/07	4034	364	4.77	13.79
323	02/20/07	6399	338	3.79	18.09
143	07/23/04	6408	282	4.2	45.62
141	02/06/06	7663	271	3.68	64.47
156	09/19/05	7751	233	3.86	46.5
56	02/24/04	4615	210	4.32	33.23
140	07/20/04	10183	179	2.99	45.51
284	04/28/06	3116	168	4.33	23.04
126	10/13/04	1329	166	3.01	7.53
324	11/28/06	3890	165	3.86	21.52
328	01/01/06	5360	159	3.98	20.03
50	06/22/04	9704	156	4.13	96.31
159	02/28/06	2764	148	5.01	23.05
239	03/21/05	3673	145	3.98	34.39
276	10/16/07	2897	138	3.44	28.91
292	03/05/08	2095	117	4.21	61.74
407	03/17/09	1722	109	2.95	21.02
327	10/06/06	3607	107	4.34	23.42
413	12/07/07	1435	103	4.45	26.99
412	06/19/08	1408	93	4.18	23.25
353	05/20/04	2902	93	5.02	20.72
Averages		1975	102	4.24 (2.89)	36.19 (136.59)

TABLE A.2: Summary statistics of those forums that contain 80 % of the total number of threads generated. Ordered by the number of threads generated per month since the starting date

Forum id	% No reply	Average TTS (day)	% Solved in same day	% Solved	% Answered	% Solved or answered
246	7	0.3	94	22	35	57
245	7	0.38	94	25	33	57
144	3	0.29	91	29	33	62
142	11	0.82	89	17	43	60
264	1	0.55	9	44	13	57
323	9	0.44	89	22	38	60
143	11	0.99	83	18	51	68
141	14	0.76	91	15	45	59
156	16	1.27	83	17	47	64
56	3	1.74	87	47	18	64
140	26	1.16	91	11	63	74
284	7	0.78	88	2	39	59
126	29	0.37	85	9	7	79
324	7	1.09	9	22	4	61
328	9	0.27	94	19	46	64
50	16	0.32	92	14	55	69
159	4	0.46	86	26	35	62
239	6	0.72	9	38	31	69
276	15	0.6	86	18	43	61
292	12	1.48	79	18	4	58
407	27	0.97	84	13	55	68
327	4	0.48	92	21	11	32
413	8	0.93	86	21	44	65
412	1	1.44	82	24	44	67
353	3	1	86	41	18	59
Averages	9	86	88	23	39	62

TABLE A.3: Summary statistics for the same list of forums as in Table 2 on a thread level, where % No reply indicates the percentage of thread within the forum that has gotten no reply up until the last observed time as of data extraction. The column "% Solved in same day" is only a percentage for threads that has been solved up until the last observed time.

Appendix B

B.1 Estimated probability matrix under different constraints

0.8913	0.0150	0	0	0.0039	0	0.0216	0.0346	0.0147	0.0033	0.0155
0	0.9252	0.0365	0.0000	0.0000	0.0013	0.0002	0.0001	0.0006	0.0357	0.0002
0	0.0352	0.8909	0.0002	0	0.0454	0.0005	0.0004	0.0013	0.0257	0.0003
0	0.0005	0.0025	0.9137	0	0.0043	0.0066	0.0292	0.0270	0.0107	0.0055
0.0294	0	0	0	0.9265	0	0	0.0392	0.0049	0	0
0	0.0020	0.0504	0.0009	0	0.9240	0.0010	0.0001	0.0033	0.0181	0.0004
0.0014	0.0090	0.0092	0.0243	0	0.0141	0.8593	0.0136	0.0313	0.0056	0.0322
0.0031	0.0028	0.0069	0.0456	0.0034	0.0004	0.0067	0.8575	0.0229	0.0207	0.0301
0.0009	0.0015	0.0061	0.0204	0	0.0121	0.0139	0.0085	0.9328	0.0005	0.0033
0	0.0022	0.0017	0.0001	0	0.0014	0.0001	0.0005	0.0000	0.9938	0.0003
0.0016	0.0054	0.0041	0.0045	0.0003	0.0051	0.0103	0.0277	0.0014	0.0362	0.9034

TABLE B.1: The expected value of **P** for forum 353 using 100 observations. A 0 indicates that the edge does not exist and 0.0000 is some small value

0.5011	0.4835	0	0	0	0	0	0.0154	0	0	0
0.0018	0.7571	0.1771	0	0.0015	0.0538	0	0.0087	0	0	0
0	0.1498	0.7887	0	0	0	0	0	0	0.0616	0
0	0.0435	0	0.8586	0	0	0	0	0.0979	0	0
0	0	0.2031	0.2794	0.5141	0	0.0034	0	0	0	0
0	0.0738	0	0	0	0.9051	0	0	0	0.0211	0
0	0	0	0.0902	0	0	0.8285	0.0813	0	0	0
0.0013	0	0	0.0023	0	0	0	0.6705	0.2289	0	0.0971
0.0103	0.0739	0	0.0438	0	0	0.0098	0.0648	0.7973	0	0
0	0	0	0	0	0.0015	0	0	0	0.9985	0
0	0	0	0.0364	0	0	0.0695	0	0	0	0.8941

TABLE B.2: The estimated $\mathbf{P}$ matrix using the natural bounds of $[0, 1]$ for forum 353 using 100 observations of $\mathbf{P}$

0.5494	0.3744	0	0	0	0	0	0.0761	0	0	0
0	0.9041	0.0791	0	0.0004	0.0108	0	0.0056	0	0	0
0	0.0546	0.8650	0	0	0.0156	0	0.0004	0	0.0643	0
0	0.0160	0	0.8783	0	0	0	0	0.1048	0	0.0008
0.1069	0	0	0	0.9765	0	0	0.0165	0	0	0
0	0.0177	0.0251	0	0	0.9421	0	0	0	0.0151	0
0	0	0	0.0790	0	0	0.8322	0.0846	0.0043	0	0
0	0	0.0640	0	0	0	0	0.6704	0.1716	0	0.0940
0.0157	0.0217	0.0086	0.0432	0	0	0.0135	0.0741	0.8231	0	0
0	0	0	0	0	0.0009	0	0	0	0.9991	0
0	0	0	0.0357	0	0	0.0619	0	0	0	0.9023

TABLE B.3: The estimated $\mathbf{P}$ matrix using bounds defined by the past observed rate $\min_{t:1,\ldots,T}\{\mathbf{P}_{i,j}(t)\}$ and $\max_{t:1,\ldots,T}\{\mathbf{P}_{i,j}(t)\}$ for forum 353 using 100 observations of $\mathbf{P}$

B.2 Comparison of Formulations

	I			**W**		
	S1	**S2**	**S3**	**S1**	**S2**	**S3**
Average MSE	76	105	67	40	42	35
Number of lowest MSE	38	3	50	18	33	40

TABLE B.4: The average error and the number of times a scenario has achieved the lowest error over a 90 week period for forum 56

	I			**W**		
	S1	**S2**	**S3**	**S1**	**S2**	**S3**
Average MSE	55	61	54	19	18	16
Number of lowest MSE	42	14	35	34	24	33

TABLE B.5: The average error and the number of times a scenario has achieved the lowest error over a 90 week period for forum 256

	I			W		
	S1	**S2**	**S3**	**S1**	**S2**	**S3**
Average MSE	127	163	124	59	57	47
Number of lowest MSE	42	5	44	20	24	47

TABLE B.6: The average error and the number of times a scenario has achieved the lowest error over a 90 week period for forum 264

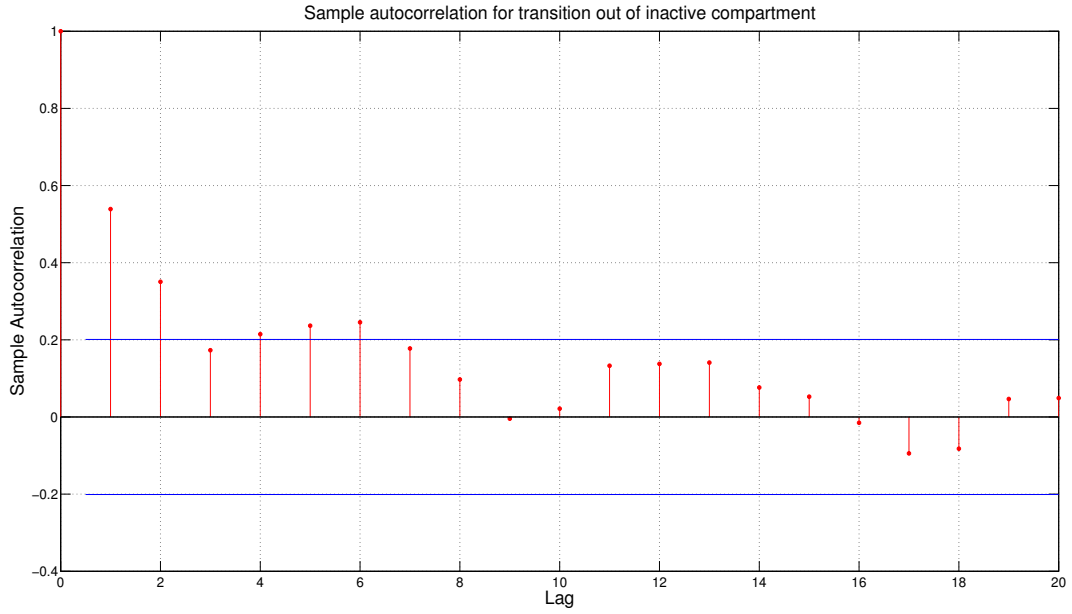


FIGURE B.1: Autocorrelation plot of the first 20 lags for the transition $\mathbf{P}_{0,0}$, the number of user remains in the inactive compartment.

