Estimation in generalized linear mixed models is challenging because the marginal likelihood is an integral without closed form. In many of the leading approaches such as Laplace approximation and Monte Carlo integration, the marginal likelihood is approximated, and the maximum likelihood estimate (MLE) can only be reached with error. An alternative, the simultaneous perturbation stochastic approximation (SPSA) algorithm is designed to maximize an integral and can be employed to find the exact MLE under the same circumstances. However, the SPSA does not directly provide an error estimate if the algorithm is stopped in a number of finite steps. In order to estimate the MLE properly with a statistical error bound, we propose the stationary SPSA (SSPSA) algorithm. Assuming that the marginal likelihood, objective function, is quadratic around the MLE, the SSPSA takes the form of a random coefficient vector autoregressive process. Under mild conditions, the algorithm yields a strictly stationary sequence where the mean of this sequence is asymptotically unbiased to the MLE and has a closed-form variance. Also, the SSPSA sequence is ergodic providing certain constraints on the step size, a parameter of the algorithm, and the mechanism that directs the algorithm to search the parameter space. Sufficient conditions for the stationarity and ergodicity are provided as a guideline for choosing the step size. Several implementation issues are addressed in the thesis: pairing numerical derivative, scaling, and importance sampling. Following the simulation study, we apply the SSPSA on several GLMMs: Epilepsy seizure data, lung cancer data, and salamander mating data. For the first two cases, SSPSA estimates are similar to published results whereas, for the salamander data, our solution greatly differs from others.
A Stationary Stochastic Approximation Algorithm for Estimation in the GLMM

by

Sheng-Mao Chang

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirement for the degree of Doctor of Philosophy

STATISTICS

Raleigh, North Carolina

2007, May

APPROVED BY:

______________________________
Dr. John F. Monahan
Chair of Advisory Committee

______________________________
Dr. Bibhuti Bhattacharyya

______________________________
Dr. Dennis D. Boos

______________________________
Dr. Daowen Zhang
Dedication

To my parents, my grand mothers, my sister
and in memory of my younger brother
Biography

Chang, Sheng-Mao was born in Taipei, Taiwan, on August 15th, 1974. He earned his Bachelor degree in Statistics in 1996 and Master degree in Statistics in 1998 from National Taipei University (transformed from Law and Commerce College, National Chung-Hsing University). After graduation, he is joining the Department of Statistics at the National Cheng Kung University, Taiwan, as an Assistant Professor starting Fall 2007.
Acknowledgements

I would like to thank the members of my advisory committee for their helpful and constructive advice.

I am extremely grateful to my advisor Prof. John F. Monahan for his knowledgeable guidance, unlimited patience, and priceless time. Every time we met, I feel inspired. On the wall of my “Free Expression Tunnel”, there is no complain but many thanks to him. I firmly believe that he is the blessing from the God that leads me to fulfill my faith.

I would like to say thank you to Dr. B.B., Dr. Boos, Dr. Davidian, Dr. Tsiatis, Dr. Genton and Dr. Zhang for their wonderful teaching. In addition, I would like to show my sincere appreciation to Dr. Stefanski and Dr. Lindsay because they changed the way I look a problem. Most importantly, I appreciate Dr. Zeng for his motivating direction and long term financial support. Without his help, I can not complete this long journey.

I also appreciate Dr. Jung-Ying Tzeng for her advice and encouragement during my most depressed days. Her support really made a remarkable change of my life. A special appreciation should go to my dear friends from Taiwanese Student Association. The friendship they provide to me is more precious than ever. Finally, I would like to thank my parents, my grand mothers and my sister for their endless love.

“You are the wind beneath my wings.”
# Contents

List of Tables vii

List of Figures viii

1 Introduction 1

2 A Review of GLMM 6
  2.1 Linear Mixed Model .............................................. 7
  2.1.1 General Formulation ............................................. 7
  2.1.2 Restricted Maximum Likelihood ............................... 8
  2.2 GLMM: Approximate Integral ..................................... 9
  2.2.1 Gaussian Quadrature Approximation ........................... 9
  2.2.2 Monte Carlo Approximation .................................... 11
  2.2.3 Radial-Spherical Approximation ............................... 12
  2.3 GLMM: Approximate Integrand .................................... 14
  2.3.1 Penalized Quasi-Likelihood ................................... 14
  2.3.2 Pseudo-Likelihood ............................................. 16
  2.3.3 Logistic-Normal Approximation ............................... 17
  2.4 GLMM: EM Algorithm ............................................. 18
  2.5 GLMM: Semiparametric, Nonparametric, and Bayesian .......... 21

