Summer 7-12-2017

Comparison of two methods in estimating standard error of simulated moments estimators for generalized linear mixed models

Danielle K. Duran
University of New Mexico

Danielle K. Duran

Candidate

Statistics

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

Dr. Yan Lu , Chairperson

Dr. Guoyi Zhang

Dr. James Degnan

Comparison of two methods in estimating standard error of simulated moments estimators for generalized linear mixed models

by

Danielle K Duran

B.S., Sociology/Statistics, Bradley University, 2009

M.S., Sociology, Illinois State University, 2011

THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Statistics

The University of New Mexico

Albuquerque, New Mexico

July, 2017

Dedication

This thesis is the product of hard work, dedication, long nights, and endless support from my parents, husband, siblings, and friends. Here are some quotes that got me through this time:

"Lead, follow, or get the hell out of the way" – Dad

"It is what it is" – Mom

"Si te puedes corazon!" – Cesar

"Oh, graduating? That's cool" – Andrew

"School is important, dontcha know?" – All of my Minnesota relatives

"When are you graduating[1] *??" –Everyone*

"Why didn't I get a quote?" –Kaitlyn

[1]Now.

Acknowledgments

I would like to thank my advisor, Dr. Yan Lu, for her support and infallible advice and guidance. I would also like to thank my cat, Naki, for being there all those time I needed to nap while writing this. I have several other people I would like to thank, as well.[2]

[2]You know who you are.

Comparison of two methods in estimating standard error of simulated moments estimators for generalized linear mixed models

by

Danielle K Duran

B.S., Sociology/Statistics, Bradley University, 2009

M.S., Sociology, Illinois State University, 2011

M.S., Statistics, University of New Mexico, 2017

Abstract

We consider standard error of the method of simulated moment (MSM) estimator for generalized linear mixed models (GLMM). Parametric bootstrap (PB) has been used to estimate the covariance matrix, in which we use the estimates to generate the simulated moments. To avoid the bias introduced by estimating the parameters and to deal with the correlated observations, (Lu, 2012) proposed a multi-stage block nonparametric bootstrap to estimate the standard errors. In this research, we compare PB and nonparametric bootstrap methods (NPB) in estimating the stan-

dard errors of MSM estimators for GLMM. Simulation results show that when the group size is large, NPB and PB perform similarly; when group size is medium, NPB performs better than PB in estimating the mean. A data application is considered to illustrate the methods discussed in this paper, using productivity of plantation roses. The data application finds that, the person caring for the roses is associated with the productivity of those beds. Furthermore, we did an initial study in applying random forests to predict the productivity of the rose beds.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

General linear mixed models (GLMMs) are extensions of the generalized linear model, introduced by (McCullagh & Nelder, 1989). GLMMs integrate random effects into the fixed portion of the model. However, correlated observations within GLMMs present a computational difficulty in solving the maximum likelihood estimator, due to the high-dimensional integrals in the likelihood function. Therefore, approximation methods are used to solve the parameters of interest, for example, approximating the data (Penalized quasi-likelihood (PQL)), approximating the integral, and approximating the moments (MSM) etc.

Penalized quasi-likelihood (PQL) approach (Breslow & Clayton, 1993) is a commonly used estimation procedure for GLMM. However, it has been noticed that PQL tends to underestimate variance components as well as regression coefficients. Lin

and Breslow (1996) proposed bias corrected PQL approach for solving the likelihood estimators. However, Jiang (1998) showed that the estimators from the above two methods are both inconsistent. Method of simulated moments (MSM) is an extention of method of moments that could be used to estimate the parameters for GLMM, which approximates the moments by Monte Carlo simulation when direct computation are not possible. Jiang (2009) showed that MSM estimator of GLMM is consistent; the precision and efficiency of the MSM estimator are competitive to PQL type estimator while the computation is relatively simple.

The bootstrap method was introduced by Efron (1979). Since then, there have been enormous applications and adaptations of bootstrap method for inference problems under various data generating mechanisms. For example, Krishnamoorthy, Lu, and Mathew (n.d.) introduced PB method for testing equality of factor means in one-way ANOVA. Zhang (2015(a), 2015(b)) investigated the multiple comparison problems in one way and two way ANOVA with unequal variances and unequal group sizes. Jiang (2009) suggested a parametric bootstrap for estimating the covariance matrix for MSM estimators for GLMM. Kunsch (1989) proposed block bootstrap for analyzing the time series data sets. Liu and Singh (1992) independently suggested "Moving Block Bootstrap" (MBB). Lu (2012) proposed a two-stage block nonparametric bootstrap to estimate the standard errors of MSM estimators for GLMM. Hall (1992), Davison and Hinkley (1997), Shao and Tu (1995) and Lahiri (2003)

have thoroughly discussed different aspects of the bootstrap method.

The interest of this study is to compare the performance of two methods (Jiang, 2009 and Lu, 2012) in estimating standard error of MSM estimators for GLMM with an application to productivity of plantation roses via GLMM and random forests. This thesis is organized around 4 sections: in Section 2, a background on the theory behind general linear models, generalized linear models, linear mixed models, generalized linear mixed models, random forests, and standard error estimation of GLMMs; Section 3 is dedicated opt the simulation comparison between PB and NPB bootstrap estimation. We first review the procedure for both methods, then follow with the results of the simulation. The results depend on the group size; when group size is medium, NPB performs better than PB when estimating the mean; on the other hand, PB performs better than NPB in estimating the variance (when group size is medium). When group size is large, both PB and NPB perform similarly in estimating both the mean and standard deviation. Next, we introduce the rose data and demonstrate selected methods PB and NPB in a data application, followed by an application of the random forests in predicting the productivity of the rose data. The last section is a conclusion and considerations for future research.

Chapter 2

Background

2.1 General Linear Model

A general linear model refers to the potential linear dependency of the outcome/response on more than one explanatory variable, compared to the simple linear model. A general linear model has the form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{i(p-1)} + \epsilon_i \tag{2.1}$$

where the response $Y_i, (i = 1, 2, \cdots, n)$ is modeled by a linear function of the explanatory variables $X_j; j = 1, \cdots, p-1$ plus an error term. The systematic portion of the model may be written as

$$E(Y_i) = \mu_i = \beta_0 + \sum_{j=1}^{p-1} X_{ij}\beta_j; \quad i = 1, \cdots, n \tag{2.2}$$

where X_{ij} is the value of the jth covariate for the observation i. We can also write this relationship in matrix form:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ is called a design matrix with the i^{th} row $\mathbf{X}_i' = (1, X_{i1}, \cdots, X_{i(p-1)})$, i.e.,

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1(p-1)} \\ 1 & X_{21} & \cdots & X_{2(p-1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & X_{n(p-1)} \end{pmatrix},$$

$\boldsymbol{\beta}$ is an $p \times 1$ vector of population parameters:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix},$$

$\boldsymbol{Y}$ is an $n \times 1$ vector of responses:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix},$$

and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of error terms:

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

The general linear model assumes that the errors ϵ_i are independent and identically distributed, such that $E[\epsilon_i] = 0$ and $var[\epsilon_i] = \sigma^2$, where the variance σ^2 is assumed to be constant. We also assume that the errors follow a normal distribution: $\epsilon_i \sim N(0, \sigma^2)$ as a basis for inferential tests.

Take for example modeling monthly productivity at a rose plantation. The continuous outcome productivity, as measured by the number of exportable rose stems cut from each flower bed, could be modeled as a linear function of relative humidity, temperature, etc.. The general linear model may look like this:

$$Productivity = \beta_0 + \beta_1(RelativeHumidity) + \beta_2(Temperature) + \epsilon$$

General linear models are very useful for a variety of situations, but do come with some restrictions. General linear models are not appropriate when the range of Y is restricted, as with binary or count variables.

2.2 Generalized Linear Models

In this section, we will first review Generalized Linear Models. Next, we will review the maximum likelihood estimators for GLM.

Review of Generalized Linear Models

Generalized linear models extend the general linear model framework to address the issue of restricted range of Y. For example, the outcome Y_i is binary with 0 or 1. $X_1, X_2, \cdots, X_{p-1}$ are explanatory variables defined as before. $E(Y_i) = \mu_i$ for $i = 1, 2, \cdots, n$. For the n independent observations, the distribution of each Y_i is is an exponential family with density

$$f(Y_i, \theta_i, \phi) = \exp\left\{\frac{Y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(Y_i, \phi)\right\},$$

where θ_i is usually called a natural or canonical parameter and ϕ is a scale parameter (known or seen as a nuisance) and $a_i(\phi)$, $b(\theta_i)$, and $c(Y_i, \phi)$ are known functions.

It can then be shown that Y_i has mean and variance as follows

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{2.3}$$

$$var(Y_i) = \sigma_i^2 = b''(\theta_i)a_i(\phi)$$

where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$. When $a_i(\phi) = \frac{\phi}{p_i}$,

where p_i is a known prior weight, usually 1. The variance has the simpler form

$$var(Y_i) = \sigma_i^2 = \phi b''(\theta_i)/p_i.$$

The variance of Y_i is thus the product of two functions; the first, $b''(\theta_i)$ is dependent on the mean via the canonical parameter and is known as the variance function. The other function depends only on ϕ, and is independent of θ_i.

Generalized linear models are characterized by three parts: the random component, the systematic component, and a link function. The generalization portion of the linear model is the *link function*. Instead of modeling the mean, as in the general linear model, we introduce a one-to-one continuous differentiable transformation $g(\mu_i)$ such that

$$\eta_i = g(\mu_i). \tag{2.4}$$

Examples of the link function include log, reciprocal, logit, and probit. The logistic link function is given as:

$$g(x) = ln\frac{x}{1-x}.$$

We further assume that the transformed mean follows a linear model, so that

$$\eta_i = \mathbf{X}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)}.$$

η_i is called the linear predictor. Given the simplicity of the model of η_i, we can invert

it to obtain

$$\mu_i = g^{-1}(\mathbf{X}_i'\boldsymbol{\beta})$$

The response Y_i is not transformed, but rather the expected value of μ_i.

