

# Inference for epidemics on networks

Brock Hermans

*Thesis submitted for the degree of*

*Masters of Philosophy*

*in*

*Statistics and Applied Mathematics*

*at*

*The University of Adelaide*

*(Faculty of Engineering, Computer and Mathematical Sciences)*

School of Mathematical Sciences



THE UNIVERSITY  
of ADELAIDE

May 14, 2016



# Contents

<b>Signed Statement</b>	<b>x</b>
<b>Acknowledgements</b>	<b>xi</b>
<b>Abstract</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Background material</b>	<b>5</b>
2.1 Maximum likelihood estimate for $\beta$ , based on final epidemic size data . . .	7
<b>3 Gilbert-Erdős-Rényi network</b>	<b>11</b>
3.1 Network theory . . . . .	11
3.1.1 Epidemics on networks . . . . .	13
3.1.2 Gilbert-Erdős-Rényi network . . . . .	15
3.2 Assumed knowledge of the network . . . . .	18
3.3 A model based on an auxiliary estimate . . . . .	19
3.3.1 Auxiliary likelihoods and auxiliary estimates . . . . .	20
3.3.2 Fitting an auxiliary estimate model . . . . .	23
3.3.3 Reducing the model . . . . .	28
3.3.4 Random effects and the issue of collinearity . . . . .	33
3.3.5 Alternate model selection . . . . .	40
3.3.6 Assumption checking . . . . .	47
3.4 Fitting a regression spline to final epidemic sizes . . . . .	51
3.4.1 From MARS to <code>earth</code> . . . . .	51
3.4.2 Applying MARS to the final epidemic sizes . . . . .	55
3.5 Comparison of the models . . . . .	58

3.6	Discussion . . . . .	62
<b>4</b>	<b>Keeling network</b>	<b>67</b>
4.1	The Keeling network . . . . .	67
4.2	Auxiliary estimate vs regression splines . . . . .	70
4.2.1	Principal Component Analysis (PCA) [23] . . . . .	77
4.3	Further model analysis and checking . . . . .	90
4.4	Discussion . . . . .	98
<b>5</b>	<b>Model mis-specification</b>	<b>99</b>
5.1	Applying the Gilbert-Erdős-Rényi model to Keeling data . . . . .	101
5.2	Applying the Keeling model to Gilbert-Erdős-Rényi data . . . . .	103
5.3	Improving the Gilbert-Erdős-Rényi model . . . . .	106
5.4	Discussion . . . . .	108
<b>6</b>	<b>Building a model for time-series data</b>	<b>109</b>
6.1	Maximum likelihood estimate for time-series data . . . . .	110
6.2	Building an auxiliary estimate model . . . . .	111
6.3	Extensions . . . . .	118
<b>7</b>	<b>Conclusion</b>	<b>120</b>
7.1	Diary-based studies . . . . .	123
7.2	Summary . . . . .	124
<b>A</b>	<b>Model errors</b>	<b>125</b>
<b>B</b>	<b>Keeling plots for the auxiliary estimate model assumptions</b>	<b>128</b>
	<b>Bibliography</b>	<b>134</b>

# List of Tables

3.1	<i>Model 3.1, Model 3.2, and Model 3.3 and their corresponding normalised residual sum of squares (NRSS) and residual mean squared error (RMSE).</i>	28
3.2	<i>The estimated regression coefficients for the terms after we reduce the model on a <math>p</math>-value criterion.</i>	30
3.3	<i>The model covariates with associated VIF values.</i>	34
3.4	<i>The process for removing terms based on the highest VIF with no concern for breaking the Principle of Marginality.</i>	35
3.5	<i>The process for removing terms based on the highest VIF where we never remove a term that violates the Principle of Marginality.</i>	36
3.6	<i>Errors for the best model so far (3.4) and the three models reduced on a VIF criterion. The errors are normalised residual sum of squares (NRSS) and residual mean squared error (RMSE)</i>	36
3.7	<i>Normalised residual sum of squares (NRSS) and residual mean squared error (RMSE) for the 10 auxiliary estimate models (built from ten independent training set simulations) based on the same test set.</i>	39
3.8	<i>The value of <math>\beta_{min}</math> used in Algorithm 2 with the corresponding maximum likelihood estimate and likelihood value that we obtain from Algorithm 2</i>	41
3.9	<i>Error terms for the <math>\beta</math>-<math>p</math>-restriction model, and the four final epidemic size restriction models. The errors include the normalised residual sum of squares (NRSS) and residual mean squared error (RMSE)</i>	44
3.10	<i>The minimum and maximum values for <math>\beta</math> that give an average final epidemic size between 10 and 90. The ‘-’ refer to not having a minimum or maximum <math>\beta</math> within the range 1 to 7; this means that we don’t simulate any data for <math>p = 0.05, 0.1</math> or <math>0.15</math>.</i>	46
3.11	<i>The two error terms for the two models with <math>k = 1, \dots, 10</math>.</i>	58