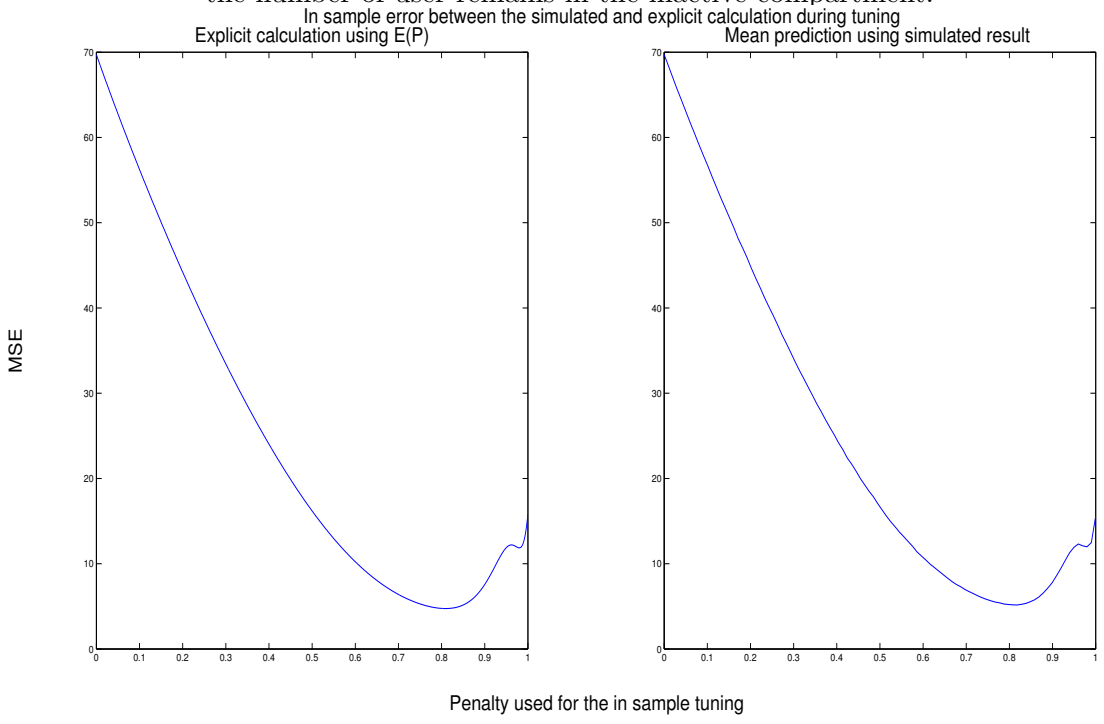


FIGURE B.2: MSE over a set of α values using the inactive compartment for a 10 step in–sample tuning using 100 observations for forum 353

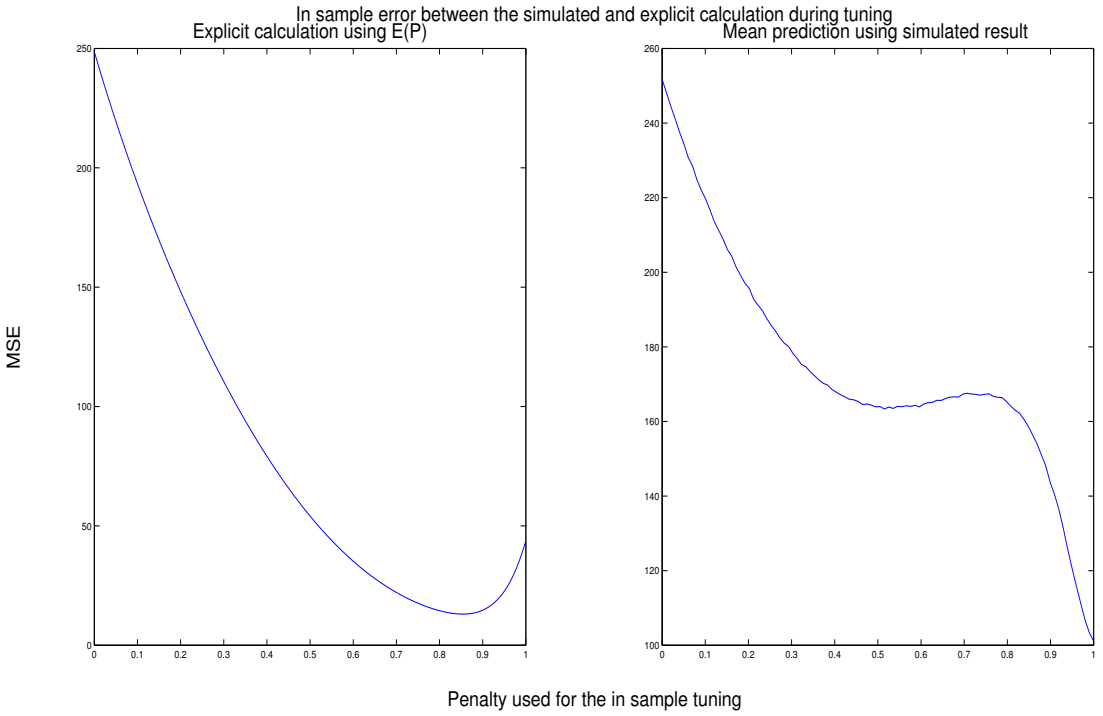


FIGURE B.3: MSE over a set of α values without the inactive compartment for a 10 step in–sample tuning using 80 observations for forum 264

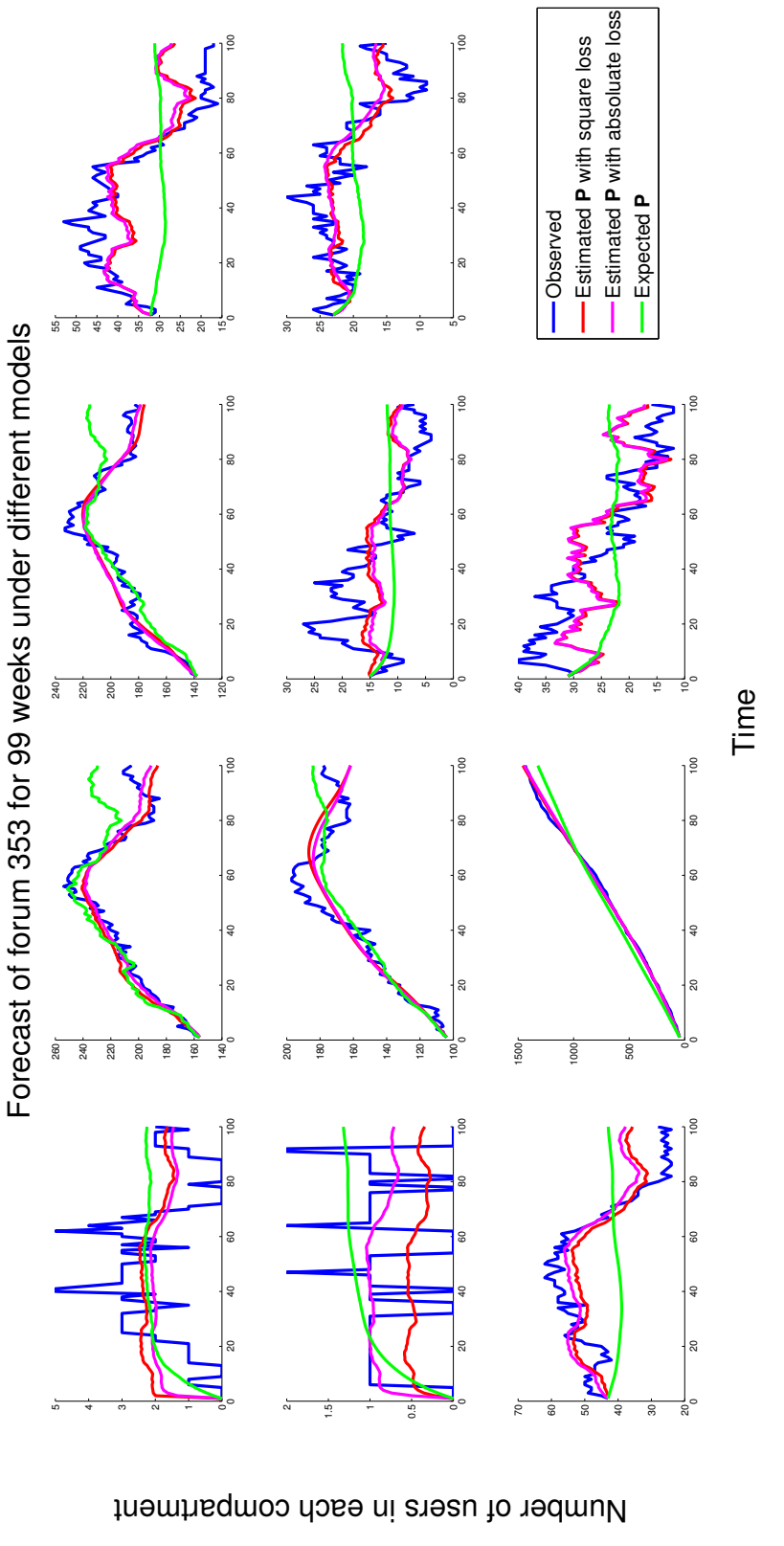


FIGURE B.4: Forecast using $\mathbf{P}$ estimated under absolute loss and square loss using the non–linear iterative formulation. The 99 steps forecasts were generated using the first observation as the initial value and the observed number of users at each time period