3 A Review of Stochastic Approximation 24
  3.1 SA – Root Finding .................................................. 25
  3.2 KWSA – Optimization ............................................. 27
  3.3 Simultaneous Perturbation SA .................................... 28
  3.4 Implementation of SPSA on GLMM ................................. 29
    3.4.1 Choosing Differencing Sequence ............................... 31
    3.4.2 Pairing ..................................................... 32
    3.4.3 Parameter Scaling ............................................ 33
    3.4.4 Importance Distribution ..................................... 33
  3.5 GLMM: using SA’s ................................................ 34
## 4 Stationary Simultaneous Perturbation Stochastic Approximation

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Motivating Example</td>
<td>37</td>
</tr>
<tr>
<td>4.2</td>
<td>SSPSA with Quadratic Objective Function</td>
<td>38</td>
</tr>
<tr>
<td>4.3</td>
<td>The Stationarity of SSPSA</td>
<td>40</td>
</tr>
<tr>
<td>4.4</td>
<td>Mean and Covariance of SSPSA</td>
<td>43</td>
</tr>
<tr>
<td>4.5</td>
<td>The Ergodicity of SSPSA</td>
<td>49</td>
</tr>
</tbody>
</table>

## 5 Simulation and Case Studies

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Simulation Studies</td>
<td>53</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Normal Regression with Normal Random Intercept</td>
<td>55</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Logistic Regression with Normal Random Intercept</td>
<td>58</td>
</tr>
<tr>
<td>5.2</td>
<td>Case Studies</td>
<td>59</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Epilepsy Seizure Data</td>
<td>60</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Lung Cancer Data</td>
<td>62</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Salamander Data</td>
<td>63</td>
</tr>
</tbody>
</table>

## 6 Future Work

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Properties of SSPSA</td>
<td>67</td>
</tr>
<tr>
<td>6.2</td>
<td>Implementation Issues</td>
<td>68</td>
</tr>
<tr>
<td>6.3</td>
<td>Possible Applications</td>
<td>68</td>
</tr>
</tbody>
</table>

## Bibliography

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>70</td>
</tr>
</tbody>
</table>

## Appendix

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Some Matrix Operations</td>
<td>80</td>
</tr>
<tr>
<td>A.1</td>
<td>Matrix Norm</td>
<td>80</td>
</tr>
<tr>
<td>A.2</td>
<td>von Neuman Series</td>
<td>81</td>
</tr>
<tr>
<td>A.3</td>
<td>Some Properties of Kronecker Product</td>
<td>82</td>
</tr>
<tr>
<td>B</td>
<td>Moments of Random Directions</td>
<td>83</td>
</tr>
<tr>
<td>B.1</td>
<td>Bernoulli Random Direction</td>
<td>83</td>
</tr>
<tr>
<td>B.2</td>
<td>Random Vector on Unit Sphere</td>
<td>83</td>
</tr>
<tr>
<td>C</td>
<td>Properties of Working Matrices</td>
<td>87</td>
</tr>
</tbody>
</table>
List of Tables

1.1 Lung Cancer Data, Booth and Hobert (1999) .................................. 2
5.1 Mean Bias and Empirical Standard Error of Normal-Normal Model . 56
5.2 Mean Square Error ($\times 100$) of Normal-Normal Model .................. 56
5.3 Multiple Comparisons with Control of N-N Model: $p$-values of Testing No Difference among Parameter Estimates .......................... 57
5.4 Mean Bias and Empirical Standard Error of Logistic-Normal Model . 60
5.5 Mean Square Error ($\times 100$) of Logistic-Normal Model .................. 60
5.6 Pairwise Multiple Comparisons of L-N Model: $p$-values of Testing No Difference of Parameter Estimates ............................. 61
5.7 Summary Results, Parameter Estimates (Standard Errors), for Epilepsy Seizure Data ................................................................. 62
5.8 Summary Results, Parameter Estimates (Standard Errors), for Lung Cancer Data ................................................................. 63
5.9 Summary Results, Parameter Estimates (Standard Errors), for Salamander Mating Data .......................................................... 66
A.1 The value of $\tau_{pq}$ when $||A||_p \leq \tau_{pq}||A||_q$, Stewart (1973). ............... 81
# List of Figures

4.1 Effect of step size $s_t$ on SPSA sequences. .......................... 38

5.1 N-N model, sample size 100: Comparisons of three methods: SSPSA, AGQ and RSPL. Exact MLEs versus approximated MLEs are plotted. 57

5.2 N-N model, sample size 30: Comparisons of three methods: SSPSA, AGQ and RSPL. Exact MLEs versus approximated MLEs are plotted. 58

5.3 N-N model, $n = 100$ (Left) and $n = 30$ (Right): Variation due to SSPSA with different numbers of iterations ............................. 59

5.4 L-N model, $n = 100$ (Left) and $n = 30$ (Right): Variation due to SSPSA with different numbers of iterations ............................. 61

5.5 Lung cancer data: Monte Carlo log-likelihood surface of $\sigma^2_u$ and $\sigma^2_v$ when $(\beta_0, \beta) = (-1.9, 1.7)$. ......................................................... 64