When the link function equates the linear predictor η_i and the canonical parameter θ_i, we say that the link is canonical. Specifically, we have,

$$g(\mu) = \eta = \theta. \tag{2.5}$$

It can be shown from (2.3) that

$$\theta = (b')^{-1}(\mu). \tag{2.6}$$

By the results of (2.5) and (2.6), we have $g = (b')^{-1}$. This link is called canonical link.

For the normal distribution, the canonical link is the identity link. Each distribution has its own canonical link, although different pairings between distributions and links are possible. A canonical link is advantageous in that it allows for the existence of minimally sufficient statistic $\boldsymbol{\beta}$.

Take for example a binary logistic regression model of the rose data. Suppose we want to model low or high productivity, where high productivity represents the top 15 percent of productivity values. The explanatory variables remain the same as in

the general linear model. The model is:

$$ln\left(\frac{P(HighProductivity)}{1-P(HighProductivity)}\right) = \beta_0 + \beta_1(RelativeHumidity) + \beta_2(Temperature),$$

which models the log odds of probability of "success" (high productivity) as a function of the explanatory variables. The systematic component of the model is composed of the explanatory variables, which can be continuous or discrete; the random component refers to the Binomial(n, p) distribution of Y; and the link function is the logistic link described above.

Maximum Likelihood Estimation

The defining feature of Generalized Linear Models is that data are all fit with the same algorithm, with a form of iteratively re-weighted least squares. Given a trial estimate of the parameters $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}^{(0)}$, we calculate the estimated linear predictor $\mathbf{X}'\hat{\boldsymbol{\beta}}^{(0)}$ and use this to obtain fitted values $\hat{\mu}_i = g^{-1}(\mathbf{X}'\hat{\boldsymbol{\beta}}^{(0)})$. Using this estimation, the working dependent variable can be calculated

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{d\eta_i}{d\mu_i}, \tag{2.7}$$

where the term $\frac{d\eta_i}{d\mu_i}$ is the derivative of the link function. Now we can calculate the iterative weights

$$w_i = p_i/[b''(\theta_i)(\frac{d\eta_i}{d\mu_i})^2], \tag{2.8}$$

where $b''(\theta_i)$ is the second derivative of $b(\theta_i)$ evaluated at the trial estimate, when $a_i(\phi)$ is assumed to have the usual form $\frac{\phi}{p_i}$. The weight is inversely proportional to the variance of z_i, the dependent variable, given the current estimates of the parameters, with proportionality factor ϕ. Using the dependent variable z_i, the weights w_i, and the predictors $\mathbf{X}_i$, we calculate the weighted least-squares estimate

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z},$$

where $\mathbf{X}$ is the model matrix, $\mathbf{W}$ is a diagonal matrix of weights with entries w_i given by (2.8) and $\mathbf{z}$ is a response vector with entries z_i given by (2.7). The procedure is iterative, and repeated until estimates vary by less than a pre-specified amount. This technique leads to maximum likelihood (ML) estimates (McCullagh and Nelder 1989). Nelder and Wederburn (1972) were the first to extend the fitting of generalized linear models to deal with maximum-likelihood estimation for exponential-family models.

2.3 Linear Mixed Effects Models

Mixed model methodology extends the general linear model to data that have a complex, multilevel, or hierarchal structure. Observations between levels or clusters are independent of one another, but observations within clusters or levels are correlated. A common example of this structure is assessments nested within individuals, or

classrooms within schools. The strength of the mixed model is the ability to model these complex data by the inclusion of multilevel random effects. Since this section, we slightly changed some of the notations, such as, $\mathbf{Y}$ to $\mathbf{y}$, $\mathbf{X}_i$ to $\mathbf{x}_i$ to make the notation be consistent with those in Jiang (2009). The notation in this section will remain the same until the end of the thesis.

The linear mixed effects (LME) model can be expressed generally as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{2.9}$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ is an $n \times 1$ vector of observations, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters defined as before; $\boldsymbol{\alpha}$ is an $m \times 1$ vector of random effects with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_m)'$. $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors; $\mathbf{X}$ is an $n \times p$ design matrix; $\mathbf{Z}$ is an $n \times m$ design matrix defined as follows:

$$\mathbf{Z} = \begin{pmatrix} 1 & z_{11} & \cdots & z_{1m} \\ 1 & z_{21} & \cdots & z_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_{n1} & \cdots & z_{n(m)} \end{pmatrix},$$

or in some textbooks (such as Demidenko (2005)),

$$\mathbf{Z} = \begin{pmatrix} \boldsymbol{Z_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Z_q} \end{pmatrix},$$

where $\mathbf{Z}_i$ is the $n_i \times m_i$ design matrix, with $\sum_{i=1}^{q} n_i = n$ and $\sum_{i=1}^{q} m_i = m$.

Assume that $\boldsymbol{\alpha}$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{G}$, and $\boldsymbol{\varepsilon}$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}$. It is also assumed that $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ are uncorrelated. For a special case $\mathbf{R} = \tau^2\mathbf{I}$, given $\boldsymbol{\alpha}$, the observations $y_1, y_2, ..., y_n$ are conditionally independent such that

$$y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\alpha}, \tau^2),$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are the ith row of the design matrices $\mathbf{X}$ and $\mathbf{Z}$ respectively. Through this article, the responses y_is are correlated in nonoverlapping blocks (or strata, clusters or groups) related to certain random effect α_r. We assume that the population blocks are all essentially infinite. The distribution of y_i does not depend on n_i, the size of the sample taken from the ith block. The variance-covariance matrix of the random effects $\mathbf{G}$ and $\mathbf{R}$ can be estimated either by maximum likelihood, or quadratic non-iterative distribution-free estimators, including MINQUE, variance least squares, and method of moments (Demidenko, 2005).

Consider the rose data as an example, again with a continuous outcome. Here, we augment the general linear model with a random effect on cutters, because we want to see if there are variabilities among the cutters regarding the productivity. The model looks like:

$$Productivity = \beta_0 + \beta_1(RelativeHumidity) + \beta_2(Temperature) + (Cutter) + \epsilon$$

This model accounts for the random effect on cutter, to control for any heterogeneity observed between cutters, while retaining relative humidity and temperature as fixed effects of primary interest.

2.4 Generalized Linear Mixed Models

Generalized linear mixed models (GLMMs) are extensions of the generalized linear mixed model family (McCullagh & Nelder, 1989), whereby random effects of the predictors are incorporated to account for the restricted range responses. This approach is especially useful for multilevel data, or models in which random cluster and/or subject effects need to be taken into account. Mixed models for the continuous normal outcomes have been developed extensively. Development have been made with nonnormal data as well, and both normal and non-normal data fall under the rubric of GLMMs.

GLMM is an extension of the general linear mixed models, in which the responses are correlated and categorical. Given a vector of random effects $\boldsymbol{\alpha}$, the responses $y_1, y_2, ..., y_n$ are conditionally independent such that the conditional distribution of y_i given $\boldsymbol{\alpha}$ is a member of the exponential family with the following probability density function

$$f(y_i|\boldsymbol{\alpha}) = \exp\left(\frac{y_i * \xi_i - b(\xi_i)}{a_i(\phi)} + c_i(y_i, \phi)\right) \tag{2.10}$$

where ϕ is a dispersion parameter and $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions. ξ_i is associated with the conditional mean $\mu_i = E(y_i|\boldsymbol{\alpha})$, which is associated with a linear predictor $\eta_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\alpha}$ through a known link function $g(\cdot)$ with $g(\mu_i) = \eta_i$. According to the properties of the exponential family, one has $b'(\xi_i) = \mu_i$. Under the canonical link, one has $\xi_i = \eta_i$, that is, $g = h^{-1}$ where $h(\cdot) = b'(\cdot)$, h^{-1} represents the inverse function of h.

Let's turn to an example. Recall the previous example of the rose productivity data. Suppose we want to model low or high productivity, where high productivity represents the top 15 percent of productivity values. Let i denote the ith cutter and j denote the flower beds gathered in by the ith cutter. Assume $i = 1, \cdots, m$ cutters and $j = 1, \cdots, n_i$ repeated observations (flower beds) nested within each cutter. The linear predictor η_{ij} can be written as

$$\eta_{ij} = \beta_0 + \beta_1(RelativeHumidity) + \beta_2(Temperature) + (Cutter),$$

therefore, the GLMM is:

$$\begin{aligned} & ln\left(\frac{P(HighProductivity)}{1 - P(HighProductivity)}\right) \\ = \ & \eta_{ij} \\ = \ & \beta_0 + \beta_1(RelativeHumidity) + \beta_2(Temperature) + (Cutter) \end{aligned}$$

where cutter is the random effect used to control for any heterogeneity observed be-

tween cutters, while retaining relative humidity and temperature as fixed effects of primary interest. This models the log odds of probability of "success" (high productivity) as a function of the explanatory variables. The outcome has Binomial(n, p) distribution, and the link function is the logistic link.