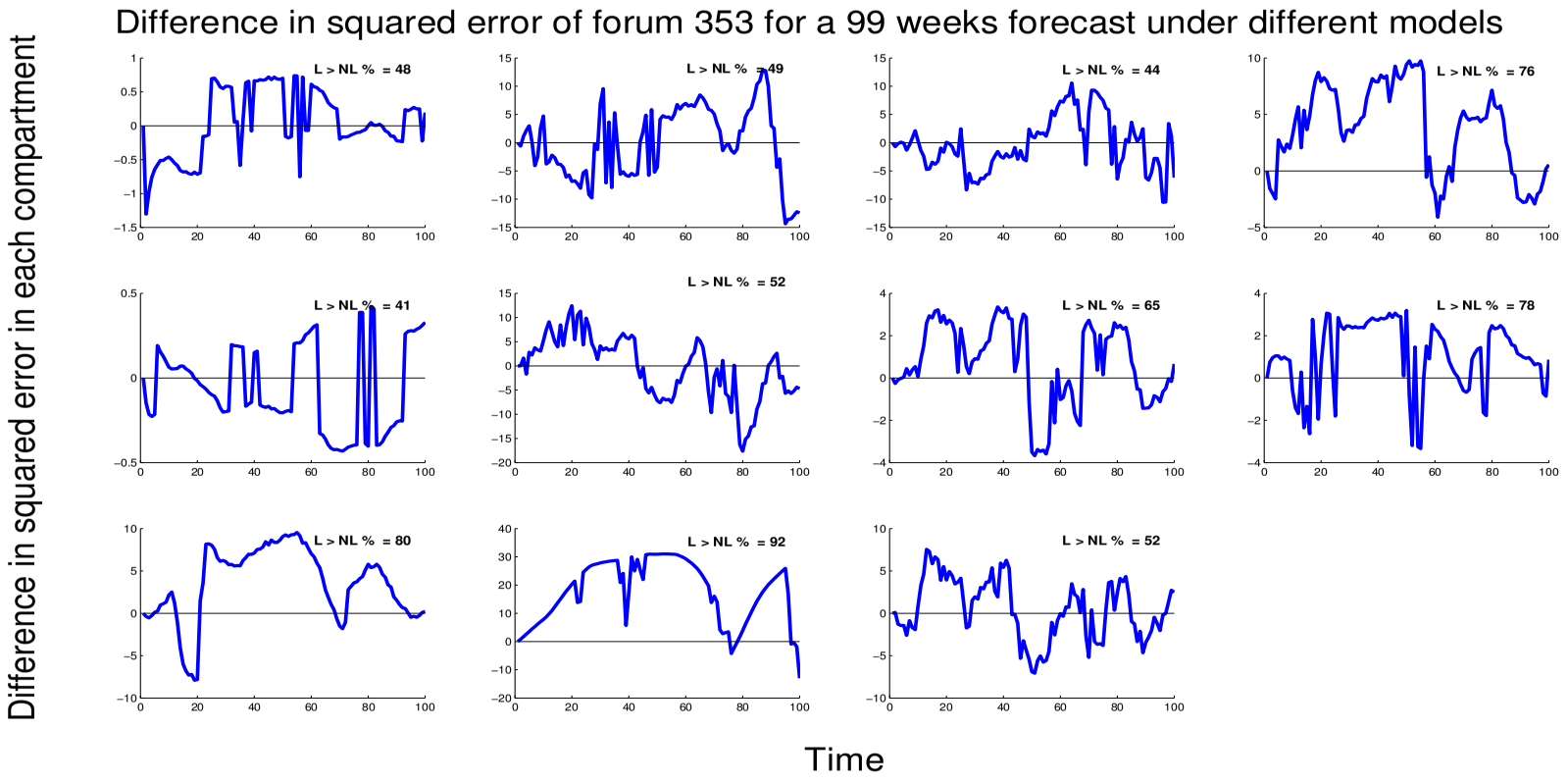


FIGURE B.5: Difference in the squared error between the forecast using $\mathbf{P}_\dagger$ estimated using the linear single step update and the non–linear iterative update under $\mathbf{I}$ formulation. A positive value indicates that the linear version has more squared error (against the observation) as compared to the non–linear version. Labels at the top right hand corner of each subplot shows the number of times the linear version is positive.

Appendix C

C.1 Comparison Between Different Proposal For Poisson Regression

The following two plots were generated using 1×10^5 samples, after a burn–in period of 1×10^3 iterations. The prior of the variance is $\sigma^{-2} \sim \mathcal{G}a(a_0, b_0)$ or $\lambda^2 \sim \mathcal{G}a(a_0, b_0)$ with $a_0 = b_0 = 0.001$ for both. The figures used the first 100 observations from forum 353 using lagged covariates.

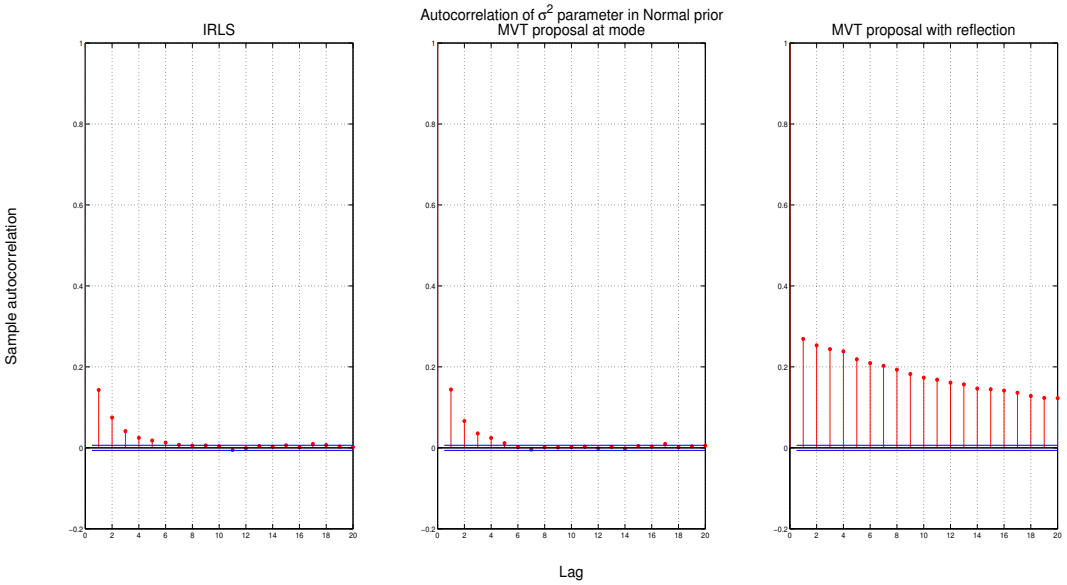


FIGURE C.1: Autocorrelation plot over 50 lags for σ^2 under Normal prior

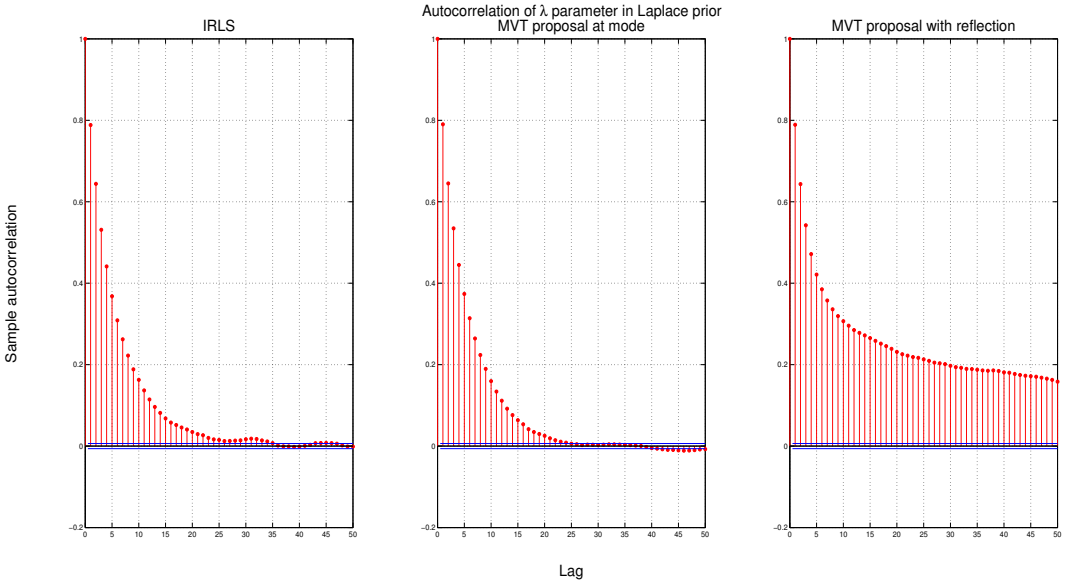


FIGURE C.2: Autocorrelation plot over 50 lags for λ under Laplace prior

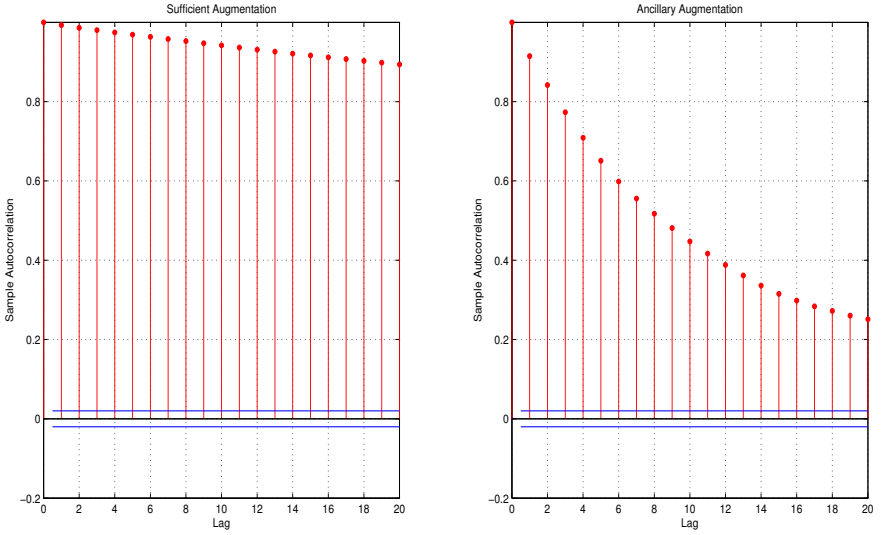


FIGURE C.3: Difference in autocorrelation between the two different sampling scheme for the Multivariate Poisson Log–normal

C.2 Conditional Mean and Variance of Overdispersed Poisson

The marginal of Y in the general case of heterogeneity (4.32) can be derived using the law of total expectation

$$\mathbb{E}(X) = \mathbb{E}_Y(\mathbb{E}_{X|Y}(X \mid Y)) \tag{C.1}$$

and the law of total variance

$$\text{Var}(Y) = \mathbb{E}_X(\text{Var}_{Y|X}(Y \mid X)) + \text{Var}_X(\mathbb{E}_{Y|X}(Y \mid X)). \tag{C.2}$$

The Poisson regression model with dispersion has mean parameter $\tilde{\lambda} = \lambda u$. Let $\mathbb{E}(U) = \mu_U, \text{Var}(U) = \sigma_U^2$ and λ is fixed. When the equal mean–variance relationship

$$\begin{aligned}\mathbb{E}(Y \mid \lambda, U) &= \tilde{\lambda} \\ \text{Var}(Y \mid \lambda, U) &= \tilde{\lambda}\end{aligned}$$

is satisfied, the expectation of Y is

$$\begin{aligned}\mathbb{E}(Y \mid \lambda) &= \mathbb{E}_U\left[\mathbb{E}(Y \mid \lambda, U)\right] = \mathbb{E}_U(\tilde{\lambda}) \\ &= \lambda\mathbb{E}(U) = \lambda\mu_U\end{aligned} \tag{C.3}$$

and the variance

$$\begin{aligned}\text{Var}(Y \mid \lambda) &= \mathbb{E}_U\left[\text{Var}(Y \mid \lambda, U)\right] + \text{Var}_U\left[\mathbb{E}(Y \mid \lambda, U)\right] \\ &= \mathbb{E}_U(\tilde{\lambda}) + \text{Var}_U(\tilde{\lambda}) \\ &= \mathbb{E}_U(\lambda u) + \text{Var}_U(\lambda v) \\ &= \lambda\mu_U + \lambda^2\sigma_U^2.\end{aligned} \tag{C.4}$$

Therefore, the mixture takes on the form of a NB2 model when $\mu_U = 1$ such that the equal mean–variance relationship is satisfied.