5.6 SSPSA sequences for Salamander Mating Data with initial value provided by Gibbs estimate, Zeger and Karim, 1991 .......................... 65

5.7 Monte Carlo marginal likelihood with (Left) SSPSA and with (Right) Gibbs fixed-effect coefficient estimates. ................................. 66

B.1 The $\xi_s^2$ function ................................................................. 85
Chapter 1

Introduction

The generalized linear mixed model (GLMM) has served as a useful tool to scientific discovery. This kind of model consists of a response following a distribution belonging to the exponential family, a link function, and covariates with random and/or fixed coefficients. Through the assigned link function, the conditional mean of the response is a function of a linear combination of covariates. Following are some examples:

- **Epileptic Seizure Data** (Thall and Vail, 1990) In a clinical trial, 59 epileptic patients were randomly assigned to have new drug or placebo (treatment). The response is the number of seizures among 4 clinics and the covariates are: baseline seizures, age, and, of course, the treatment. In addition, it is reasonable to assume that individuals have different reactions to different treatments. With multiple observations on each individual, thus random effect can be identified and estimated.

- **Lung Cancer Data** Lung cancer data in Booth and Hobert (1999) provides a good example of GLMM. In their Table 1 (reproduced in Table 1.1) the result of 14 studies of the incidence of lung cancer among smokers and non-smokers are presented. While the primary interest is on the effect of smoking on the
incidence of lung cancer, evidence presented in Section 5.2.2 suggests a random study effect.

- **Salamander Data**  
  McCullagh and Nelder (1989) reported an interesting dataset which contains information about the mating preference of two isolated salamander populations. Each female individual was assigned to mate with three males which may or may not be in the same population. The mating design is not balanced in order to avoid time effects. Since there are three measurements (successfully mated or not) per individual it is reasonable to take individual effects into account. Again, individual effect is assumed to be random.

In all of these examples the response variable is discrete and the model contains random effects.

<table>
<thead>
<tr>
<th>Study</th>
<th>Smoking Lung Cancer</th>
<th>Smoking Total</th>
<th>Non-smoking Lung Cancer</th>
<th>Non-smoking Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>83</td>
<td>155</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
<td>317</td>
<td>3</td>
<td>46</td>
</tr>
<tr>
<td>3</td>
<td>129</td>
<td>210</td>
<td>7</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>467</td>
<td>12</td>
<td>137</td>
</tr>
<tr>
<td>5</td>
<td>412</td>
<td>711</td>
<td>32</td>
<td>163</td>
</tr>
<tr>
<td>6</td>
<td>597</td>
<td>1263</td>
<td>8</td>
<td>122</td>
</tr>
<tr>
<td>7</td>
<td>88</td>
<td>262</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>1350</td>
<td>2646</td>
<td>7</td>
<td>68</td>
</tr>
<tr>
<td>9</td>
<td>60</td>
<td>166</td>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>459</td>
<td>993</td>
<td>18</td>
<td>99</td>
</tr>
<tr>
<td>11</td>
<td>724</td>
<td>970</td>
<td>4</td>
<td>58</td>
</tr>
<tr>
<td>12</td>
<td>499</td>
<td>961</td>
<td>19</td>
<td>75</td>
</tr>
<tr>
<td>13</td>
<td>451</td>
<td>2180</td>
<td>39</td>
<td>675</td>
</tr>
<tr>
<td>14</td>
<td>260</td>
<td>519</td>
<td>5</td>
<td>33</td>
</tr>
</tbody>
</table>

The GLMM provides a good statistical tool to analyze these kinds of data. The models are characterized by the response that follows a distribution from the ex-
ponential family whose mean is a one-to-one function of the linear combination of covariates (or predictors or independent variables.) For example, in the seizure data, a simplified model of Breslow and Clayton (1993) is considered. The response $y_{ij}$ is the number of seizures of patient $i$ at the $j$th visit to a clinic. It is reasonable to assume $y_{ij}$ follows Poisson distribution with mean $\mu_{ij}$ and

$$\log \mu_{ij} = X_{ij}^T \beta + u_i,$$

where $X_{ij}$ is a vector containing the fixed effect covariates, $\beta$ representing the corresponding unknown coefficient, and intercept and $u_i$ denoting the individual variation of individual $i$. The random variable $u_j$’s are normally distributed with a common variance.

In the lung cancer data, for the study $i$ and smoking group $j$ ($j = 1$ if smoking and $j = 0$ if non-smoking), the response follows the binomial($n_{ij}, p_{ij}$) distribution with the link function $\log(p_{ij}/(1 - p_{ij})) \equiv \logit(p_{ij}) = \beta_0 + \beta_1 x_{ij} + u_i + v_{ij}$. Here, $x_{ij}$ indicates the smoking status $j$ of study $i$; $u_i$ represents the random variation of study $i$; $v_{ij}$ denotes the random effect of smoking status $j$ within study $i$. Note that $\beta$’s stand for the coefficients of fixed effects and $u_i$’s and $v_{ij}$’s follow normal distributions with mean zero and unknown variance.