2.5 Random Forest

Random forests as an idea are a subset of machine learning, which allow for automated decision making. Preceding random forests were procedure such as bagging (Breiman, 1996) and random split selection (Dietterich, 1998). The term random forest refers to an ensemble (or forest) of decision trees, grown from a variant of the nodes. The "randomness" in random forests is introduced at each node when determining the split. These trees are non-parametric, meaning that they can "model arbitrarily complex relations between inputs and outputs, without any a priori assumption" (p.26) Louppe (2014). They are able to handle ordered data or categorical data, or a mix of both; are robust to outliers and errors; minimize noise in variables to exclude irrelevant data; and are easily interpretable. The theory of random forest comes from use of the strong law of large numbers, which shows that the random forests always converge and therefore overfitting is not a concern Breiman (2001). The accuracy of a random forest depends on the strength of the individual tree

classifiers, and their interdependence Amit and Geman (1997).

Take the rose data as an example: we are interested in whether high or low productivity is influenced by a variety of factors; the random forest takes in our independent variables (cutter, temperature, pests, etc) and uses each of those in estimating a node in the individual trees. We can see the overall accuracy of the forest and the importance of each variable in predicting productivity.

2.6 Method of Simulated Moments for GLMM

Method of moments (MoM) is another way used to estimate the parameters for GLMMs. The MoM begins with obtaining a set of estimating equations by equating sample moments of the sufficient statistics to their expectations. These expectations usually involve integrals, the highest dimension of which equals the number of sources of random effects Jiang (1998). The evaluation of the integrals may not be possible, therefore, a *method of simulated moments* (MSM) may be considered as an approximation. MSM was introduced by McFadden (1989) and applies to situation in which the theoretical moment function cannot be expressed. MSM approximates the moments by Monte Carlo simulation when direct computation of the moments are not possible. Jiang (2009) showed that MSM estimator of GLMM is consistent; furthermore, the MSM estimators are computationally simpler and comparable in

efficiency and precision to the PQL type estimators.

The following is a description of the general results of MSM estimator for GLMM . Much of this notation is from Jiang (2009). Assume that the conditional density of y_i given the vector of random effects $\boldsymbol{\alpha}$ has the following form,

$$f(y_i|\boldsymbol{\alpha}) = exp[(w_i/\phi)\{y_i\xi_i - b(\xi_i)\} + c_i(y_i, \phi)],$$

where ϕ is a dispersion parameter, and w_i's are known weights with

$$w_i(x) = \begin{cases} 1, & \text{for ungrouped data} \\ n_i, & \text{for grouped data if the response is an average} \\ 1/n_i, & \text{response is a group sum.} \end{cases}$$

$b(\cdot)$ and $c_i(\cdot, \cdot)$ are known functions. For ξ_i, we assume a canonical link $\eta_i = \xi_i$. Let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \cdots, \boldsymbol{\alpha}_q')'$, where $\boldsymbol{\alpha}_r$ is a $m_r \times 1$ random vector (with $m_1 + m_2 + \cdots + m_q = m$) whose components are independently distributed as $N(0, \sigma_r^2)$, $1 \leq r \leq q$. For convenience, let

$$\boldsymbol{\alpha} = \mathbf{Du} \tag{2.11}$$

where $\mathbf{D}$ is blockdiagonal with the diagonal blocks $\sigma_r \mathbf{I}_{m_r}$, $1 \leq r \leq q$, and $\mathbf{u} \sim N(\mathbf{0}, \mathbf{I}_m)$.

Suppose the linear predictor associated with the link function is $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}$. $\mathbf{Z}$ is an $n \times m$ design matrix. Let $\mathbf{Z}_r$ be an $n \times m_r$ design matrix of random effects $\boldsymbol{\alpha}_r$, so that $\mathbf{Z} = (\mathbf{Z}_1, \cdots \mathbf{Z}_q)$. For simplicity, we assume that $\mathbf{Z_r}, 1 \leq r \leq q$ are standard

design matrices that each $\mathbf{Z_r}$ consists only of 0s and 1s, and there is exactly one 1 in each row and at least one 1 in each column. We denote the ith row of $\mathbf{Z_r}$ by $\mathbf{z}'_{ir} = (\mathbf{z}_{irl})'$ with $1 \leq i \leq n$ and $1 \leq l \leq m_r$. We have $|\mathbf{z}_{ir}|^2 = 1$ and for $s \neq t$, $\mathbf{z}'_{sr}\mathbf{z}_{tr} = 0 \quad \text{or} \quad 1$.

Let $\mathbf{I}_r = \{(s,t) : 1 \leq s \neq t \leq n, \mathbf{z}'_{sr}z_{tr} = 1\} = \{(s,t) : 1 \leq s \neq t \leq n, \mathbf{z}_{sr} = \mathbf{z}_{tr}\}$. Let $\mathbf{X}_j$ be the jth column of design matrix $\mathbf{X}$. $\mathbf{W} = \text{diag}(w_i, 1 \leq i \leq n)$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_1, \cdots, \sigma_q)$, $\mathbf{u} = (\mathbf{u}'_1, \cdots, \mathbf{u}'_q)'$ with $\mathbf{u}_r = (\mathbf{u}_{rl})_{1 \leq l \leq m_r}$, $\mathbf{u}_r \sim N(\mathbf{0}, \mathbf{I}_{m_r})$. $e(\boldsymbol{\theta}, \mathbf{u}) = \{b'(\xi_i)\}_{1 \leq i \leq n}$ with $\xi_i = \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{r=1}^{q} \sigma_r \mathbf{z}'_{ir}\mathbf{u}_r$. Thus, the MM equations that do not involve ϕ are given by

$$\sum_{i=1}^{n} w_i x_{ij} y_i = \mathbf{X}'_j \mathbf{W} \mathbf{E}\{e(\boldsymbol{\theta}, \mathbf{u})\}, 1 \leq j \leq p, \tag{2.12}$$

$$\sum_{(s,t) \in I_r} w_s w_t y_s y_t = \mathbf{E}\{e(\boldsymbol{\theta}, \mathbf{u})' \mathbf{W} \mathbf{H}_r \mathbf{W} e(\boldsymbol{\theta}, \mathbf{u})\}, 1 \leq r \leq q, \tag{2.13}$$

where the expectations on the right-hand sides are with respect to $\mathbf{u} \sim N(0, \mathbf{I}_m)$. We approximate the right-hand sides by a simple Monte Carlo simulation. Let $u^{(1)}, \cdots, u^{(L)}$ be generated i.i.d. copies of $\mathbf{u}$, the right side of (2.12) and (2.13) can be approximated by Monte Carlo averages $\mathbf{X}'_j \mathbf{W} [\frac{1}{L} \sum_{l=1}^{L} e\{\boldsymbol{\theta}, \mathbf{u}^{(l)}\}], 1 \leq j \leq p$ and $\frac{1}{L} \sum_{l=1}^{L} e\{\boldsymbol{\theta}, \mathbf{u}^{(l)}\}' \mathbf{W} \mathbf{H}_r \mathbf{W} e\{\boldsymbol{\theta}, \mathbf{u}^{(l)}\}, 1 \leq r \leq q$.

2.7 Standard Error Estimation of GLMM

The primary difficulty in implementing full likelihood inference lies in the integrations needed to evaluate the quasi-likelihood ql; hence, we can turn to penalized quasi-likelihood estimation (PQL) (Breslow & Clayton, 1993). PQL estimation (Breslow & Clayton, 1993) and bias corrected PQL estimation (Lin & Breslow, 1996) have been popular solutions for addressing this problem; however, these methods have been shown to yield inconsistent estimators (Jiang, 1998).

Parametric Bootstraping (PB) resample a known distribution function, whose parameters are estimated from the sample. A problem of parametric bootstrap is that the estimators are used to generate the simulated moments instead of the true parameters, in which bias was introduced by estimating the parameters. The basic idea of non-parametric bootstrapping (introduced by Efron (1979)) is to estimate population parameters via simulation; nonparametric bootstrapping involves sampling with replacement from the observed data, without making assumptions as to the sampling distribution. The random samples are the same size as the sample itself that it draws from. Samples are taken B times, resulting in a sampling vector which consists of $1 \times \mathbf{B}$ samples. The standard errors may then be estimated over all samples. There have been numerous applications and variations of bootstrap methodology for interference problems. These aspects have been discussed by Hall

(1992), Davison and Hinkley (1997), Shao and Tu (1995) and Lahiri (2003).

2.7.1 Variance Estimation of MSM Estimators Using PB

Standard errors can be estimated easily with PB; however, this estimated standard error is also easily influenced by the parameter estimates used when generating simulated moments. Let $\hat{\boldsymbol{\theta}}$ be the MSM estimator of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_1, \sigma_2, ...\sigma_q)$, and $M(\boldsymbol{\theta})$ be the vector of moments. Let $\hat{\boldsymbol{\vartheta}}$ be the MSM estimator of $\boldsymbol{\vartheta} = (\boldsymbol{\beta}', \sigma_1^2, \sigma_2^2, ...\sigma_q^2)$, where $\boldsymbol{\vartheta}$ is a function of $\boldsymbol{\theta}$. By Taylor theorem,

$$\hat{\mathrm{M}}(\hat{\boldsymbol{\theta}}) \approx \mathrm{M}(\boldsymbol{\theta}) + \dot{\mathrm{M}}(\boldsymbol{\theta})\mathrm{J}^{-1}(\boldsymbol{\theta})(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})$$

where $\dot{\mathrm{M}}$ is the matrix of first derivatives and $\mathrm{J}(\boldsymbol{\theta}) = \mathrm{diag}(1, ..., 1, 2\sigma_1, ..., 2\sigma_q)$. The covariance matrix of $\hat{\boldsymbol{\vartheta}}$ is derived as follows

$$\mathrm{Var}(\hat{\boldsymbol{\vartheta}}) \approx \mathrm{J}(\boldsymbol{\theta})\dot{\mathrm{M}}(\boldsymbol{\theta})^{-1}\mathrm{Var}(\hat{M}) \left(\dot{\mathrm{M}}(\boldsymbol{\theta})^{-1}\right)^T \mathrm{J}(\boldsymbol{\theta}).$$

The simulated moments can be used to estimate $\dot{\mathrm{M}}(\boldsymbol{\theta})$ and a parametric bootstrap may be used to estimate the covariance matrix of $\hat{M}$. Generate K bootstrap samples from the GLMM using $\hat{\boldsymbol{\theta}}$ and compute the sample moments for all bootstrap samples. Take $\mathrm{M}^k(\hat{\boldsymbol{\theta}}) = (M_1^k, M_2^k, ...M_s^k)$ for the kth bootstrap sample, and then

$$\widehat{Var}(\hat{M}) = \frac{1}{K-1}\sum_{k=1}^{K}(M^k - \bar{M}^*)(M^k - \bar{M}^*)^T$$

where $\bar{M}^* = \frac{1}{K}\sum_{k=1}^{K} M^k$.