C.3 Poisson–Lognormal Sampling Information

C.3.1 Gradient and Hessian

For the Poisson regression with Gaussian prior $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{b}_0, \mathbf{B}_0)$, let $\boldsymbol{\lambda} = e^{\mathbf{X}\boldsymbol{\beta}}$, then posterior is

$$\mathcal{L}(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \mathbf{b}_0, \mathbf{B}_0) \propto \mathbf{y}^\top\mathbf{X}\boldsymbol{\beta} - \sum \boldsymbol{\lambda} - 0.5\log(|\mathbf{B}_0|) - 0.5(\boldsymbol{\beta} - \mathbf{b}_0)^\top\mathbf{B}_0^{-1}(\boldsymbol{\beta} - \mathbf{b}_0)$$

with gradient and Hessian as

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} \propto \mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \boldsymbol{\lambda} - \mathbf{B}_0^{-1}(\boldsymbol{\beta} - \mathbf{b}_0) \tag{C.5}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \propto -\mathbf{X}^\top \operatorname{diag}(\boldsymbol{\lambda}) \mathbf{X} - \mathbf{B}_0^{-1}. \tag{C.6}$$

Similarly, in the Poisson–Lognormal case where the mean is $\boldsymbol{\lambda} = e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{v}}$ with dispersion $\mathbf{v} \sim \mathcal{N}_D(0, \boldsymbol{\Sigma})$ of D dimension. Here, we derive the case for the general case, where $\boldsymbol{\Sigma} = \sigma^2$ in the univariate case. Both the gradient and the Hessian for $\boldsymbol{\beta}$ are the same as above as $\nabla_{\boldsymbol{\beta}} \boldsymbol{\lambda} = \boldsymbol{\lambda}$. The conditional posterior of $\mathbf{v}_i$ is

$$\mathcal{L}(\mathbf{v}_i \mid \mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto \mathbf{Y}_i^\top \mathbf{v}_i - \sum \boldsymbol{\lambda}_i - 0.5 \log(|\boldsymbol{\Sigma}|) - 0.5(\mathbf{v}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{v}_i).$$

Write $\mathbf{X}\boldsymbol{\beta} + \mathbf{v}$ as $\mathbf{X}\boldsymbol{\beta} + \mathbf{I}\mathbf{v}$, then it is obvious that the derivatives for the dispersion $\mathbf{v}$ are

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}} \propto \mathbf{Y}_i - \boldsymbol{\lambda}_i - \boldsymbol{\Sigma}^{-1} \mathbf{v}_i \tag{C.7}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{v} \partial \mathbf{v}^\top} \propto -\operatorname{diag}(\boldsymbol{\lambda}_i) - \boldsymbol{\Sigma}^{-1} \tag{C.8}$$

C.3.2 Sampling Variance Parameter Under Uniform Prior

Posterior of σ^2 given prior $p(\sigma^2)$ is

$$f(\sigma^2 \mid \mathbf{v}, \mu) \propto f_{\mathcal{N}}(\mathbf{v} \mid \mu, \sigma^2) p(\sigma) \tag{C.9}$$

where the likelihood is

$$f_{\mathcal{N}}(\mathbf{v} \mid \sigma, \mu) \propto \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (v_i - \mu)^2 \right\}.$$

We would like to have an uninformative prior and instead assign the prior $p(\sigma)$ to be a uniform distribution of some given range $\sigma \sim \mathcal{U}(0, ub)$. Then the prior for the variance only has contribution from the Jacobian

$$p(\sigma^2) = p(\sigma) \left| \frac{\partial \sigma}{\partial \sigma^2} \right| \propto \sigma^{-1}. \tag{C.10}$$

The posterior (C.9) becomes

$$f(\sigma^2 \mid \mathbf{v}, \mu) \propto (\sigma)^{-(n+1)} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(v_i-\mu)^2\right\}$$
$$\propto (\sigma^2)^{-((n-1)/2+1)} \exp\left\{-\frac{(n-1)}{2\sigma^2}\frac{\sum_{i=1}^{n}(v_i-\mu)^2}{n-1}\right\},$$

so σ^2 is of the form of a scale–inv–χ^2 distribution with d degree of freedom and τ^2 the scale,

$$\frac{(\tau^2 d/2)^{d/2}}{\Gamma(d/2)} x^{-(d/2+1)} \exp\left\{-\frac{d\tau^2}{2x}\right\},$$

that equates to $d = n-1$ and

$$\tau^2 = \frac{\sum_{i=1}^{n}(v_i-\mu)^2}{n-1}$$

is the unbiased estimator for the variance of $\mathbf{v}$. Because sampling from $X \sim$ inv–$\chi^2(d, \tau^2)$ is equal to $(\tau^2 d)^{-1}X \sim$ inv–$\chi^2(d)$, the variance is

$$(\sum_{i=1}^{n}(v_t-\mu)^2)^{-1}\sigma^2 \sim \text{inv–}\chi^2(n-1). \tag{C.11}$$

But (C.10) is also true when $p(\sigma) \propto 1$, a uniform prior over $(0, \infty]$. So the bound ub on the standard deviation is only a restriction to guard against large σ, and subsequently large $\mathbf{v}$ where the MCMC procedure may fail.

C.4 AR(1) Conditionals

C.4.1 Dispersions

The multivariate normal distribution of d dimension is denoted as (C.12) with $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}, \text{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

$$\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{C.12}$$

Let's partition the random vector $\boldsymbol{X}$ into

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$$

of size q and $d-q$ respectively, with mean and covariance structure as

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Then the conditional mean and variance of X_1 given the realization $X_2 = \mathbf{x}_2$ can be expressed as

$$\mathbb{E}(\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \mathbf{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \tag{C.13}$$

$$\mathrm{Var}(\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \mathbf{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \tag{C.14}$$

(Anderson, 1984) which is summarized as

$$\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \mathbf{x}_2 \sim \mathcal{N}_q\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right). \tag{C.15}$$

We wish to find the conditional distribution of the dispersion v_t under an AR(1) model by conditioning on all the other dispersions $\mathbf{v}_{-t}$. This is equivalent to conditioning on both v_{t+1}, v_{t-1} as higher lag dispersion do not enter into the conditional. Let the vector $\boldsymbol{X} = (v_t\ v_{t+1}\ v_{t-1})$, which is distributed as (C.12) with $\boldsymbol{\mu} = 0$ and

$$\begin{aligned} \mathrm{Cov}(v_t, v_{t+1}) &= \phi\delta^2 \\ \mathrm{Cov}(v_{t+1}, v_{t-1}) &= \phi^2\delta^2, \end{aligned}$$

where $\delta^2 = (1-\phi^2)^{-1}\sigma^2$ such that

$$\boldsymbol{\Sigma} = \delta^2 \begin{bmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{bmatrix}. \tag{C.16}$$

Substituting the corresponding elements of (C.16) into (C.15) yields

$$v_t \mid v_{t+1}, v_{t-1}, \phi, \sigma^2 \sim \mathcal{N}\left(\frac{\phi(v_{t+1} + v_{t-1})}{1+\phi^2}, \frac{\sigma^2}{1+\phi^2}\right)$$

for $t = 2, 3, \ldots, T-1$ where the first and last element of $\mathbf{v}$ is

$$v_1 \mid v_2, \phi, \sigma^2 \sim \mathcal{N}\left(\phi v_2, \sigma^2\right) \tag{C.17}$$

and

$$v_T \mid v_{T-1}, \phi, \sigma^2 \sim \mathcal{N}\left(\phi v_{T-1}, \sigma^2\right). \tag{C.18}$$

C.4.2 Sampling Variance and Autoregressive coefficients

We first derive the general case of an AR(p) model before showing the specific case of an AR(1). The joint posterior of an AR(p) model with σ^2 and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$ is

$$f(\boldsymbol{\phi}, \sigma^2 \mid \mathbf{v}) \propto \mathcal{N}(\mathbf{v}_{1:p} \mid 0, \boldsymbol{\Sigma}) \left[\prod_{t=p+1}^{T} \mathcal{N}(v_t \mid \boldsymbol{\phi} \mathbf{v}_{t-1:t-p}, \sigma^2) \right] f(\sigma^2) f(\boldsymbol{\phi}) \tag{C.19}$$

where $\mathbf{v}_{t-1:t-p} = (v_{t-1}, v_{t-2}, \ldots, v_{t-p})^\top$ and $\boldsymbol{\Sigma}$ is the p dimension covariance matrix of the joint distribution, i.e.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{bmatrix} \tag{C.20}$$

where γ_j denotes the autocovariance of lag j. $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{p-1})$ is the first p element of the first column vector of $\mathbf{A} = \sigma^2(\mathbf{I}_{p\times p} - \mathbf{F} \otimes \mathbf{F})^{-1}$, where

$$\mathbf{F} = \begin{bmatrix} \boldsymbol{\phi}_{-p} & \boldsymbol{\phi}_p \\ 1 & 0 \end{bmatrix} \tag{C.21}$$

is the first order difference equation of a p order autoregressive process (Hamilton, 1994, chap. 1). Let $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Omega}$, the first term in (C.19) can be written as

$$\mathcal{N}(v_{1:p} \mid 0, \boldsymbol{\Sigma}) = (2\pi\sigma^2 \boldsymbol{\Omega})^{-1/2} \exp\left\{ -\frac{\mathbf{v}_{1:p}^T \boldsymbol{\Omega}^{-1} \mathbf{v}_{1:p}}{2\sigma^2} \right\}, \tag{C.22}$$

sampling σ^2 is a standard Gibbs sampling of a linear regression of (C.11)

$$\Big(\sum_{t=p+1}^{T} (v_t - \boldsymbol{\phi} \mathbf{v}_{t-1:t-p})^2 + \mathbf{v}_{1:p}^T \boldsymbol{\Omega}^{-1} \mathbf{v}_{1:p} \Big)^{-1} \sigma^2 \sim \text{inv–}\chi^2(T-1). \tag{C.23}$$