In the salamander data, Karim and Zeger (1992) proposed the random effects model as follows

$$\logit(\Pr(Y_{ij} = 1| X_{ij}, u_i^f, u_j^m)) = X_{ij}^T \beta + u_i^f + u_j^m$$

where the vector $X_{ij}$ indicates the population of two individuals, the $i$th female and the $j$th male, and the $u_i^f$ and $u_j^m$ represent the corresponding individual effects. Without further information, we assume $u_i^f \sim N(0, \sigma_f^2)$, $u_j^m \sim N(0, \sigma_m^2)$ and mutual independence among individuals.

In general, one way to analyze above models is to maximize the marginal likelihood obtained by integrating the random effect from the likelihood. The marginal
likelihood of the $i$th individual has the form

$$L_i(\gamma, D) = \int_{\mathbb{R}^q} p(y_i | W_i, \mathbf{u}; \gamma) d\Phi(\mathbf{u}; 0, D)$$ (1.1)

where $p(\cdot | W_i, \mathbf{u}; \gamma)$ denotes the density function of the response, usually belongs to exponential family, $\mathbf{u}$ follows a $q$-dimensional multivariate normal distribution with mean zero and variance $D_{q \times q}$, and $\gamma$ is a vector of parameters. The response is $y_i$ and $W_i$ is a row vector of covariates which corresponds to the $i$th individual. Throughout this work, we reserve $\phi(\cdot; \mu, \Sigma)$ and $\Phi(\cdot; \mu, \Sigma)$ to denote the normal density and distribution function, respectively, with mean $\mu$ and variance-covariance $\Sigma$. Both $\gamma$ and $D$ are unknown parameters. For convenience, denote $\theta^T = (\gamma^T, \text{vec}(D)^T)$. The linear relationship is characterized by

$$E(Y_i | W_i, \mathbf{u}) = \mu_i \quad \text{and} \quad \eta(\mu_i) = X_i^T \beta_{p \times 1} + Z_i^T \mathbf{u}_{q \times 1}$$

where $W_{n \times (p+q)} = [X_{n \times p}, Z_{n \times q}]$, $W_i$ is the $i$th row of $W$, and $\eta(\cdot)$ is the so-called link function in the generalized linear model literature. Additionally, the link function should be one-to-one and be pre-specified before analysis.

The remainder of this work is arranged as follows. In Chapter 2, we review the literature of GLMM, beginning with the LMM and then show the difficulties in its generalized version GLMM. Several major approaches are mentioned, including approximate integral, approximate integrand, and EM. In Chapter 3, we review the literature on stochastic approximation (SA). SA is designed for root finding or optimization of a non-deterministic objective function. Several of its extensions are reviewed as well. In Section 3.4, we discuss some implementation issues about applying simultaneous perturbation SA on marginal likelihood like (1.1). We close chapter 3 with a short survey of some SA methods applied to GLMM. In Chapter 4, we define the stationary simultaneous perturbation SA. With the assertion that, around the maximum, the objective function can be well approximated by a quadratic function, we show that the proposed algorithm has the form of random coefficient autoregressive (RCA).
model. Also, we provide the sufficient conditions for stationarity and ergodicity, and demonstrate the unbiasedness and consistency of our estimator. Next, in Chapter 5, simulation studies and three aforementioned cases are analyzed.
Chapter 2

A Review of GLMM

The most well studied subclass of GLMMs is the linear mixed model (LMM) – replacing \( p(\cdot) \) by \( \phi(\cdot) \) in (1.1) and taking the link function as identity, \( i.e. \eta(x) = x \). For a detailed survey, we refer to Verbeke and Molenberghs (2001) and McCulloch and Searle (2001). Similar conjugate classes in the statistical literature are the Beta-binomial model (Lee and Nelder, 1996), Poisson-Inverse Gamma model (Breslow, 1984), and other over-dispersion models. A favorable feature of these models is that their marginal likelihood has a closed form so the MLE is tenable by taking derivatives. On the contrary, if \( p(\cdot) \) in (1.1) is chosen so that the corresponding integral has no closed form then approximations or numerical methods are required. Several popular solutions are examined below. Note that most of leading methods are combination of them.

The rest of this chapter will be arranged as follows. First of all, Section 2.1 introduces linear mixed models. Several fundamental issues are addressed. Most of them are benchmarks of GLMM as well. Second, as shown in (1.1), the integral has no analytical solution so some remedies are required. Methods that rely on approximating the integration are summarized in Section 2.2. On the other hand, one can approximate the integrand so that the integral has a closed-form representation (Section 2.3.) Last, we discuss the use of EM algorithm in Section 2.4 and in Section...
2.5, Bayesian and non-fully-parametric methods are briefly mentioned.