2.7.2 Variance Estimation of MSM Estimators Using Block NonParametric Bootstrap (NPB)

For correlated observations, such as those observed in GLMM, single observation resampling fails to work. Block bootstrapping for time series data sets was first proposed by Kunsch (1989). Many block bootstrap techniques have been proposed since then: Liu and Singh (1992) (Carlstein, 1986) (Politis & Romano, 1992) (Carlstein, Do, Hall, Hesterberg, & Kunsch, 1998) (Paparoditis & Politis, 2001). Lu (2012) proposed a block bootstrap for use in MSM estimators for GLMM; a review of the procedure follows.

Consider a general GLMM,

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}, \tag{2.14}$$

and a simple blocked sample

$\mathbf{y}_1 : y_{11}, y_{12}, \cdots, y_{1n_1}$,

$\mathbf{y}_2 : y_{21}, y_{22}, \cdots, y_{2n_2}$,

$\vdots$

$\mathbf{y}_t : y_{t1}, y_{t2}, \cdots, y_{tn_t}$,

where t is number of groups with $\sum_{i=1}^{t} n_i = n, i = 1, 2, \cdots, t$. The following gives a procedure of nonparametric block bootstrap standard error estimation for MSM estimators:

Step 1: Sample t numbers from 1 to t with replacement, say $t_1, t_2, \cdots, t_t$. t_i may be equal to t_j since sampling is with replacement.

Step 2: Sample n_{t_1} observations from $\mathbf{y}_{t_1}$; n_{t_2} observations from $\mathbf{y}_{t_2}$ etc until n_{t_t} observations from $\mathbf{y}_{t_t}$ with replacement to form a block bootstrap sample.

Step 3: Calculate MSM estimators of μ and σ^2 for the bootstrap sample by equations (2.12) and (2.13). The right side of equations was approximated by a simple Monte Carlo simulation introduced in Section 2.6.

Step 4: Repeat step 1 and 3 L times. Calculate the sampling variance of the MSM estimates from the L bootstrap samples.

Chapter 3

Proposal: Comparison of PB and NPB Bootstrap Estimation

In this section, we perform simulation studies to compare the PB and nonparametric bootstrap in estimating the standard errors of MSM estimators for GLMM. Simulation results are given in Tables (3.1), (3.2) and also in Figures (3.1), (3.2), (3.3) and (3.4). Logistic normal model (4.1) is used in simulation study:

$$\text{logit}(P(Y_{ij} = 1)) = \mu + \alpha_i, \tag{3.1}$$

where $1 \leq i \leq m$ (in this case, number of groups t is equal to the number of random effects m), $1 \leq j \leq n_i$ for each i, n_i is the number of observations within each group i,

and $\alpha_i's$ are i.i.d. normally distributed random variables with mean zero and variance σ^2. For simplicity, we assume that $n_i = k$, for $i = 1, 2, \cdots, m$. Simulation set up mainly follows from Jiang (2009):

(1) Set $\mu = .2$, $\sigma = 1$, $m = 30$ and $k = 6$. Generate a sample by (4.1);

(2) Find MSM estimates for μ and σ^2 for the bootstrap samples. The MM estimating equations by (2.12) and (2.13) for logistic normal model (4.1) are as follows,

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{y}_{i.} = E(\mathbf{y}_{i.}),$$

and

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{y}_{i.}^2 = E(\mathbf{y}_{i.}^2),$$

where $E(\mathbf{y}_{i.}) = kE(f(u))$, $E(\mathbf{y}_{i.}^2) = kE(f(u)) + k(k-1)E(f^2(u))$, $f(u) = \exp(\mu + \sigma u)/(1 + \exp(\mu + \sigma u))$. Recall that u is a standard normal random variable. Generate a sequence of standard normal random variables, $u_1, u_2, ..., u_L$. The replication number L for the simulated moments is set to 100. We approximate the expectations by

$$E(f(u)) = \frac{\sum_{i=1}^{L} f(u_i)}{L}$$

and

$$E(f^2(u)) = \frac{\sum_{i=1}^{L} f^2(u_i)}{L}.$$

The MSM estimators of μ and σ can be solved from the system MM estimating equations.

(3) Find parametric bootstrap variance estimators Parametric $\hat{\text{SE}}(\hat{\mu})$ and Parametric $\hat{\text{SE}}(\hat{\sigma}^2)$ (refer to Section 2.7.1);

(4) Find nonparametric block bootstrap variance estimates Nonparametric $\hat{\text{SE}}(\hat{\mu})$ and Nonparametric $\hat{\text{SE}}(\hat{\sigma}^2)$ (refer to Section 2.7.2), the replication number is 200 for this step;

(5) Repeat steps (1) to (4) for $n = 1000$ replications, record the average values of the MSM estimators ($\hat{\mu}$ and $\hat{\sigma}^2$), standard errors ($\hat{\text{SE}}(\hat{\mu})$ and $\hat{\text{SE}}(\hat{\sigma}^2)$), parametric bootstrap standard errors and nonparametric block bootstrap standard errors of MSM estimators;

(6) Repeat steps (1) to (5) for different settings with $m = 30$, $k = 20$; $m = 80$, $k = 6$ and $m = 80$, $k = 20$. Table 3.1 gives the results. Figures (3.1) and (3.2) are representation of Table 3.1.

(7) Repeat steps (1) to (6) for $\mu = 1.0$, $\sigma = 1.0$. Table 3.2 gives the results. Figures (3.3) and (3.4) are representation of Table 3.2.

In both tables, $\hat{\mu}$ and $\hat{\sigma}^2$ are the average values of the MSM estimates from the 1000 replications. They are considered as the true mean and variance. $\text{SE}(\hat{\mu})$ and $\text{SE}(\hat{\sigma}^2)$ are the average value of standard errors. They are considered as the true standard error estimates of $\hat{\mu}$ and $\hat{\sigma}^2$. Parametric $\hat{\text{SE}}(\hat{\mu})$ and Parametric $\hat{\text{SE}}(\hat{\sigma}^2)$ are parametric bootstrap variance estimates. And Nonparametric $\hat{\text{SE}}(\hat{\mu})$ and Nonparametric $\hat{\text{SE}}(\hat{\sigma}^2)$ are nonparametric block bootstrap variance estimates. Numbers in paren-

Table 3.1: Simulation results: $\mu = .2$, $\sigma = 1$ in model (4.1)

m	k	$\hat{\mu}$	SE($\hat{\mu}$)	Parametric $\hat{SE}(\hat{\mu})$	Nonparametric$\hat{SE}(\hat{\mu})$
30	6	.2022	.2545	$.4929_{(5.2014)}$	$.2720_{(.0535)}$
30	20	.1927	.2096	$.2032_{(.0338)}$	$.2260_{(.0356)}$
80	6	.2052	.1539	$.1525_{(.0170)}$	$.1608_{(.0195)}$
80	20	.2066	.1311	$.1252_{(.0129)}$	$.1334_{(.0144)}$
		$\hat{\sigma}^2$	SE($\hat{\sigma}^2$)	Parametric$\hat{SE}(\hat{\sigma}^2)$	Nonparametric$\hat{SE}(\hat{\sigma}^2)$
30	6	1.0563	.6423	$.6098_{(.2675)}$	$.7736_{(.3720)}$
30	20	1.0021	.3999	$.3633_{(.1338)}$	$.4689_{(.1668)}$
80	6	1.0271	.3560	$.3617_{(.0906)}$	$.3990_{(.1047)}$
80	20	.9969	.2382	$.2205_{(.0493)}$	$.2465_{(.0520)}$

thesis are the sample variances.

From Tables (3.1), (3.2) and Figures (3.1), (3.2), (3.3) and (3.4), we see that both methods are acceptable in estimating μ and σ^2. When the group size is large, nonparametric bootstrap and PB perform similarly. We also notice that when group size is medium, nonparametric bootstrap performs better than PB in estimating the mean; while PB performs better than nonparametric bootstrap in estimating the variance.