For $\boldsymbol{\phi}$, recognize that the second term in (C.19)

$$\prod_{t=p+1}^{T} \mathcal{N}(v_t \mid \boldsymbol{\phi} \mathbf{v}_{t-1:t-p}, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{\sum_{t=p+1}^{T} (v_t - \boldsymbol{\phi} \mathbf{v}_{t-1:t-p})^2}{2\sigma^2} \right\},$$

which is again a linear regression with $\boldsymbol{\beta} = \boldsymbol{\phi}$, $\mathbf{y} = (v_T, v_{T-1}, v_{p+1})^\top$ and $\mathbf{X} = \left[\mathbf{v}_{T-1:T-p}^\top ; \mathbf{v}_{T-2:T-p-1}^\top ; \dots ; \mathbf{v}_{p+p:p+1}^\top\right]$. The distribution of the regression coefficients can be found by applying standard OLS theory (i.e. (4.12)) with a Gibbs sampling step

$$\boldsymbol{\beta}^* \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}) \tag{C.24}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. Evidently, a sample from (C.24) does not guarantee stationarity and a proposal is only valid if the largest eigenvalues of (C.21) is smaller than 1. Then, sampling from (C.19) using a Metropolis–Hastings has acceptance probability

$$\min\left\{\frac{\mathcal{N}(\mathbf{v}_{1:p} \mid 0, \boldsymbol{\Sigma}^*)\left[q(\boldsymbol{\phi} \mid \boldsymbol{\phi}^*)\prod_{t=p+1}^{T}\mathcal{N}(v_t \mid \boldsymbol{\phi}^*\mathbf{v}_{t-1:t-p}, \sigma^2)\right]}{\mathcal{N}(\mathbf{v}_{1:p} \mid 0, \boldsymbol{\Sigma})\left[q(\boldsymbol{\phi}^* \mid \boldsymbol{\phi})\prod_{t=p+1}^{T}\mathcal{N}(v_t \mid \boldsymbol{\phi}\mathbf{v}_{t-1:t-p}, \sigma^2)\right]}, 1\right\}, \tag{C.25}$$

where $\boldsymbol{\Sigma}^*$ is the adjusted covariance matrix given $\boldsymbol{\phi}^*$. If we use the Gibbs proposal of (C.24), then the acceptance probability is reduced to

$$\min\left\{\frac{\mathcal{N}(\mathbf{v}_{1:p} \mid 0, \boldsymbol{\Sigma}^*)}{\mathcal{N}(\mathbf{v}_{1:p} \mid 0, \boldsymbol{\Sigma})}, 1\right\} \tag{C.26}$$

because (C.24) is a Gibbs step, i.e. the acceptance probability is 1, and it replaces the terms in the square brackets (both the nominator and denominator) of (C.25).

Now for the AR(1) model, we proceed by first using the knowledge of (C.24) and (C.26). We sample $\phi^{(t+1)}$ by first generating our proposal via

$$\phi^* \sim \mathcal{N}(\hat{\phi}, \hat{s}^{-1}\sigma^2)\mathbf{1}\left\{\phi \in (-1, 1)\right\}, \quad \hat{\phi} = \hat{s}^{-1}\sum_{t=2}^{T} v_t v_{t-1}, \quad \hat{s} = \sum_{t=1}^{T-1} v_t^2,$$

where $\mathbf{1}\left\{\cdot\right\}$ is an indicator function that ensure the AR process remains stationary given our proposal. Then accept ϕ^* with probability

$$\min\left\{\frac{\mathcal{N}(\mathbf{v}_1 \mid 0, (1-(\phi^*)^2)^{-1}\sigma^2)}{\mathcal{N}(\mathbf{v}_1 \mid 0, (1-\phi^2)^{-1}\sigma^2)}, 1\right\}. \tag{C.27}$$

For the variance, it is a direct application of (C.23). Substituting in the corresponding terms where $\gamma_0 = (1-\phi^2)^{-1}\sigma^2$ gives

$$(\sum_{t=2}^{T}(v_t - \phi v_{t-1})^2 + (1-\phi^2)^{-1}v_1^2)\sigma^2 \sim \text{inv–}\chi^2(T-1). \tag{C.28}$$

Appendix D

D.1 Graphs of Combined Result

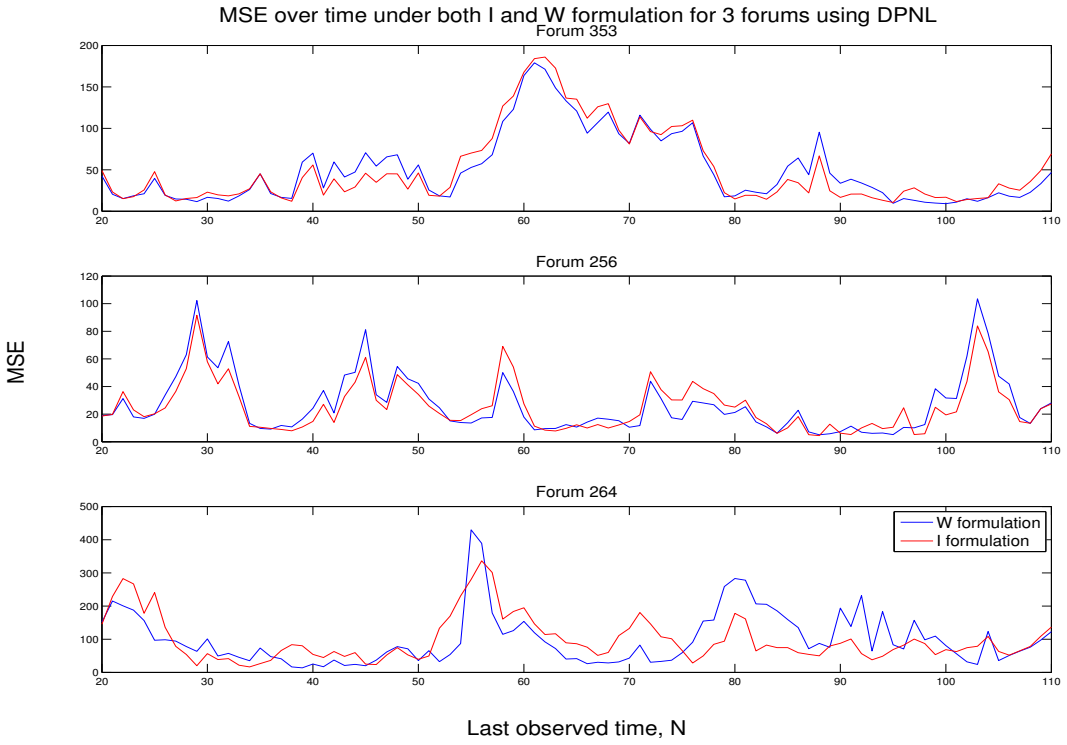


FIGURE D.1: The MSE over time between the two formulation for all three forums. Only the non–linear estimation of **SP** is demonstrated.

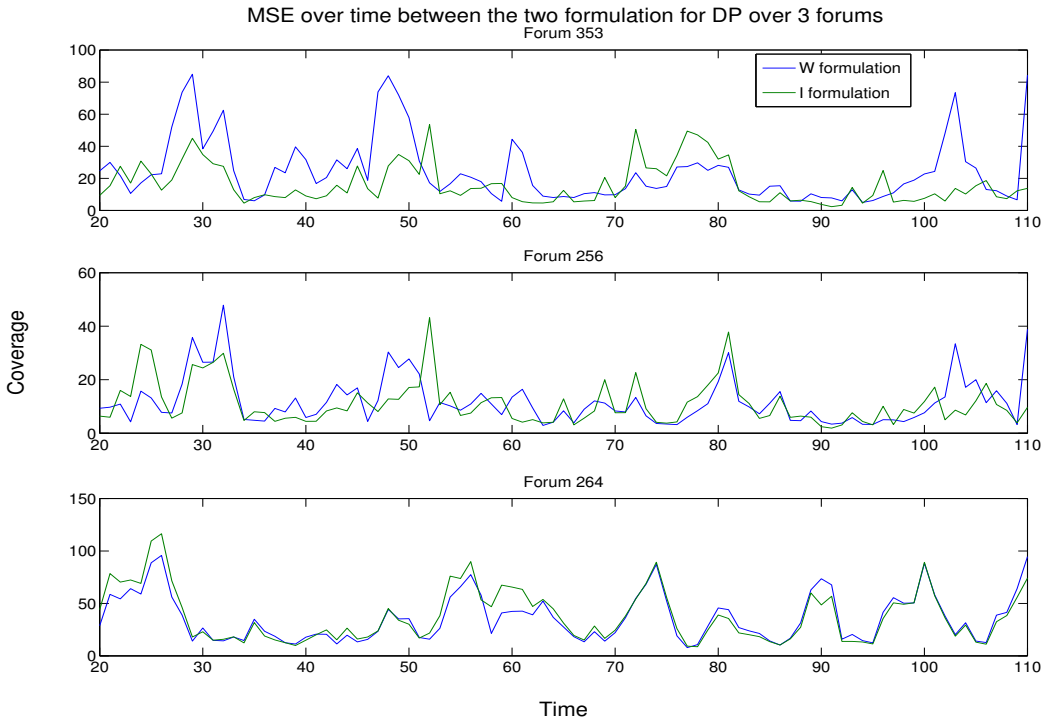


FIGURE D.2: MSE under the two formulation of **W** and **I** using **SP** for 3 forums over time