2.1 Linear Mixed Model

The linear mixed model can describe a variety of statistical problems such as analysis of variance (Searle et al., 1991), longitudinal experiment (Verbeke and Molenberghs, 2001), linear models with latent variables (Fuller, 1987), and others. In this section, we consider its general formulation and the construction and merits of restricted maximum likelihood estimate (REML). These ideas remain important in GLMM inference and computation.

2.1.1 General Formulation

The linear mixed model can be defined as $y = X\beta + Zu + \epsilon$ where $\epsilon_{n \times 1} \sim N(0, R_{n \times n})$, $u_{q \times 1} \sim N(0, D_{q \times q})$ and $\epsilon$ and $u$ are independent. Note that, $(y_{n \times 1}, X_{n \times p}, Z_{n \times q})$ is defined as before; $\beta_{p \times 1}$ denotes the fixed effect coefficient vector whereas $u_{q \times 1}$ denotes the random effect coefficient vector. Taking advantage of conditional expectation, we have

$$E(Y|X, Z) = E(E(Y|u)) = E(X\beta + Zu) = X\beta$$

and

$$Var(Y|X, Z) = E(Var(Y|u)) + Var(E(Y|u)) = R + ZDZ^T.$$

Thus, we have $Y$ follows normal distribution with mean $X\beta$ and variance $R + ZDZ^T$. It turns out that, analogous to (1.1), the marginal likelihood is

$$(2\pi)^{-n/2}|R + ZDZ^T|^{-1/2} \exp \left\{ -\frac{1}{2} (y - X\beta)^T(R + ZDZ^T)^{-1}(y - X\beta) \right\}.$$ 

Although the marginal likelihood is simple, deriving the maximum likelihood estimator (MLE) can be complicated. A detailed derivation can be seen from Wolfinger et al. (1994). They also provided two ready-to-use optimization methods: Newton-Raphson and Fisher-scoring.
2.1.2 Restricted Maximum Likelihood

In random effect methods, sometimes we are more interested in the variance components, \( D \) and \( R \), rather than the other parameters, \( \gamma \). Suppose that variance components are our major interest so we can treat the fixed effect coefficient \( \beta \) as a nuisance parameter. One way to deal with the nuisance parameter is projecting the model onto a subspace so that the estimation of \( \beta \) does not affect the estimation of variance components. For example, let \( K_{n \times n} \) be a real matrix such that

\[
K_y = KX\beta + KZu + K\epsilon = KZu + K\epsilon,
\]

and \( KX\beta = 0 \) for all \( \beta \in \mathbb{R}^p \). Then the distribution of \( K_y \) is normal with mean \( 0 \) and variance \( KZDZ^TK^T + KRK^T \), free of \( \beta \). The ML inference using this kind of transformed model is called REML, Thompson (1962). A possible \( K \) is \( I_n - P_X \) where \( P_X = X(X^TX)^{-1}X^T \), where \( "-" \) denotes a generalized inverse. It turns out that the transformed response

\[
K_y = (I - P_X)y
\]

is exactly the residual of linear regression with fixed effects only. So the “RE” in REML stands for the “residual” as well.

The benefit of using REML is that the estimator of variance component is unaffected by the estimation of fixed effects. Recall that the MLE has good properties such as asymptotic normality, consistency, and efficiency. However, it is not guaranteed to be unbiased and this bias may not disappear in certain mixed models. The reason why it is biased (for variance component estimate) is that it does not take into account the degrees of freedom (df) used to estimate fixed effects. On the other hand, REML not only removes the fixed effect but also lowers the rank of its design matrix, and attempts to correct the degrees of freedom. The REML estimate is usually computed iteratively. We first assign an initial values for \( R \) and then estimate \( \beta \) by weighted least squares (WLS), the residuals are used to compute the estimate of \( R \) and \( D \). Then we recalculate residuals by WLS and iterate until convergence.
2.2 GLMM: Approximate Integral

The first solution is to approximate the integral in (1.1). We review three approaches. First, one may approximate the integral with fixed abscissas such as trapezoid rule, Simpson’s rule, and Gaussian quadrature (GQ). This kind of approximation usually has low errors but it is not unbiased and it is not easy to control the accuracy. We introduce GQ in Section 2.2.1. Second, one may construct an unbiased estimator of the integral, e.g., Monte Carlo (MC) and importance sampling. This kind of estimation is unbiased and the corresponding variance is easy to derive, see Section 2.2.2. However, the convergence rate of MC is slow. The third possibility is combining the above two together, e.g., the Radial-Spherical (RS) method, and thus, inheriting those merits at the same time, see Section 2.2.3. In the following, we are going to introduce these three major methods to approximate or to estimate the integration. Hereafter, denote $p_i(u)$ as the shorthand of $p(y_i|W_i, u; D)$ if there is no confusion.