Figure 3.1: Simulation results: $\mu = .2$, $\sigma = 1$ in model (4.1)

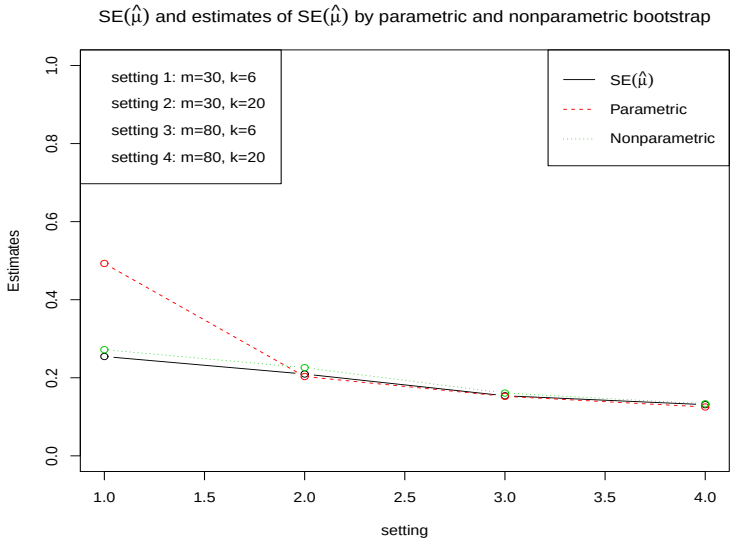


Figure 3.2: Simulation results: $\mu = .2$, $\sigma = 1$ in model (4.1)

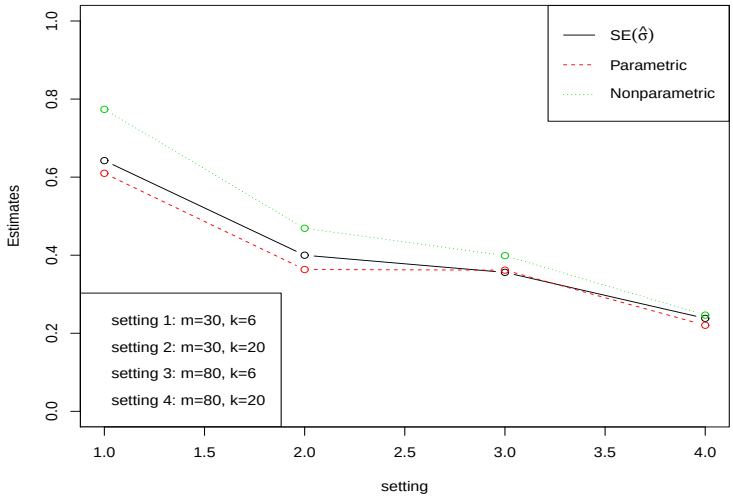


Table 3.2: Simulation results: $\mu = 1.0$, $\sigma = 1.0$ in model (4.1)

m	k	$\hat{\mu}$	SE($\hat{\mu}$)	Parametric $\hat{SE}(\hat{\mu})$	Nonparametric$\hat{SE}(\hat{\mu})$
30	6	1.0088	.2889	$.3610_{(2.0846)}$	$.2983_{(.0651)}$
30	20	.9804	.2205	$.2078_{(.0358)}$	$.2359_{(.0391)}$
80	6	1.0064	.1777	$.1649_{(.0206)}$	$.1734_{(.0236)}$
80	20	.9979	.1357	$.1295_{(.01343)}$	$.1389_{(.0148)}$
		$\hat{\sigma}^2$	SE($\hat{\sigma}^2$)	Parametric$\hat{SE}(\hat{\sigma}^2)$	Nonparametric$\hat{SE}(\hat{\sigma}^2)$
30	6	1.033	.6921	$.6505_{(.3116)}$	$.8211_{(.4504)}$
30	20	.9756	.4050	$.3778_{(.1419)}$	$.4816_{(.1890)}$
80	6	1.0088	.3964	$.3870_{(.1078)}$	$.4237_{(.1269)}$
80	20	.9968	.2418	$.2352_{(.0533)}$	$.2611_{(.0613)}$

Figure 3.3: Simulation results: $\mu = 1.0$, $\sigma = 1.0$ in model (4.1)

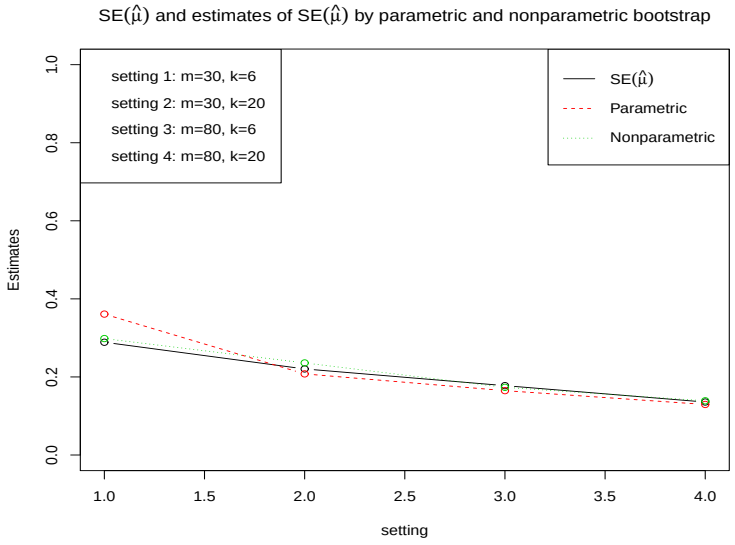
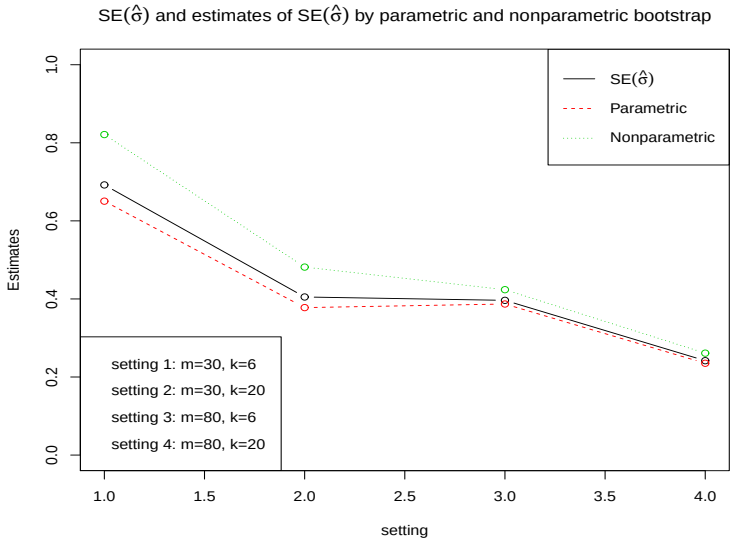


Figure 3.4: Simulation results: $\mu = 1.0$, $\sigma = 1.0$ in model (4.1)



Chapter 4

Data Application

Throughout this document we have referred to the rose data. Here, we will introduce more thoroughly the data and apply selected methods. The dataset in this paper come from a commercial rose plantation located outside of Biblian, Ecuador, in the province of Canar. The plantation began collecting data on flower productivity in 2009, and slowly added other variables of interest over the following years, including flower bed number and length, flower variety, greenhouse conditions (temperature, humidity, dew point), which worker was caring for and cutting the flowers, whether plastic was applied, and pests and infections that the rose plants sometimes acquire. By 2013, all of these variables were being collected for at least some of the greenhouses (but not all of the greenhouses, due to cost and time restrictions). Figure 4.1, below, shows an aerial view of the plantation.

Figure 4.1: Trebol Roses Plantation.

Multiple flower varieties are housed in any greenhouse; between one and fourteen flower varieties can be found in a single greenhouse. Similarly, multiple cutters work within any greenhouse; at least 5 cutters are required, with a maximum of 12 cutters present in one greenhouse. Cutters may work in more than one greenhouse; similarly, a popular flower variety may be grown in more than one greenhouse.

Productivity, the main variable of interest, was collected every month of the year except for October and November. The plants grow in 90-day cycles; due to this, it is important that the plants rest

Figure 4.2: A cutter gathering stems.

for these two months in preparation for the heavy production required by Valentine's Day in February. Temperature, relative humidity, and dew point were collected every 30 minutes. This data was aggregated to a monthly average. Different workers are assigned to sets of flower beds. Workers, or cutters as we refer to them here, are responsible for all aspects of care for the plants, including cutting the stems for exportation. All cutters are trained in the same manner. The workers were assigned to certain beds within greenhouses for one year.This could potentially change, but often workers remained with the same assigned beds the following year. Worker turnover is not a concern as working conditions and wages are good; therefore, many of the same workers have remained with the plantation throughout the data collection period. Plastic is removed or added to all beds within greenhouses on certain dates. This date is aggregated to the month level; for example, if plastic was changed on 1/8/2014, it would be coded as a change during January of 2014. The variables, productivity, flower variety, and worker were collected at the flower bed level. The variables temperature, humidity, dew point, and plastic were collected at the greenhouse level. The pest and infection variables were collected for flower varieties within greenhouse. Productivity refers to the total number of exportable stems gathered for each row of plants for a given month. After all of the potentially exportable stems have been cut from the plants, they are then assessed for quality and counted in post-production if they are chosen for exportation. Temperature is measured in degrees Celsius, with low alarms at 0.0 degrees and high alarms at 40.0 degrees. Rel-

ative humidity is the amount of water vapor present in air at a given temperature, expressed as a percentage. Low alarms sound at 35.0, while a relative humidity of 85.0 triggers high alarms. Dew point is an indication of the amount of water in the air, and is measured in degrees Celsius. Stem class was used to group flower varieties together; categories were: less than 50 cm, 50 to 53 cm, 53 to 55 cm, 55 to 58 cm, and 58 to 61 cm.

Figure 4.3: Inside one of the greenhouses.

A number of other variables are involved in producing roses in a plantation setting which are not accounted for: plant position within the greenhouse (those plants closer to the door experience more variation in greenhouse conditions) and how much water each bed received. No data are available for these and similar variables. Similarly, some variables are controlled for across greenhouses. For example, the plantation

makes its own fertilizer, which is applied to every flower bed. All beds are treated with any desired soil additives. All greenhouses are exposed to the same ambient light, wind, and weather conditions as they are all situated within about 30 hectares. Table 4.1 summarizes the variables used and how they were measured.