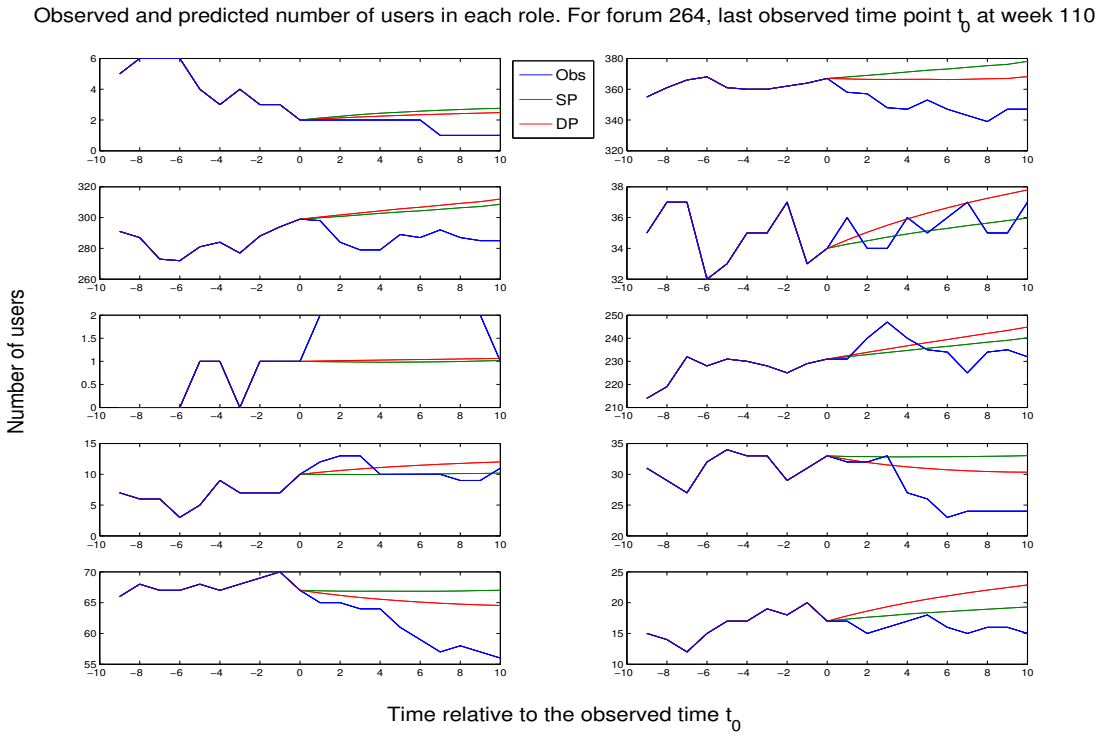


FIGURE D.3: Plots of all 10 roles of interests for the observed and prediction number of users using **DP** and **SP**. Time on the x–axis is relative to the last observed time point t_0, at week 110 such that 0 is observed.

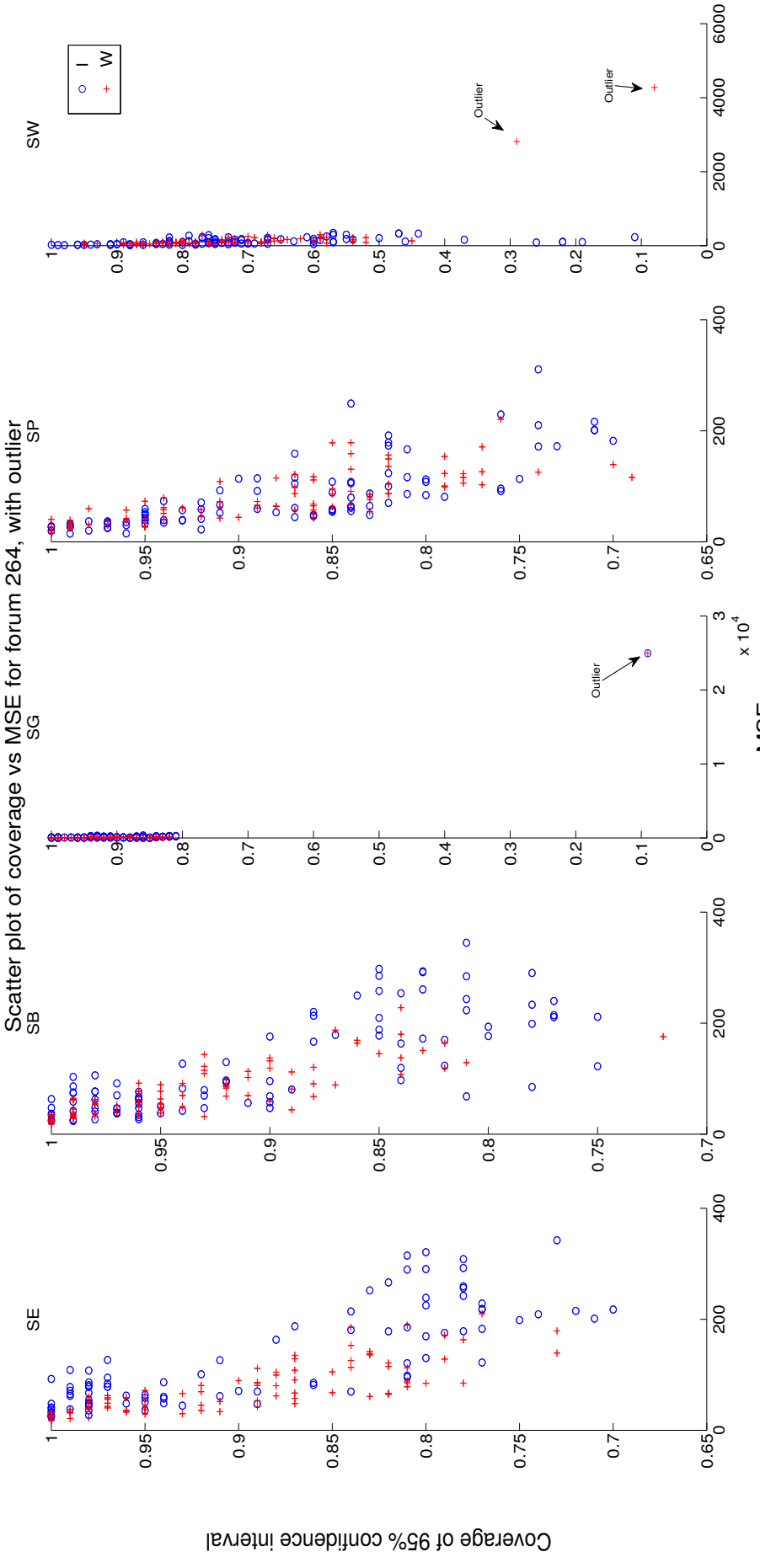


FIGURE D.4: The coverage against MSE of different methods and formulations over all 90 forecasts for forum 264

References

Agichtein, E., Liu, Y., and Bian, J. Modeling information-seeker satisfaction in community question answering. *ACM Transactions on Knowledge Discovery from Data*, 3(2):1–27, 2009.

Aitchison, J. and Ho, C. H. The Multivariate Poisson-Log Normal Distribution. *Biometrika*, 76(4):643–653, 1989.

Aitchison, J. and Shen, S. M. Logistic-Normal Distributions: Some Properties and Uses. *Biometrika*, 67(2):pp. 261–272, 1980.

Al-Osh, M. A. and Alzaid, A. A. First-Order Integer-Valued Autoregressive (INAR(1)) Process. *Journal of Time Series Analysis*, 8(3):261–275, 1987.

Albert, J. H. A Bayesian Analysis of a Poisson Random Effects Model for Home Run Hitters. *The American Statistician*, 46(4):246–253, 1992.

Anderson, A. and Huttenlocher, D. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 850–858, 2012.

Anderson, T. W. *An Introduction to Multivariate Statistical Analysis*. Wiley series in probability and mathematical statistics. Wiley, 2nd edition, 1984.

Anderson, T. W. and Goodman, L. A. Statistical Inference about Markov Chains. *Ann. Math. Statist.*, 28(1):89–110, 1957.

Andrews, D. F. and Mallows, C. L. Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 36 (1):99–102, 1974.

Anscombe, F. J. The Transformation of Poisson, Binomial and Negative-Binomial Data. *Biometrika*, 35(3/4):246–254, 1948.

Avramidis, A. N., Channouf, N., and L'Ecuyer, P. Efficient correlation matching for fitting discrete multivariate distributions with arbitrary marginals and normal-copula dependence. *INFORMS Journal on Computing*, 21(1):88–106, 2009.

Bahr, D. B., Browning, R. C., Wyatt, H. R., and Hill, J. O. Exploiting Social Networks to Mitigate the Obesity Epidemic. *Obesity*, 17(4):723–728, 2009.

Bailey, N. T. J. *The Mathematical Theory of Infectious Diseases and Its Applications*. Griffin, 2nd edition, 1975.

Banfield, J. D. and Raftery, A. E. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821, 1993.

Barabási, A.-L. and Albert, R. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.

Barbieri, M. M. and Berger, J. O. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897, 2004.

Bardenet, R., Doucet, A., and Holmes, C. Towards scaling up Markov chain Monte Carlo : an adaptive subsampling approach. In *Proceedings of The 31st International Conference on Machine Learning*, number 4, pages 405–413, 2014.

Barnard, J., McCulloch, R., and Meng, X.-l. Modeling Covariance Matrices In Terms Of Standard Deviations and Correlations, With Application to Shrinkage. *Statistica Sinica*, 10:1281–1311, 2000.

Bartlett, M. S. The Square Root Transformation in Analysis of Variance. *Supplement to the Journal of the Royal Statistical Society*, 3(1):68–78, 1936.

Boorman, S. A. and White, H. C. Social Structure from Multiple Networks. II. Role Structures. *American Journal of Sociology*, 81(6):1384–1446, 1976.

Brandtzæ g, P. B. Towards a unified Media-User Typology (MUT): A meta-analysis and review of the research literature on media-user typologies. *Computers in Human Behavior*, 26(5):940–956, 2010.

Brandtzæ g, P. B. Social Networking Sites: Their Users and Social Implications A Longitudinal Study. *Journal of Computer-Mediated Communication*, 17(4): 467–488, 2012.

Brauer, F. and Castillo-Chavez, C. *Mathematical Models in Population Biology and Epidemiology*. Springer, 1st edition, 2001.

Brauer, F., van den Driessche, P., and Wu, J. *Mathematical Epidemiology*. Springer Berlin Heidelberg, 2008.

Brockwell, P. J. and Davis, R. A. *Time Series: Theory and Methods*. Springer, 2nd edition, 2009.

Brown, L., Cai, T., Zhang, R., Zhao, L., and Zhou, H. The rootunroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability Theory and Related Fields*, 146(3-4):401–433, January 2009.

Bulmer, M. G. On Fitting the Poisson Lognormal Distribution to Species-Abundance Data. *Biometrics*, 30(1):101–110, 1974.