2.2.1 Gaussian Quadrature Approximation

Gaussian quadrature (or Gauss-Hermite quadrature) serves as an approximation of the integral having the form

$$
\int_{\mathbb{R}} f(z) e^{-z^2/2} dz
$$

(2.1)

where $f(\cdot)$ is a smooth function. The $m$th order GQ approximation is defined as

$$
\sum_{j=1}^{m} w_j f(z_j) \approx \int_{\mathbb{R}} f(z) e^{-z^2/2} dz
$$

(2.2)

where $\{z_j\}$ are roots of $m$th order Hermite polynomial and $\{w_j\}$ are suitable weights. This approximation is useful, in particular, when the integral has the form like (1.1). Consider (1.1) with one-dimensional $u \sim N(\mu, \sigma^2)$ for example. Let $z = (u - \mu)/\sigma$. The likelihood in (1.1) can be rewritten as

$$
L_i = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} p_i(\sigma z + \mu) e^{-z^2/2} dz
$$
so GQ can be applied directly. For a more general class like \( \int g(x)dx, g : \mathbb{R} \to \mathbb{R} \),

one can also apply GQ to approximate the integral (Liu and Pierce, 1994). A general

form is as follow:

\[
\int g(x)dx = \int \frac{g(x)}{\phi(x; \mu, \sigma^2)} \phi(x; \mu, \sigma^2)dx = \int \frac{g(\mu + \sigma z)}{\phi(z; 0, 1)} \times \frac{1}{\sqrt{2\pi}} \times e^{-z^2/2}dz.
\]

So we go back to the form of (2.1) and hence, weights \( \{w_j\} \)'s in (2.2) are still valid.

In practice, the random variable \( x \) usually has higher dimension. Fortunately, the GQ technique remains similar; simply apply the product rule. For example, let

\( z = (z_1, z_2)^T \) and \( f : \mathbb{R}^2 \to \mathbb{R} \) then the corresponding GQ has the form

\[
\int_{\mathbb{R}^2} f(z)e^{-z^Tz}dz \approx \sum_{i=1}^n \sum_{j=1}^m w_{in}w_{jm}f(z_{in}, z_{jm})
\]

where \( \{z_{in}, z_{jm}\} \)'s are relevant abscissas and \( \{w_{in}, w_{jm}\} \)'s are corresponding weights.

If the \( d \times d \) dimensional random variable \( u \) follows a normal distribution with mean \( \mu_d \times 1 \) and variance \( D_{d \times d} \) which is positive definite, then the relevant transformation is \( z = L^{-1}(u - \mu) \) where \( D = LL^T \) is the Cholesky decomposition and \( L \) is a lower triangular \( d \times d \) matrix. Consequently,

\[
\int_{\mathbb{R}^d} p_i(u)\phi(u; \mu, D)du = \int_{\mathbb{R}^d} p_i(Lz + \mu)\phi(z; 0, I)dz.
\]

For a general introduction and implementation, we refer to Monahan (2001) and for a detailed treatment, we refer to Davis and Rabinowitz (1984).

A more efficient way to apply GQ in GLMM is adaptive Gaussian quadrature (AGQ) proposed by Pinheiro and Bates (1995) for nonlinear mixed models. Unlike GQ which centers \( u \) at the mean \( \mathbf{0} \), AGQ centers it at \( \hat{u} \), a reasonable estimate of the mean of random vector \( u \). This modification works better as long as \( \hat{u} \) is much closer to the sample center than the population center \( \mathbf{0} \).
2.2.2 Monte Carlo Approximation

A straightforward method of approximating an integral is MC integration. That is, sample $u_j$’s i.i.d. $\Phi(u; 0, D)$ so that the sample mean gives

$$\frac{1}{m} \sum_{j=1}^{m} p_i(u_j) \approx \int p_i(u) d\Phi(u; 0, D).$$

According to the strong law of large numbers, the left-hand side converges to the right-hand side. If sampling from $\phi(\cdot)$ is difficult or inefficient, then importance sampling can be used. With importance density $g(\cdot)$, the approximation becomes

$$\frac{1}{m} \sum_{j=1}^{m} p_i(u_j) \phi(u_j; 0, D) \frac{1}{g(u_j)}$$

so that

$$E \left\{ \frac{1}{m} \sum_{j=1}^{m} p_i(u_j) \phi(u_j; 0, D) \right\} = \int p_i(u) d\Phi(u; 0, D)$$

where the $u_j$’s are i.i.d. samples having density function $g(\cdot)$. The choice of $g(\cdot)$ is crucial. At a minimum, samples need to be drawn from it and $g$ has to be absolutely continuous with respect to $\phi$. In practice, when the underlying distribution is normal a Student-$t$ distribution can be the importance distribution which provides a more robust result (although it is not the most efficient choice), Evans and Swartz (1996). Other choices for sampling efficiency and practical issues can be seen in Srinivasan (2002). For references discussing GLMM and nonlinear mixed model with approximate likelihood, we refer to Durbin and Koopman (1997), Pinheiro and Bates (1995), and references therein.