Table 4.1: Data Variable Description.

Variable	Description
Productivity	Collected per flower bed, a count of how many usable stems were gathered for export
Variety	The variety of rose
Temperature	The temperature inside of the greenhouse, measured in Celsius. Measured every 30 minutes and aggregated to month
Humidity	The amount of water in the air, measured in percent. Measured every 30 minutes and aggregated to month
Dew point	The atmospheric temperature below which water droplets begin to condense and form dew, measured in Celsius. Measured every 30 minutes and aggregated to month
Cutter	A number assigned to the person taking care of particular sets of flower beds
Plastic Change	An indicator variable which states whether plastic was either applied or removed from flower beds during a particular month
Spiders	A count of how many spiders were found on the plants within a given greenhouse for a given month
Aphids	A count of how many aphids were found on the plants within a given greenhouse for a given month
Botritis	A count of how many instances of botrytis were found on the plants within a given greenhouse for a given month
Mold	A count of how many instances of mold were found on the plants within a given greenhouse for a given month
Velloso	A count of how many instances of velloso were found on the plants within a given greenhouse for a given month

A total of 17 greenhouses are represented in this dataset, with between 43 and 249 flower beds in each, from years 2013 to 2016. Data on the other variables discussed varies by whether the data was collected at the time. Table 4.2 displays the descriptive statistics for the continuous variables in the full dataset.

Table 4.2: Descriptive Statistics for the Full Dataset.

Year	Mean	Std Dev	Min	Max
Productivity	349.05	295.08	1	2,988
Temperature	15.75	1.22	12.93	18.50
Humidity	75.37	5.89	9.43	82.85
Dew Point	10.94	5.40	7.25	78.03
Spiders	21.76	18.92	0	95.68
Aphids	0.68	1.76	0	27.20
Botritis	0.36	1.53	0	23.57
Mold	20.25	26.49	0	98.64
Velloso	1.12	4.19	0	56.83

We can see that there's great variability, especially among the productivity of the flowers; for example, the standard deviation is nearly as large as the mean, indicating that there are greenhouses or beds that sometimes produce inordinately high amounts of flower stems. Temperature and dew point are measured in degrees Celsius, and are both within the normal range for greenhouses. Humidity is measured as a percent of the water in the air, and also tends to be within normal range. The pest variables, spiders, aphids, botrytis, mold, and velloso, are all interpreted as rates. If 10 spiders were found during the month of January in Greenhouse 31 (which has 200 beds), for example, the rate would be recorded as $10/200 = 0.05$ spiders (or aphids, mold,

etc.). The pest data aren't necessarily interpreted as within a range;" the plantation simply tries to manage their numbers as best they can, ideally as low as possible. Next, we'll examine some descriptive statistics for the categorical variables in Table 4.3.

Table 4.3: Tabulation of Categorical Variables in Full Data.

Categories	Frequency	Percent
Stem Class		
Less than 50 cm	7,754	7.33
50 to 53 cm	4,545	4.29
53 to 55 cm	20,306	19.19
55 to 58 cm	59,285	56.02
58 cm or more	13,931	13.16
Instances of Plastic Change		
No plastic change	121,297	95.74
Change of Plastic	5,403	4.26

We can see that the flower varieties with stems greater than 55 cm and less than 58 cm are the most populous, followed by those varieties greater than 53 cm and less than 55 cm. Regarding plastic, most of the year, there is no change of plastic. Recall that change of plastic refers to just that: a change. It may have been on the plants in the greenhouse for the previous month, two months, etc.; we were not provided with this information, just the fact that it was either added or removed.

The full dataset is huge, and unsuitable for research within one to two years. Even data cleaning will take a huge amount of time. Therefore, in this thesis, we

have done an initial study of the productivity and related variables using a selected subset of data. Specifically, we have applied PB and NPB standard error estimation methods of MSM for GLMM by using 2015 rose data, and applied random forests to predict productivity by using greenhouse 12 data in 2015.

4.1 Application of MSM with PB and NPB standard error estimators

In this section, we applied the MSM, PB and NPB standard error estimators to a real data example. The data set we used in this example is from year 2015 with 21649 observations. The variables we considered for this example is productivity and cutter. The dataset has a lot of missing data which prevents reliable statistical inference. In this example, since we mainly want to illustrate the use of the methods we've studied in this research. Therefore, for simplicity, we've deleted the missing cases for both of these two variables. We also notice that there are zero productivity. These cases are also excluded from the study. After rough cleaning of the data, we are left with 19206 observations.

The rose plantationers are interested in if there is difference among the cutters regarding high productivity. Therefore, we treated productivity as a categorical variable with high productivity (above or equal 85 quantile) be 1, and not high

productivity (lower than 85 quantile) be 0. Since cutters are randomly employed by the company, and we are interested in the variability among them, we consider cutters (32 cutters available) as a random factor. Tables 4.4, 4.5 and 4.6 give some summary statistics of the data set with 19206 observations.

Table 4.4: Mean Productivity by Cutters.

1	2	3	4	5	6	7	8
371.575	397.067	406.461	292.0415	435.157	332.828	347.396	365.417
9	10	11	12	13	14	15	16
334.313	339.757	362.699	310.639	218.202	322.733	264.864	310.284
17	18	19	20	21	22	23	24
273.888	374.913	328.633	394.026	386.523	228.029	321.164	587.703
25	26	27	28	29	30	31	32
405.476	463.335	513.433	436.973	534.222	609.125	563.512	515.249

Table 4.5: Productivity Percentiles from 19206 observation in 2015.

0	25th	50th	75th	85th	100th
1	191	317	473	597	3115

Using Table 4.5, high productivity is coded as 1 if productivity is higher than 597, and low productivity is coded as 0 if productivity is lower or equal than 597. The high productivity percentage of the cutters are listed as follows:

Table 4.6: Percent High Productivity by Cutters.

1	2	3	4	5	6	7	8
0.1634	0.1767	0.1779	0.0533	0.2767	0.1235	0.1140	0.1424
9	10	11	12	13	14	15	16
0.0923	0.1191	0.1485	0.0700	0.0189	0.1005	0.0543	0.0684
17	18	19	20	21	22	23	24
0.0041	0.0855	0.1753	0.1622	0.1575	0.0422	0.0846	0.3913
25	26	27	28	29	30	31	32
0.2145	0.2349	0.3198	0.1795	0.3184	0.3305	0.3677	0.3510

From the above tables, we've seen variability among the cutters. For example, cutter 24 realized nearly 40 percent of their beds as "high" productivity, while cutter 17 only harvested 0.41 percent of their beds at or above the 85th percentile for productivity. The percent of high productivity varies widely among cutters.

Since the data set is huge with a lot of computation by using PB and NPB methods, plus the data set is only with rough cleaning, therefore, we consider a stratified random sample in order to reduce the computation and to select a representative sample from our rough data. Within each cutter, we randomly selected 60 beds, with a total of $32 * 60 = 1920$ observations. We now apply the glmm logistic model to analyze the data

$$\text{logit}(P(Y_{ij} = 1)) = \mu + \alpha_i, \tag{4.1}$$

where $1 \leq i \leq m$, $1 \leq j \leq k_i$ for each i, and $\alpha_i's$ are i.i.d. normally distributed random variables with mean zero and variance σ^2. Here $m = 32$ is the number of

the cutters and $k_i = k = 60$ is the number of observations within each cutter i.

Following the simulation steps we've described in Section 2.5.3, we set up $n = 1$ (simulation run is 1 since we have a real data); the replication number L for the simulated moments is set to be 1000 in order to find MSM estimators μ and σ^2; the replication number for NPB methods is set up to be 200. We've found the following estimates:

Table 4.7: Results, with $m = 32$ and $k = 60$.

$\hat{\mu}$	PB $\hat{SE}(\hat{\mu})$	NPB $\hat{SE}(\hat{\mu})$	$\hat{\sigma}^2$	PB $\hat{SE}(\hat{\sigma}^2)$	NPB $\hat{SE}(\hat{\sigma}^2)$
-1.7422	.1526	0.2240	0.5907	0.1777	0.1809

In Table 4.7, $\hat{\mu}$ and $\hat{\sigma}^2$ are the average values of the MSM estimators from the 1000 replications. PB $\hat{SE}(\hat{\mu})$ and PB $\hat{SE}(\hat{\sigma}^2)$ are standard error estimates of $\hat{\mu}$ and $\hat{\sigma}^2$ by PB respectively. NPB $\hat{SE}(\hat{\mu})$ and NPB $\hat{SE}(\hat{\sigma}^2)$ are standard error estimates of $\hat{\mu}$ and $\hat{\sigma}^2$ by NPB respectively. The expected proportion is calculated by $p = e^{-1.7422}/(1 + e^{-1.7422}) = 0.1490$, which is close to 15% (above the 85th quantile). The estimated standard errors by PB and NPB are pretty close to each other. $\hat{\sigma}^2/se = 0.5907/0.1777 = 3.32$, which is three standard deviation away from the center. Therefore, we consider the variability between cutters is significant.

4.2 Random Forest

The Random Forest utilized the data from 2015 for Greenhouse 12. Greenhouse 12, with 148 flower beds, was chosen from the other greenhouses after a selection process. We needed a greenhouse with relatively fewer beds, and relatively little missing data across variables; Greenhouse 12 fit these conditions. Full data on cutter in Greenhouse 12 was available for 114 beds over ten months (data on cutter for 34 beds was not recorded by the plantation). Of the remaining 1,140 observations, the first three months (Jan, Feb and Mar) of 2015 were selected for inspection and cleaning, for a subset of 444 observations. Fifty-three percent of the data were missing for the pest variables. Before applying the random forest, missing data were imputed using a regression model via the random.forest package in the RStudio environment. The continuous variable of productivity was recoded into a binary variable indicating high productivity. Criterion for "high" productivity was falling in the top 15th percentile of productivity. Let's first take a look at similar descriptive statistics for the subset of 2015 data for Greenhouse 12, displayed in Table 4.8.