Cameron, A. and Trivedi, P. K. Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, 46(3):347–364, 1990.

Cameron, A. and Trivedi, P. K. *Regression Analysis of Count Data*, volume 41 of *Econometric Society monographs*. Cambridge University Press, 2nd edition, 1998.

Cario, M. C. and Nelson, B. L. Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix. Technical report, Department of Industrial Engineering and Mangement Sciences, Northwestern University, Evanston, Ill., 1997.

Casella, G. Empirical Bayes Gibbs sampling. *Biostatistics*, 2(4):485–500, 2001.

Casella, G. and George, E. I. Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174, 1992.

Cha, M., Mislove, A., and Gummadi, K. P. A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 721–730, New York, NY, USA, 2009. ACM.

Chan, J., Hayes, C., and Daly, E. M. Decomposing Discussion Forums and Boards Using User Roles. In *ICWSM*, 2010.

Chan, K. S. and Ledolter, J. Monte Carlo EM Estimation for Time Series Models Involving Counts. *Journal of the American Statistical Association*, 90(429): 242–252, 1995.

Channouf, N. and L'Ecuyer, P. Fitting a Normal Copula for a Multivariate Distribution with Both Discrete and Continuous Marginals. In *Proceedings of the Winter Simulation Conference, 2009*, WSC '09, pages 352–358. Winter Simulation Conference, 2009.

Chib, S. Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241, 1998.

Chib, S. and Greenberg, E. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335, 1995.

Chib, S. and Winkelmann, R. Markov Chain Monte Carlo Analysis of Correlated Count Data. *Journal of Business & Economic Statistics*, 19(4):428–435, 2001.

Chib, S., Greenberg, E., and Winkelmann, R. Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics*, 86(1):33–54, June 1998.

Christakis, N. A. and Fowler, J. H. Social Network Sensors for Early Detection of Contagious Outbreaks. *PLoS ONE*, 5(9):e12948, 2010.

Cohen, A. C. J. Estimating the Mean and Variance of Normal Populations from Singly Truncated and Doubly Truncated Samples. *The Annals of Mathematical Statistics*, 21(4):557–569, 1950.

Cook, R. D. Influential Observations in Linear Regression. *Journal of the American Statistical Association*, 74(365):169–174, 1979.

Damien, P. and Walker, S. G. Sampling Truncated Normal, Beta, and Gamma Densities. *Journal of Computational and Graphical Statistics*, 10(2):206–215, 2001.

Davis, J. A. Clustering and structural balance in graphs. *Human Relations*, 20 (2):181–187, 1967.

Delamater, J. D. and Myers, D. J. *Social Psychology*. Wadsworth Publishing, 7th edition, 2010.

DeLone, W. H. and McLean, E. R. Information Systems Success: The Quest for the Dependent Variable. *Information Systems Research*, 3(1):60–95, 1992.

DeLone, W. H. and McLean, E. R. The DeLone and McLean Model of Information Systems Success : A Ten-Year Update. *Journal of Management Information Systems*, 19(4):9–30, 2003.

Devroye, L. *Non-Uniform Random Variate Generation*. Springer, 1986.

Dhillon, I. S., Guan, Y., and Kulis, B. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 551–556, New York, NY, USA, 2004. ACM.

Ding, C., He, X., and Simon, H. D. On the equivalence of nonnegative matrix factorization and spectral clustering. In *in SIAM International Conference on Data Mining*, 2005.

Duan, N. Smearing Estimate: A Nonparametric Retransformation Method. *Journal of the American Statistical Association*, 78(383):605–610, 1983.

Durbin, J. and Watson, G. S. Testing for Serial Correlation in Least Squares Regression. III. *Biometrika*, 58(1):1–19, 1971.

Durbin, J. and Koopman, S. J. *Time Series Analysis by State Space Methods*. Clarendon Press, 2001.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least Angle Regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Embrechts, P., J McNei, A., and Straumann, D. *Correlation and Dependence in Risk Mangement: Properties and Pitfalls*. Cambridge University Press, Cambridge, 2002.

Erdös, P. and Rényi, A. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

Everitt, B. S., Landau, S., and Leese, M. *Cluster Analysis*. Wiley, 2009.

Facebook. Facebook statistic, 2013.

Fraley, C. and Raftery, A. E. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458): 611–631, 2002.

Fraser, C. Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS ONE*, 2(8), 2007.

Freeman, L. C. Visualizing Social Networks. *Journal of Social Structure*, 1(1), 2000.

Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical software*, 33(1), 2010.

Frühwirth-Schnatter, S. and Wagner, H. Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *Biometrika*, 93(4):827–841, 2006.

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

Gelman, A. Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analysis*, 1(3):515–533, 2006.

Gelman, A., Bois, F., and Jiang, J. Physiological Pharmacokinetic Analysis Using Population Modeling and Informative Prior Distributions. *Journal of the American Statistical Association*, 91(436):1400–1412, 1996.

Gelman, A., Carlin, J. B., Stern, H., and Rubin, D. B. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd edition, 2003.

Geritz, S. a. H. and Kisdi, E. On the mechanistic underpinning of discrete-time population models with complex dynamics. *Journal of theoretical biology*, 228 (2):261–9, May 2004.

Ghosh, S. and Henderson, S. G. Behavior of the NORTA method for correlated random vector generation as the dimension increases. *ACM Transactions on Modeling and Computer Simulation*, 13(3):276–294, July 2003.

Ghosh, S. and Henderson, S. G. Chessboard Distributions and Random Vectors with Specified Marginals and Covariance Matrix. *Operations Research*, 50(5): 820–834, 2002.

Gibson, G. J. Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology*, 15(1):19–40, March 1998.

Girolami, M. and Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Godfrey, K. *Compartmental model and its application*. Academic Press, 1983.

Goeman, J. J. L1 penalized estimation in the Cox proportional hazards model. *Biometrical journal Biometrische Zeitschrift*, 52(1):70–84, 2010.

Goffman, E. *The Presentation of Self in Everyday Life*, volume 21 of *Anchor books*. Doubleday, 1959.

Golder, S. A. and Donath, J. Social roles in electronic communities. *Internet Research*, 5:1–25, 2004.

Gouriéroux, C. and Monfort, A. *Simulation Based Econometric Methods*. Oxford University Press, 1997.

Granovetter, M. S. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

Granovetter, M. S. The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*, 1(1983):201–233, 1983.

Gutmann, H.-M. A Radial Basis Function Method for Global Optimization. *Journal of Global Optimization*, 19(3):201–227, 2001.

Halperin, M. Estimation in the Truncated Normal Distribution. *Journal of the American Statistical Association*, 47(259):457–465, 1952.

Hamilton, J. D. *Time Series Analysis*. Princeton University Press, 1994.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, March 2007.

Hanneke, S., Fu, W., and Xing, E. P. Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605, 2010.

Hansell, S. Weak Ties , and the Groups , Cooperative of Peer Friendships Integration. *Social Psychology Quarterly*, 47(4):316–328, 1984.

Hardin, J. and Hilbe, J. *Generalized Linear Models and Extensions*. Stata Press, 2nd edition, 2007.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2009.

Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.

Hautz, J., Hutter, K., Fuller, J., Matzler, K., and Rieger, M. How to Establish an Online Innovation Community? the Role of Users and Their Innovative Content. *2014 47th Hawaii International Conference on System Sciences*, 0:1–11, 2010.

Hayashi, F. *Econometrics*. Princeton University Press, 2000.

Himelboim, I., Gleave, E., and Smith, M. Discussion catalysts in online political discussions: Content importers and conversation starters. *Journal of Computer-Mediated Communication*, 14(4):771–789, July 2009.

Hobert, J. P., Roy, V., and Robert, C. P. Improving the Convergence Properties of the Data Augmentation Algorithm with an Application to Bayesian Mixture Modeling. *Statistical Science*, 26(3):332–351, 2011.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97 (460):1090–1098, December 2002.

Hoffman, M. D. and Gelman, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

Holland, P. W. and Leinhardt, S. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76 (373):33–50, 1981.

Holt, C. C. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10, January 2004.

Iriberri, A. and Leroy, G. A Life-cycle Perspective on Online Community Success. *ACM Computing Surveys*, 41(2):11:1—-11:29, February 2009.

Jacquez, J. A. *Compartmnetal analysis in biology and medicine*. Elsevvier Science, 1st edition, October 1972.

Jain, A. K., Murty, M. N., and Flynn, P. J. Data Clustering : A Review. *ACM Computing Surveys*, 31(3):264–323, 1999.

James, W. and Stein, C. Estimation with Quadratic Loss, 1961.

Jin-Guan, D. and Yuan, L. The Integer-Valued Autoregressive (INAR(p)) Model. *Journal of Time Series Analysis*, 12(2):129–142, 1991.

Johnson, N. L., Kotz, S., and Balakrishnan, N. *Discrete Multivariate Distributions*. Wiley, 1997.

Jones, D., Schonlau, M., and Welch, W. J. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global optimization*, (13):455–492, 1998.

Jung, R. C. and Winkelmann, R. Two aspects of labor mobility: A bivariate Poisson regression approach. *Empirical Economics*, 18(3):543–556, 1993.

Jung, R. C., Kukuk, M., and Liesenfeld, R. Time series of count data: modeling, estimation and diagnostics. *Computational Statistics & Data Analysis*, 51(4): 2350–2364, 2006.

Karlis, D. An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics*, 30(1):63–77, 2003.

Karlis, D. and Ntzoufras, I. Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52 (3):381–393, October 2003.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. Markov Chain Monte Carlo in Practice: A Roundtable Discussion. *The American Statistician*, 52(2): 93–100, 1998.

Kaufman, L. and Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Blackwell, 2nd edition, 2005.

King, J. L., Grinter, R. E., and Pickering, J. M. The rise and fall of netville: The saga of a cyberspace construction boomtown in the great divide. In *Culture of the Internet*, pages 3–34. Psychology Press, 1997.

Krivitsky, P. N. Exponential-family random graph models for valued networks. *Electronic Journal of Statistics*, 6:1100–1128, 2012.

Kruskal, W. H. Ordinal Measures of Association. *Journal of the American Statistical Association*, 53(284):814–861, 1958.

Law, R. and Blackford, J. C. Self-Assembling Food Webs: A Global Viewpoint of Coexistence of Species in Lotka-Volterra Communities. *Ecology*, 73(2):567–578, 1992.