Following this line, different techniques of drawing samples yield different approximations. In literature, Wei and Tanner (1990) and Chan and Ledolter (1995) discussed the MC integration method, Zeger and Karim (1991) and McCulloch (1994) proposed a Gibbs sampler, McCulloch (1997) offered a Metropolis-Hastings algorithm, and later, Booth and Hobert (1999) defined a rejection sampler and an importance sampler. These approximations together with the EM algorithm or Newton-Raphson...
algorithm have been successfully proposed as solutions of GLMM problem (see Chen et al., 2002, for example.) Among these methods, the efficiency of algorithms is rooted in the sample size drawn from the underlying sampler. We will come back to this issue when we discuss the EM algorithm.

As Monahan and Genz (1997) pointed out in a $d$–dimensional integration problem, the numerical quadrature may converge with rate $O(m^{-2/d})$ or $O(m^{-4/d})$ where $m$ is the number of abscissas, whereas the MC method converges with $O(n^{-1/2})$ where $n$ represents the number of function evaluation. The curse of dimensionality operates on fixed quadrature severely but not on the MC. The other drawback of quadrature is the difficulty of evaluating its accuracy. An idea that fuses GQ and MC together has been proposed by Siegel and O’Brien (1985) and Genz and Monahan (1998). We demonstrate this idea in next subsection; however, we refer Chapter 12 of Monahan (2001) for a detailed treatment.

2.2.3 Radial-Spherical Approximation

Consider integrating the function $f : [0, 1] \to \mathbb{R}$ using

$$
\hat{I}_f = \frac{1}{n} \sum_{i=1}^{n} f \left( U + \frac{i - 1}{n} \right)
$$

where $U$ follows a uniform distribution at interval $[0, 1/n]$. Now, since $\hat{I}_f$ is a random variable, we have

$$
\mathbb{E}_U[\hat{I}_f] = \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{1/n} n \times f \left( u + \frac{i - 1}{n} \right) du = \int_{0}^{1} f(v) dv.
$$

In words, $\hat{I}_f$ is an unbiased estimator of $\int_{0}^{1} f(v) dv$. Together with the idea of quadrature, Siegel and O’Brien (1985) suggested an unbiased integral rule which integrates over the interval $[-1, 1]$:

$$
T_f(R) = w_0(R)f(0) + w_1(R)h(-R) + w_1(R)h(R)
$$
where \((w_0(R), w_1(R)) = (2 - 2/3R^2, 1/3R^2)\) and \(R\) has the density function \(3r^2\) on 
\([0, 1]\). For any function \(f\), \(T_f(R)\) is unbiased, i.e. \(E_R T_f(R) = \int_{-1}^{1} f(r) dr\) and \(T_f(R)\) is exact if \(f\) is a cubic polynomial.

Further, Genz and Monahan (1998) proposed another integral rule \(M_f^k(R)\) which is an unbiased estimator of

\[
\frac{2^{-d/2}}{\Gamma(d/2)} \int_{-\infty}^{\infty} f(r)|r|^{d-1}e^{-r^2/2}dr = \frac{1}{2} \int_{-\infty}^{\infty} f(r)(f(r) + f(-r))dr
\]

where the superscript of \(M\) denotes the degree of rule. Take the third-order rule for example. Let \(R \sim \chi_{d+2}\) and weights \(w_0(R) = 1 - d/R^2\) and \(w_1(R) = d/4R^2\). So \(M_f^3(R)\) has the form

\[
M_f^3(R) = w_0(R)f(0) + w_1(R)(f(R) + f(-R))/2.
\]

Formulas for the \(w_i\)’s are given in Genz and Monahan (1998) for various \(k\).

The other useful integral rule is for the integration over the surface of the \(d\)—dimensional unit ball denoted as \(U_d\). Stewart (1980) proposed an efficient algorithm to compute a random orthogonal matrix \(Q\) such that

\[
\int_{U_d} f(z)dz = \mathbb{E} \left[ \sum_{i=1}^{k} u_i f(Qv^{(i)}) \right]
\]

where \(v^{(i)}\) denotes the \(i\)th abscissa (a \(d\)—dimensional vector) and \(u_i\) is its corresponding weight. Suppose \(X\) is a matrix where each element in \(X\) is a random sample from standard normal distribution. Its QR decomposition can be denoted as \(X = Q_1R_1\) where columns of \(Q_1\) are orthogonal and \(R_1\) is upper triangular. Then, \(Q_1\) is a valid random matrix of \(Q\) for radial-spherical approximation. Actually, \(Q\) randomly rotates \(v^{(i)}\)’s so that the weighted sum (2.4) becomes a random variable and is unbiased.