3.12	<i>Residual mean squared error for both approaches, for each value of <math>p</math>.</i>	60
4.1	<i>Error terms for the three polynomial models, where NRSS=Normalised residual sum of squares, and RMSE=Residual mean squared error.</i>	73
4.2	<i>Regression estimates for the auxiliary estimate approach for both the Keeling and Gilbert-Erdős-Rényi networks.</i>	74
4.3	<i>Clusters for the number of pairs, triples and triangles.</i>	84
4.4	<i>Comparison of the three auxiliary estimate models, given by splitting the training set into three sub-datasets. All models are reduced using the <math>p</math>-value criteria.</i>	85
4.5	<i>The error terms for splitting based on different network and different numbers of clusters, for the auxiliary estimate model.</i>	87
4.6	<i>The error terms for splitting based on different network and different numbers of clusters, for the regression splines model.</i>	89
4.7	<i>The two error terms for the two Keeling models with <math>k = 1, \dots, 10</math>.</i>	93
5.1	<i>Residual mean squared errors for both modelling approaches and both network types.</i>	100
5.2	<i>Comparison of the three models for the Gilbert-Erdős-Rényi auxiliary estimate approach.</i>	107
6.1	<i>Example of epidemic time-series data.</i>	109
6.2	<i>Model coefficients for the bivariate model.</i>	113
A.1	<i>Table of different models used in Chapter 3 including their normalised sum of squared error and residual mean square error.</i>	126
A.2	<i>Table of different models used in Chapter 4 including their normalised sum of squared error and residual mean square error.</i>	127

# List of Figures

2.1	<i>Black and Ross' co-loxicographical ordering of the SIR states, for a population of <math>n = 3</math> (from Back and Ross' paper [6]). . . . .</i>	9
3.1	<i>A network with four nodes, and some edges between these nodes. . . . .</i>	13
3.2	<i>Upper plot: A sample network with 5 nodes, 5 edges, and one infected node; Lower plot: Node 3 is infected by Node 2. . . . .</i>	16
3.3	<i>A network with 10 nodes . . . . .</i>	19
3.4	<i><math>\beta</math> against <math>\beta_{Aux}</math>, separated by the value of <math>p</math>. . . . .</i>	24
3.5	<i>The true value of <math>\beta</math> against the predicted value of <math>\beta</math> using Model 3.1. . . . .</i>	27
3.6	<i>Residuals vs betaAux for the simple model. . . . .</i>	27
3.7	<i>Top plot: Cross validation error for all seven models and values of <math>k = 2, \dots, 10</math>; Bottom plot: Cross validation error for the two models with smallest cross validation error (Full model and PVal model). . . . .</i>	32
3.8	<i>The GCV sum of squares against the value of <math>\lambda</math>. . . . .</i>	38
3.9	<i>Box-plot of the 10 estimates for the regression coefficients, separated by the different covariates. . . . .</i>	38
3.10	<i>Residuals vs Fitted values plot for the auxiliary estimate model. . . . .</i>	48
3.11	<i>Residuals vs covariates plot for the auxiliary estimate model. . . . .</i>	49
3.12	<i>Normal Q-Q plot for the auxiliary estimate model. . . . .</i>	50
3.13	<i>Left plot: Example of regressing <math>y</math> on <math>x</math> using simple linear regression; Middle plot: Example of regressing <math>y</math> on <math>x</math> with a quadratic and cubic term; Right plot: Example of regressing <math>y</math> on <math>x</math> using regression splines. . . . .</i>	53
3.14	<i>The results for a regression spline with squared terms. . . . .</i>	54
3.15	<i>The MARS model with interaction terms between the final epidemic sizes and network properties. . . . .</i>	56