Lee, K., Marin, J.-M., Mengersen, K., and Robert, C. P. Bayesian Inference on Mixtures of Distributions. *Handbook of Statistics*, 25(05):24, 2008.

Lee, S.-i., Lee, H., Abbeel, P., and Ng, A. Y. Efficient L1 Regularized Logistic Regression. In *In Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, 2006.

Lewis, R. M., Shepherd, A., and Torczon, V. Implementing Generating Set Search Methods for Linearly Constrained Minimization. *SIAM Journal on Scientific Computing*, 29(6):2507–2530, October 2007.

Lindley, D. V. and Smith, A. F. M. Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34(1):1–41, 1972.

Lotka, A. J. *The Elements of Physical Biology*. Williams & Williams Co, 1925.

Macqueen, J. Some Methods For Classification And Analysis Of Multivariate Observation. *Proceedings of the Berkley Symposium on Mathematical Statistics and Probability*, 1(233):281–297, 1967.

Maia, M., Almeida, J., and Almeida, V. Identifying user behavior in online social networks. In *Proceedings of the 1st workshop on Social network systems SocialNets 08*, pages 1–6. ACM Press, 2008.

Marin, J., Mengersen, K., and Robert, C. P. Bayesian Modelling and Inference on Mixtures of Distributions. In Rao, C. R. and Dipak, D., editors, *Handbook of statistics*, volume 25, chapter 16, pages 15840–15845. Elsevier, 2005.

Matis, J. H. and Wehrly, T. E. Stochastic Models of Compartmental Systems. *International Biometric Society*, 35(1):199, March 1979.

McCabe, B. P. M., Martin, G. M., and Harris, D. Efficient probabilistic forecasts for counts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):253–272, 2011.

McCullagh, P. Quasi-Likelihood Functions. *The Annals of Statistics*, 11(1):59–67, 1983.

McCullagh, P. and Nelder, J. A. *Generalized Linear Models*. Chapman and Hall, 2nd edition, 1989.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. *Generalized, Linear, and Mixed Models*. Wiley-Interscience, 2nd edition, 2008.

McLachlan, G. J. and Peel, D. *Finite Mixture Model*. Wiley, 2000.

Meng, X.-L. and van Dyk, D. The EM Algorithm–An Old Folk-Song Sung to a Fast New Tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):511–567, 1997.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6), 1953.

Moreno, J. L. *Who shall survive?: A new approach to the problem of human interrelations*. Nervous and Mental Disease Publishing Co., 1934.

Morris, M. and Ogan, C. The Internet as Mass Medium. *Journal of Communication*, 46(1):39–50, 1996.

Narsky, I. and Porter, F. C. *Statistical Analysis Techniques in Particle Physics: Fits, Density Estimation and Supervised Learning*. Wiley, 2013.

Neal, R. M. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

Nelson, J. F. Multivariate Gamma-Poisson Models. *Journal of the American Statistical Association*, 80(392):828–834, 1985.

Nelson, R. B. *An Introduction to Copulas. Lecture Notes in Statistics*. Springer-Verlag, New York, 1999.

Newman, M., Forrest, S., and Balthrop, J. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101, September 2002.

Newman, M. E. J. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167, 2003.

Nolker, R. D. and Zhou, L. Social computing and weighting to identify member roles in online communities. In *The 2005 IEEEWICACM International Conference on Web Intelligence WI05*, volume 28 of *WI '05*, pages 87–93. IEEE, 2005.

Nonnecke, B. and Preece, J. Lurker demographics: Counting the silent. In Turner, T., Szwillus, G., Czerwinski, M., Paterno, F., and Pemberton, S., editors, *Proceedings of the ACM Conference on Human Factors in Computing Systems*, volume 2 of *CHI '00*, pages 73–80. ACM New York, NY, USA, ACM, 2000.

O'Malley, A. J. and Zaslavsky, A. M. Domain-Level Covariance Analysis for Multilevel Survey Data With Structured Nonresponse. *Journal of the American Statistical Association*, 103(484):1405–1418, 2008.

Osborne, M. R., Presnell, B., and Turlach, B. A. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3): 389–403, 2000.

Ovadia, S. The Role of Big Data in the Social Sciences. *Behavioral & Social Sciences Librarian*, 32(2):130–134, 2013.

Park, M. Y. and Hastie, T. L1 - regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007.

Park, T. and Casella, G. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Poor, H. V. and Hadjiliadis, O. *Quickest Detection*. Cambridge University Press, 2008.

Preece, J. *Online Communities: Designing Usability and Supporting Sociability*. John Wiley & Sons, 2000.

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

Robert, C. P. and Casella, G. *Monte Carlo Statistical Methods*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2nd edition, 2005.

Roberts, G. O. and Rosenthal, J. S. Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

ROBUST. Risk and Opportunity management of huge-scale BUSiness communiTy cooperation. *Business*, pages 1–119, 2009.

Rogers, E. M. *Diffusion of Innovations*. Simon & Schuster International, 5th edition, 2003.

Rowe, M., Angeletou, S., and Alani, H. Anticipating Discussion Activity on Community Forums. In *2011 IEEE Third Intl Conference on Privacy Security Risk and Trust and 2011 IEEE Third Intl Conference on Social Computing*, pages 315–322. IEEE, 2011.

Rowe, M., Fernandez, M., Angeletou, S., and Alani, H. Community Analysis through Semantic Rules and Role Composition Derivation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 18, 2013.

Rubin, D. B. The Bayesian Bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.

Sampson, S. Crisis in a cloister. *Unpublished doctoral dissertation, Department of Socialogy, Cornell University*, 1969.

SAP. Recognition Program, 2012.

Schaeffer, S. E. Graph clustering. *Computer Science Review*, 1(1):27–64, August 2007.

Scheffer, M. and van Nes, E. H. Self-organized similarity, the evolutionary emergence of groups of similar species. *Proceedings of the National Academy of Sciences of the United States of America*, 103(16):6230–5, April 2006.

Scheffer, M., Carpenter, S., Foley, J. A., Folke, C., and Walker, B. Catastrophic shifts in ecosystems. *Nature*, 413(6856):591–6, October 2001.

Schmidt, A. and Rodriguez, M. Modelling multivariate counts varying continuously in space. In *Bayesian Statistics 9*, pages 611–638. Oxford University Press, 2011.

Schölkopf, B., Smola, A., and Müller, K. Kernel principal component analysis. In *Artificial Neural NetworksICANN'97*, pages 327–352. MIT Press, 1997.

Schölkopf, B., Smola, A., and Müller, K. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

Shen, H. and Huang, J. Z. Interday Forecasting and Intraday Updating of Call Center Arrivals. *Manufacturing & Service Operations Management*, 10(3):391–410, June 2008.

Sneath, P. H. The application of computers to taxonomy. *Journal of General Microbiology*, 17(1):201–226, 1957.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. New Specifications For Exponential Random Graph Models. *Sociological Methodology*, 36 (1):99–153, 2006.

Soong, T. T. and Dowdee, J. W. Pharmacokinetics with uncertainties in rate constants III: the inverse problem. *Mathematical Biosciences*, 19(34):343–353, 1974.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

Stein, C. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution, 1956.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

Tönnies, F. *Gemeinschaft und Gesellschaft*. Fues, Leipzig, 1887.

Toral, S. L., Martínez-Torres, M. R., Barrero, F., and Cortés, F. An empirical study of the driving forces behind online communities. *Internet Research*, 19 (4):378–392, 2009.

Trusov, M., Bucklin, R., and Pauwels, K. Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. *Journal of Marketing*, 73(5):90–102, 2009.

Tsionas, E. G. Bayesian analysis of the multivariate poisson distribution. *Communications in Statistics - Theory and Methods*, 28(2):431–451, 1999.

Turchin, P. *Complex population dynamics: A Theoretical/Empirical Synthesis*. Princeton University Press, 2003.

van Duijn, M. A. J., Snijders, T. A. B., and Zijlstra, B. J. H. p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, 58(2): 234–254, 2004.

Vanderbei, R. J. *Linear Programming: Foundations and Extensions*, volume 49. Kluwer, internatio edition, March 1998.

Wang, Y. J. and Wong, G. Y. Stochastic Blockmodels for Directed Graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

Wasserman, S. and Faust, K. *Social Network Analysis*. Cambridge University Press, 1994.

Wasserman, S. and Pattison, P. E. Logit models and logistic regressions for social networks: II. Multivariate relations. *The British journal of mathematical and statistical psychology*, 52 (Pt 2):169–193, November 1999.

Wasserman, S. and Pattison, P. E. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p. *Psychometrika*, 61(3): 401–425, September 1996.

Welser, H. T., Gleave, E., Fisher, D., and Smith, M. Visualizing the Signatures of Social Roles in Online Discussion Groups. *Journal of Social Structure*, 8(2): 1–32, 2007.

West, M., Harrison, P. J., and Migon, H. S. Dynamic Generalized Linear Bayesian Models and Forecasting. *Journal of American Statistical Association*, 80:73–96, 1985.

White, H. C., Boorman, S. A., and Breiger, R. L. Social-structure from multiple networks. I: Blockmodels of roles and positions. *American Journal of Sociology*, 81(4):730–780, 1976.

Whittaker, S., Terveen, L., Hill, W., and Cherny, L. The dynamics of mass interaction. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work CSCW 98*, volume 29, pages 257–264. ACM Press, 1998.

Wild, S. M. and Shoemaker, C. Global Convergence of Radial Basis Function Trust-Region Algorithms for Derivative-Free Optimization. *SIAM Review*, 55 (2):349–371, January 2013.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2007.

Yoshida, T., Jones, L. E., Ellner, S. P., Fussmann, G. F., and Hairston, N. G. Rapid evolution drives ecological dynamics in a predator-prey system. *Nature*, 424(6946):303–6, July 2003.

Yu, J.-w. and Tian, G.-l. Efficient algorithms for generating truncated multivariate normal distributions. *Acta Mathematicae Applicatae Sinica, English Series*, 27 (4):601–612, 2011.

Yu, Y. and Meng, X.-L. To Center or Not to Center: That Is Not the Questio-nAn AncillaritySufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570, 2011.

Zeger, S. L. and Karim, M. R. Generalized Linear Models With Random Effects; A Gibbs Sampling Approach. *Journal of the American Statistical Association*, 86(413):79–86, 1991.

Zeger, S. L. and Qaqish, B. Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, 44(4):1019–1031, 1988.