Productivity for Greenhouse 12 in the first three months of 2015 was, on average, lower than the overall productivity across all greenhouses and years, but with a larger standard deviation. This can be accounted for by the high yields that Greenhouse 12 occasionally experienced during this time. The temperature, humidity, and dew

Table 4.8: Descriptive Statistics for Greenhouse 12 for the first three months in 2015.

	Mean	Std Dev	Min	Max
Productivity	444.56	388.23	0	1954
Temperature	16.62	0.53	15.87	17.04
Humidity	76.78	3.03	73.97	80.98
Dew Point	11.51	0.33	11.04	11.78
Spiders	15.18	9.90	3.77	35.16
Aphids	0.06	0.08	0	0.25
Botritis	0	0	0	0
Mold	16.10	12.05	5.06	48.44
Velloso	0	0	0	0

point are all within reason. Greenhouse 12 had less of a problem with spiders, aphids, and mold than greenhouses overall; and no instances of botrytis or velloso were observed during this time. Next, we look at the categorical data for Greenhouse 12, Year 2015, in Table 4.9.

Table 4.9: Tabulation of Categorical Variables in Greenhouse 12.

Categories	Frequency	Percent
Stem Class		
Less than 50 cm	99	22.30
50 to 53 cm	111	25.00
53 to 55 cm	132	29.73
58 cm or more	102	22.97
High and Low Productivity		
Low Productivity	378	85.14
High Productivity	66	14.86

This table displays the frequencies and percentages for stem class and high/low productivity. There were no instances of plastic change during this time period for

Table 4.10: Percent High Productivity by Cutter for Greenhouse 12.

Cutter Number	Percent High Productivity
23	18.06
24	0
25	21.74
26	33.33
28	0

Greenhouse 12, so that variable is eliminated. This greenhouse has a more balanced representation of stem classes, compared to greenhouses overall. Productivity was coded high if it fell in the 85th percentile or higher; for Greenhouse 12 during the first three months of 2015, this corresponded to productivity of 586 or higher. Since we are particularly interested in cutter as a random variable, we also examine the distribution of low/high productivity among the cutters, in Table 4.10:

Notice that Table 4.10 showed that cutters 24 and 28 did not cut in the top 15% of productivity; one may be likely to think that this is due to the number of beds assigned, but we see in Table 4.11 that both cutters are assigned very similar number of beds. These 5 cutters are more or less evenly distributed across flower beds:

Table 4.11: Distribution of Beds among Cutters.

Cutter Number	Number of Beds	Percent
23	72	21.05
24	69	20.18
25	69	20.18
26	69	20.18
28	63	18.41

The random forest procedure fit a model with n=500 trees, with 6 variables tried at each split. Five hundred trees in a forest is more than sufficient to ensure correct classification. In using the rose data, we were interested in the percent of cases correctly classified as either low or high productivity. Our random forest did reasonably well in classifying observations of high or not high productivity based on the independent variables. The out of bag (OOB) estimate of error rate was 5.56 percent. Recall that bagging refers to a method whereby the dataset is split into a training and test group; the OOB error rate refers to the rate of correct classification of the actual observation in the test data based on the data in the training set. We're mostly interested in how accurately our observations of high and not high productivity were classified by the random forest. Table 4.12 shows the confusion matrix:

Table 4.12: Random Forest Confusion Matrix.

	Predicted: Low Prod	Predicted: High Prod	Classification Error
Actual: Low Prod	257	8	0.03018868
Actual: High Prod	11	66	0.14285714

Our random forest was much more accurate in predicting low productivity than high productivity. Our tree did very well at predicting low productivity, with an error rate of 0.03. On the other hand, about 14 % of the observations that it classified as "low" productivity were actually high productivity. Figure 4.4, below, shows the

error rate over the trees, where the black line represents the overall classification error.

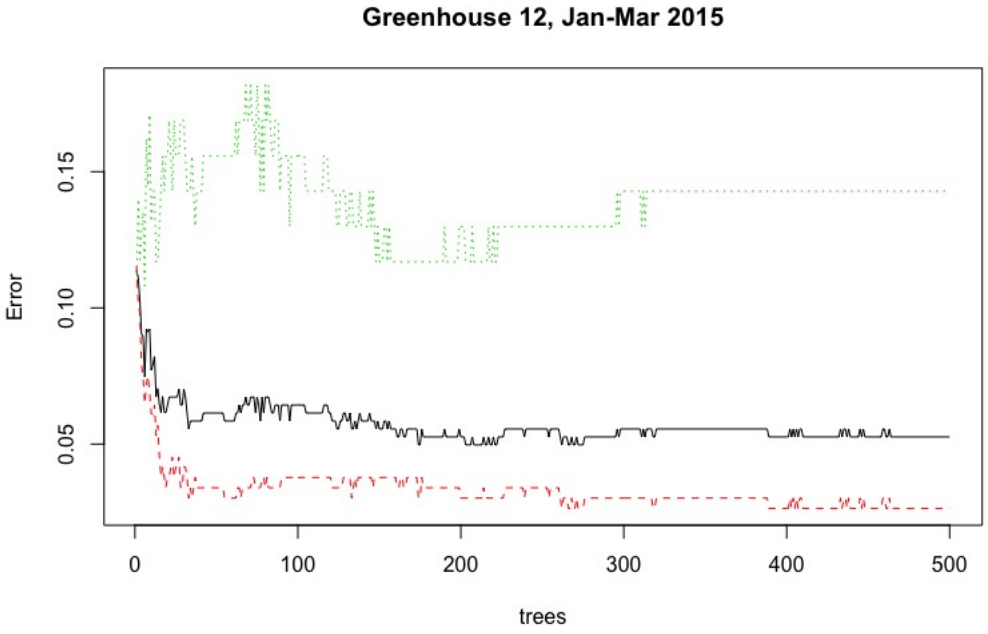


Figure 4.4: Error Rate of Classification over Trees.

The black line represents the overall classification error (around 5.56%); the green line represents the classification error when predicting the high productivity beds (around 14%); the red line represents the classification error when predicting the low productivity beds (around 3%). The green line, indicating inaccurate low productivity classification, bounced around for the first 150 trees and then settled out a little below 0.14. The error rate for high productivity classification started low, around 0.6, and quickly dropped to its average rate of about 0.02. This coincides with

the confusion matrix above, which indicates that 14.3 percent of outcomes classified as low productivity were actually high productivity, and 3.02 percent of outcomes classified as high productivity were actually observations for low productivity. The error rate over trees decreases around 20 trees and stabilizes quickly. The prediction (classification) accuracy can be examined in Figure 4.5, below.

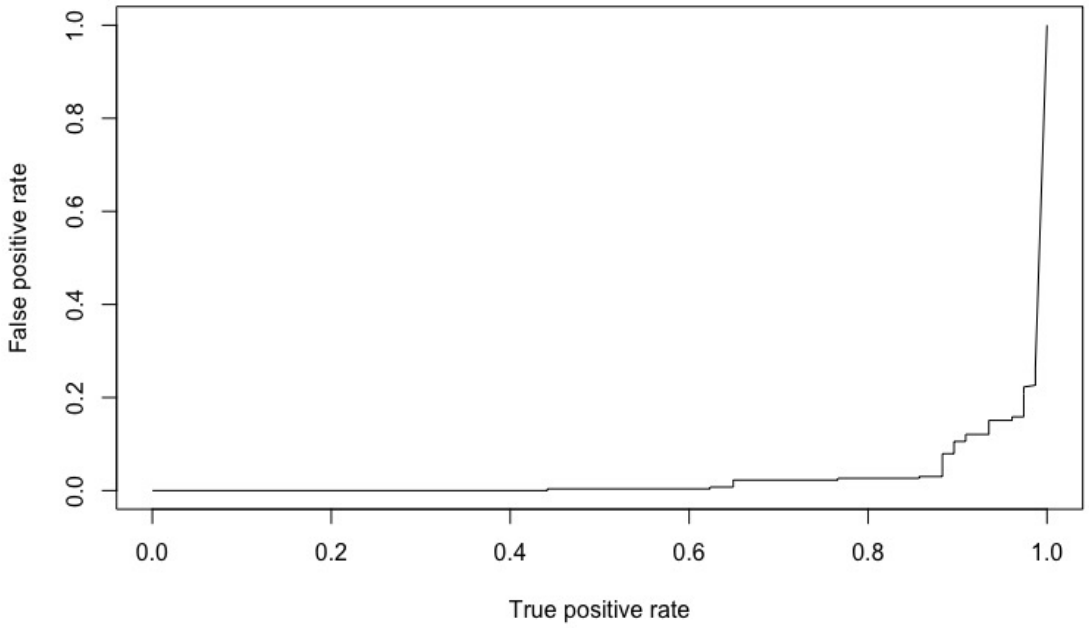


Figure 4.5: False/True Positive Rate for Random Forest.

Figure 4.5 shows the true and false positive classification rate over all trees. A true positive rate of 1.0 would indicate that 100 percent of the trees correctly classified the outcome as either low or high productivity, while a false positive rate of 1.0 would indicate perfect misclassification. Our random forest does not classify with

100 percent accuracy, but as the true positive rate is above 0.90 for the most part, we can say that the random forest is reasonably accurate in classification.

Now that we've seen how accurate our random forest is, we want to know, which variables matter the most in predicting high or not high productivity? The most important predictors are shown in Figure 4.6.