3.16	<i>The MARS model with interaction terms between the final epidemic sizes and network properties and interactions within the final epidemic sizes. . .</i>	57
3.17	<i>Normalised residual sum of squares against varying numbers of final epidemic sizes, separated by the model used. . . . .</i>	59
3.18	<i><math>\beta_{Aux}</math> against <math>\beta</math>, separated by <math>p</math>, where <math>\beta_{Aux}</math> is calculated using only one final epidemic size. . . . .</i>	60
3.19	<i>Box-plots of squared error, separated by <math>p</math> and <math>\beta_{Aux}</math>, for auxiliary estimate approach. . . . .</i>	63
3.20	<i>Box-plots of normalised squared error, separated by <math>p</math> and <math>\beta_{Aux}</math>, for auxiliary estimate approach. . . . .</i>	64
3.21	<i>Box-plots of raw error, separated by <math>p</math> and <math>\beta_{Aux}</math>, for auxiliary estimate approach. . . . .</i>	65
4.1	<i>Residuals vs betaAux for the simple model with no polynomial terms and interactions between betaAux and the three network properties. . . . .</i>	72
4.2	<i>The true value of <math>\beta</math> against the predicted value using the <math>\beta_{Aux}</math> approach, separated along the rows by <math>\delta</math> and along the columns by <math>\alpha</math>. . . . .</i>	76
4.3	<i>The true value of <math>\beta</math> against the predicted value using regression splines, separated along the rows by <math>\delta</math> and along the columns by <math>\alpha</math>. . . . .</i>	76
4.4	<i>11 plots of <math>\beta</math> against <math>\beta_{4,2}^*</math> coloured by a range of values for <math>f</math>. . . . .</i>	78
4.5	<i>Two dimensional plots of the principal components. Left plot: PC1 against PC2; middle plot: PC1 against PC3; right plot: PC2 against PC3. . . . .</i>	80
4.6	<i>Plot of the three principal components. . . . .</i>	81
4.7	<i>Top plot: 3-dimensional PCA plot, coloured by <math>f</math>; Middle plot: 3-dimensional PCA plot, coloured by <math>\delta</math>; Bottom plot: 3-dimensional PCA plot, coloured by <math>\alpha</math>. Red represents <math>f, \delta, \alpha = 0.1, 0.2, 0.3</math>, blue represents <math>f, \delta, \alpha = 0.4, 0.5, 0.6</math> and green represents <math>f = 0.7, 0.8, 0.9</math> and <math>\delta, \alpha = 0.7, 0.8, 0.9, 1</math>. . . . .</i>	82
4.8	<i>Top plot: 3-dimensional PCA plot, coloured by pairs; Middle plot: 3-dimensional PCA plot, coloured by triples; Bottom plot: 3-dimensional PCA plot, coloured by triangles. Red represents the smallest third of the network property (Cluster 1), blue represents middle third (Cluster 2), and top represent highest third (Cluster 3). . . . .</i>	83

4.9	<i>Normalised residual sum of squares against varying numbers of final epidemic sizes, separated by the different models. . . . .</i>	93
4.10	<i>Box-plots of raw errors for auxiliary estimate model. Plots are separated by their number of pairs into three groups (along the columns) and the estimated <math>f</math> (along the rows). The values of 1, 2, and 3 along the columns represents whether the number of pairs is <math>\leq 3923</math>, <math>&gt; 3923</math> and <math>\leq 6676</math> or <math>&gt; 6676</math>. . . . .</i>	97
5.1	<i>Predicted values of <math>\beta</math> for Keeling models against Gilbert-Erdős-Rényi model. Darker shaded points represent heavy clustering of points, and lighter shaded points represent low clustering of points. . . . .</i>	101
6.1	<i>True <math>\tilde{\beta}</math> against the predicted value of <math>\tilde{\beta}</math> using the linear model, <math>\tilde{\beta}_{6.1}^*</math>. Plots are separated by <math>p</math>. . . . .</i>	114
6.2	<i>True <math>\gamma</math> against the predicted value of <math>\gamma</math> using the linear model, <math>\gamma_{6.1}^*</math>. Plots are separated by <math>p</math>. . . . .</i>	114
6.3	<i>Box-plots for <math>\tilde{\beta}</math>, separated by <math>p</math> and <math>\beta_{Aux}</math>. . . . .</i>	116
6.4	<i>Box-plots for <math>\gamma</math>, separated by <math>p</math> and <math>\gamma_{Aux}</math>. . . . .</i>	117
6.5	<i>Plot of true <math>\beta</math> (<math>\tilde{\beta}/\gamma</math>) against predicted value of <math>\beta</math> (<math>\tilde{\beta}_{6.1}^*/\gamma_{6.1}^*</math>), separated by <math>p</math>. . . . .</i>	118
6.6	<i>Plot of true <math>\beta</math> (<math>\tilde{\beta}/\gamma</math>) against predicted value of <math>\beta</math> (<math>\tilde{\beta}_{6.2}^*/\gamma_{6.2}^*</math>), separated by <math>p</math>, for Model (6.2). . . . .</i>	118
B.1	<i>Normal Q-Q plot for Sub-model 1. . . . .</i>	128
B.2	<i>Normal Q-Q plot for Sub-model 2. . . . .</i>	129
B.3	<i>Normal Q-Q plot for Sub-model 3. . . . .</i>	129
B.4	<i>Residuals vs Fitted values for Sub-model 1. . . . .</i>	129
B.5	<i>Residuals vs Fitted values for Sub-model 2. . . . .</i>	130
B.6	<i>Residuals vs Fitted values for Sub-model 3. . . . .</i>	130
B.7	<i>Residuals vs Predictors for Sub-model 1. . . . .</i>	131
B.8	<i>Residuals vs Predictors for Sub-model 2. . . . .</i>	132
B.9	<i>Residuals vs Predictors for Sub-model 3. . . . .</i>	133

# List of Algorithms

1	SIR simulation . . . . .	8
2	Maximum likelihood estimate . . . . .	10
3	Simulating a Gilbert-Erdős-Rényi network . . . . .	17
4	Approximate Bayesian Computation (ABC) algorithm [24] . . . . .	21
5	Calculating the auxiliary estimate for $\beta, betaAux$ . . . . .	23
6	Simulating a dataset for $betaAux$ approach . . . . .	25
7	$k$ -fold Cross Validation . . . . .	29
8	Simulating $(p, \beta)$ restrictions . . . . .	43
9	Multivariate Adaptive Regression Splines . . . . .	55
10	Simulating a Keeling network . . . . .	68
11	Simulating the Keeling network dataset . . . . .	71
12	Principal Component Analysis [23, page 23] . . . . .	79
13	Estimating the network parameter $f$ . . . . .	96
14	Simulating a training and test set for time-series data on Gilbert-Erdős-Rényi networks . . . . .	111

# Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

SIGNED: ..... DATE: .....

# Acknowledgements

I would like to begin by thanking my supervisors, Dr. Jonathan Tuke and Dr. Joshua Ross, and not just for their help over the past two years. Prior to commencing a Masters of Philosophy, Dr. Tuke and Dr. Ross had given me a wealth of support and guidance in my academic studies. Without fostering my interest in epidemiology I might never have undertaken a Masters of Philosophy, and so for that I am grateful.

I would also like to acknowledge Associate Professor Garique Glonek for his assistance in helping my understanding of multivariate adaptive regression splines, and how I could use them in my thesis.

Finally I would like to acknowledge my family for always supporting me when I needed it: my father and mother, Brian and Terri, for always believing I can do anything I set my mind to; my brother Luke, for always taking a keen interest in what I was doing; and my sister Amy for always trying her hardest to understand what in the world I was talking about. Special mention goes to my partner, Tessa Longstaff, for putting up with me in times of stress, especially over the last two months before submitting my thesis.

# Abstract

One of the motivating questions for many epidemiologists is “how quickly or widely will a particular infection spread?” To answer this question, often epidemic models are used to model the spread of a disease, with different epidemic models making different assumptions about the development of the disease; for example, two similar epidemic models might differ in whether they assume that people develop immunity after recovering from the disease. The advantage of these epidemic models is that they can be used to quickly estimate the epidemic model parameters, given an observed outbreak of a disease.

One assumption of most standard epidemic models is that an infectious individual has an equal probability of spreading the disease to any susceptible person in a population. The classical Susceptible-Infective-Recovered (SIR) model is an epidemic model with this assumption, which says that any pair of individuals has an equal probability of having an interaction [1]; when this interaction is between a susceptible and infectious individual, we call this interaction *adequate contact* [1] if the interaction results in a susceptible individual contracting the disease.

This assumption of equal probability of interaction (called homogeneous-mixing) can be a restrictive and unreasonable assumption in situations where within a population there are some pairs of individuals that never interact or some pairs of individuals that interact with higher probability. A more general model considers networks, in which nodes represent people and edges represent a possible path of infection; that is, if Node A (infectious) and Node B (susceptible) don’t share an edge then Node A cannot directly infect Node B. However without this homogeneous-mixing assumption, inference for the epidemic model parameters can be computationally intensive.

This thesis will answer two questions:

1. given observed properties of the network and the final epidemic size(s), can we efficiently estimate the epidemic model parameters; and,

2. given observed properties of the network and known times and types of events, can we efficiently estimate the epidemic model parameters?

We will answer these questions by assuming the data came from the homogeneous-mixing SIR model, and then estimating the epidemic model parameters. We then use a linear model to adjust these estimates, which provides a fast way to estimate epidemic model parameters. We will also show that the error of the models we present are, on average, never greater than 0.2; that is, the estimated epidemic model parameters are on average within 0.2 of the true epidemic model parameters.