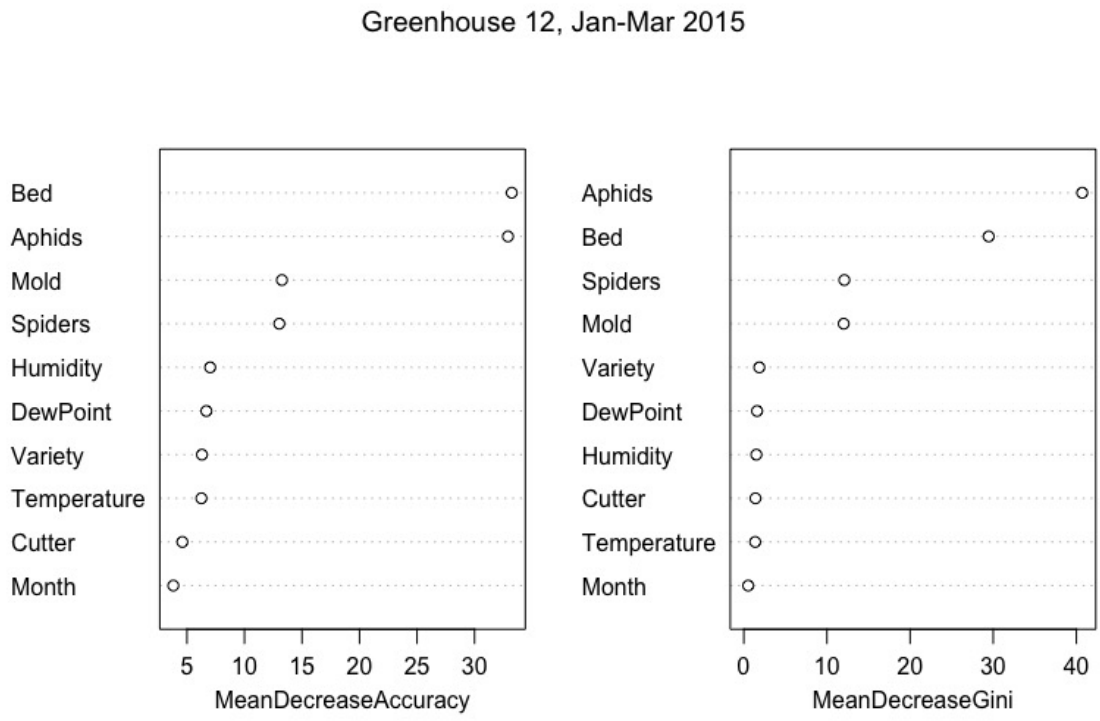


Figure 4.6: Variable Importance Plot for Random Forest. Higher number means the variable is more important.

The predictors are examined in terms of mean decrease in Gini coefficient, which is a measure of how each variable contributes to the homogeneity of the nodes in the random forest. Each time a predictor is selected for to split a node, the Gini

coefficient for that node is calculated and compared to the original node. A Gini coefficient of 1 indicates complete homogeneity, while a coefficient of 0 indicates complete heterogeneity. According to Table 4.13, below, the most important predictors in our random forest were presence of aphids, flower bed, and the presence of spiders.

Table 4.13: Mean Decrease in Gini for Predictors in Random Forest.

Predictor	Mean Decrease Gini
Aphids	40.7090649
Bed	29.4553331
Spiders	12.0904459
Mold	12.0138108
Variety	1.8635322
DewPoint	1.5973287
Humidity	1.5078361
Cutter	1.3987558
Temperature	1.3658488
Month	0.5264053

4.3 Summary Data Application

This thesis has provided a first look at the type of data that is gathered by commercial rose exporters by using random forest and simulation of the standard errors of the GLMM. We applied the MSM with NP and NPB standard error estimators to a subset of data available for 2015. We selected productivity of greater than 0 as a rough cleaning measure, as we are only interested here in simulating the standard errors. Any missing data on cutter was also excluded, as we wanted to show the

standard error estimation among this group. Preliminary inspection showed great variability in the percent of yields that were classified as "high", by cutter. We selected a total of 60 random observations per cutter; with 32 cutters, our n was 1,920 observations for the simulation study. The simulation used here is $n = 1$ because we are using actual observed data. The standard error estimates from PB and NPB were quite similar, most likely because the group of cutters was relatively large. The variability between cutters was also found to be significant in the simulation. Next, we fit a random forest model to Greenhouse 12 for the first three month of 2015. The random forest procedure fit our data rather well, with an OOB error rate of 5.56 percent and low rates of classification error for low productivity. It was harder to classify high productivity, and that's likely due to the artificial creation of a boundary above and below 85 percent. It is likely that the random forest had trouble predicting values that were just above the 85th percentiles of productivity, and incorrectly classified those observations as low productivity. This procedure found that aphids, flower bed, and spiders were the most important variables; after these three variables, mold and humidity were the most important.

Chapter 5

Conclusion/Future Research

In this thesis, we compared the PB and NPB methods of estimating standard error of MSM estimators. The simulation showed that both methods work well when groups are relatively large, but when group size is medium, NPB performs better than PB in estimating the mean, and PB does a better job of estimating the variance than NPB. We also considered a data application of some of the models reviewed. The application of MSM with PB and NPB standard error estimators to our observed data yielded interesting results similar to those of the simulation: when group size is relatively large, PB and NPB estimation methods perform similarly. Another direction is how to work out an algorithm to save time on computation. Current computation is quite expensive. We need to run about 20 hours for calculating the standard errors for each setting. This is an important first look as to how these

methods can be applied to a commercial dataset such as this one.

We've also tried to provide a concise yet sufficient review of the topics covered; from the general linear model to the more complicated GLMM, random forest, and methods of GLMM standard error estimation. We've applied two selected methods to our data, but it is possible that another nonparametric method may suit the data better.

This thesis was a preliminary look at the rose data. Much remains to be done with data of this nature; for starters, we used just one year of data, and through a rough cleaning process selected data where productivity was at least one stem. It may be the case that productivity was actually 0, but at the behest of the plantation, and not because the roses did not produce any stems for that month. Secondly, a more thorough analysis would clean the data carefully before selection. Next, much more could be done with the data in respect to time and greenhouse; an analysis using more than one year of data would yield useful information about trends over time; furthermore, using more than one greenhouse would allow tracking of productivity within and between greenhouses. An application such as this would eventually allow the plantation to see the factors impacting productivity, possibly correct those factors to increase productivity and in turn increase revenue. Since cutter was identified as an important variable, it's imperative to explore more thoroughly the impact of different cutters by including cutter as a nested effect within greenhouse. Another

route to more further explore the data would involve taking the top three or four most important variables identified by the Random Forest and using these as explanatory factors in a GLMM. This thesis has provided a first glance at using this type of data, which is well within the realm of interest of commercial international rose exporters.

References

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with random trees. *Neural Computation*, *9*, 1545-1588.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *26*(2), 123-140.

Breiman, L. (2001). Random forests. *Working Paper*.

Breslow, N. E., & Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal American Statistical Association*, *88*, 9-25.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Annuals of Statistics*, *14*, 1171-1179.

Carlstein, E., Do, K.-A., Hall, P., Hesterberg, T., & Kunsch, H. (1998). Matched-block bootstrap for dependent data. *Bernoulli*, *4*, 305-328.

Davison, A., & Hinkley, D. (1997). *Bootstrap methods and their application.* Cambridge, UK: Cambridge University Press.

References

Demidenko, E. (2005). *Mixed models: Theory and applications.* Hoboken, NJ: John Wiley and Sons, Inc.

Dietterich, R. (1998). An experimental comparison of three methods for costructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning.*

Efron, B. (1979). Bootstrp methods: Another look at the jackknife. *Annuals of Statistics*, *7*, 1-26.

Hall, P. (1992). *The bootstrap and edgeworth expansion.* New York: Springer-Verlag Inc.

Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *Journal American Statistical Association*, *93*, 720-729.

Jiang, J. (2009). *Linear and generalized linear mixed models and their applications.* New York: Springer.

Krishnamoorthy, K., Lu, F., & Mathew, T. (n.d.).

Kunsch, H. R. (1989). The jackknife and bootstrap for general stationary observations. *Annuals of Statistics*, *17*, 1217-1241.

Lahiri, S. (2003). *Resampling methods for dependent data.* New York: Springer-Verlag Inc.

Lin, X., & Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal American Statistical Association*, *91*, 1007-1016.

References

Liu, R. Y., & Singh, K.(1992). Moving blocks jackknife and bootstrap capture weak dependence. *In exploring the limits of bootstrap, Edited by R. Lepage and L. Billard*, 225-248.

Louppe, G.(2014). *Understanding random forests: From theory to practice*. Unpublished doctoral dissertation, University of Liège.

Lu, Y. (2012). Standard error of the method of simulated moment estimator for generalized linear mixed models. *Communications in statistics-simulation and computation*, *42*, 1-7.

McCullagh, P., & Nelder, J.(1989). *Generalized linear models*. Boca Raton: Chapman and Hall.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, *57 (5)*, 995-1026.

Paparoditis, E., & Politis, D. (2001). Tappered block bootstrap. *Biometrika*, *88*, 1105-1119.

Politis, D., & Romano, J.(1992). A circular block-resampling procedure for stationary data. *In exploring the limits of bootstrap*, *263-270*, 1171-1179.

Shao, J., & Tu, D.(1995). *The jackknife and bootstrap*. New York: Springer-Verlag Inc.

Zhang, G. (2015(a)). A parametric bootstrap approach for one-way ANOVA under unequal variances with unbalanced data. *Communications in Statistics-*