Monahan and Genz (1997) proposed a stochastic version of integration called the Radial-Spherical (RS) approximation. The idea is decompose the the whole integral, say \(d\) dimensional, into two: radial integral, 1—dimensional, and spherical integral,
\[(d - 1)\)-dimensional,

\[
\int_{R^d} g(u)\phi(u)du = \int_{R^+} \int_{U_d} g(rz)\phi(rz)dzd^{d-1}r.
\]

Obviously, the inner integral can be approximated by (2.4) and the outer can be approximated by (2.3). This suggests that the inner integral can be estimated by

\[
G_{n_1}(r) = \frac{1}{n_1} \sum_{j=1}^{n_1} \sum_{i=1}^{k_1} u_i g(rQ^{(j)}v^{(i)}) \phi(rQ^{(j)}v^{(i)})
\]

and thus, the estimator for the whole integration becomes

\[
\frac{1}{n_2} \sum_{j=1}^{n_2} \sum_{i=0}^{k_2-1} w_i(r_j)(G_{n_1}(r_j) + G_{n_1}(-r_j))/2.
\]

As suggested by Monahan and Genz (1997), a better approximation is a mixed RS approximation: adopting (2.4) for estimating the inner integral and using GQ for the outer integral. Clarkson and Zhan (2002) implemented GLMM fitting by the mixed RS approximation. To sum up, RS approximation is very efficient and precise as long as the integrand \(g\) is smooth or symmetric.

### 2.3 GLMM: Approximate Integrand

In this section, we consider methods that replace \(g\) by its approximated version \(\hat{g}\) in (1.1) so that the integral has a closed form. Popular choices are penalized quasi-likelihood (PQL) and pseudo-likelihood (PL). For logistic regression with a random effect, we describe the the logistic-normal approximation. Next, we review these methods and investigate how they work.

#### 2.3.1 Penalized Quasi-Likelihood

Breslow and Clayton (1993) is perhaps the most cited paper in the GLMM literature. Their milestone proposal, penalized quasi-likelihood (PQL), takes advantage
of a locally quadratic approximation of the integrand and the Laplace approximation. The approximated likelihood can be derived by following steps. First, we define

\[ p(u) = \exp\{k(u)\} \]

and expand \( k \) at \( u^\ast \) where \( \partial k(u)/\partial u|_{u=u^\ast} = 0 \), i.e. \( u^\ast \) is the maximum. By doing so, we can define

\[ \hat{p}(u) = \exp\left\{ k(u^\ast) - \frac{1}{2}(u - u^\ast)^T K^\ast (u - u^\ast) \right\} \]

where

\[ K^\ast = -\frac{\partial^2 k(u)}{\partial u \partial u^T}|_{u=u^\ast}. \]

Then, PQL approximates (1.1) by replacing \( p_i \) by \( \hat{p}_i \) and consequently, the approximated version of (1.1) becomes

\[
\int \hat{p}(u) \phi(u; 0, D) du \\
\propto p(u^\ast) |D|^{-1/2} \int \exp\left\{ -\frac{1}{2}(u - u^\ast)^T K^\ast (u - u^\ast) - \frac{1}{2} u^T D^{-1} u \right\} du \\
= p(u^\ast) |D|^{-1/2} \int \exp\left\{ -\frac{1}{2} u^T (K^\ast + D^{-1}) u + u^T K^\ast u \right\} du
\]

where \(|A|\) denotes the determinant of matrix \( A \). In order to make the integrand a Gaussian kernel, let \( A = I + K^\ast^{-1/2} D^{-1} K^\ast^{-1/2} \) and rewrite the exponent as

\[-\frac{1}{2} u^T K^\ast^{-1/2} A K^\ast^{-1/2} u + u^T K^\ast^{-1/2} A^{1/2} A^{-1/2} K^\ast^{-1/2} u^\ast \]

and thus

\[
\int \hat{p}(u) \phi(u; 0, D) du \\
\propto p(u^\ast) |D|^{-1/2} |D^{-1} + K^\ast|^{-1/2} \exp\left\{ -\frac{1}{2} u^\ast^T (D + K^\ast^{-1})^{-1} u^\ast \right\}.
\]

This procedure is also called the Laplace approximation of order two.

Although (2.5) is elegant, it can be expected that when the second-order Taylor expansion performs badly, the approximation fails to mimic (1.1) closely. From another viewpoint, \( \hat{p}_i(u) \) is proportional to the normal distribution with mean \( u^\ast \) and variance \( K^\ast^{-1} \). Thus, the condition that PQL performs well is that \( p_i(u) \) acts like a normal. Unfortunately, binomial (logistic regression) as well as Poisson (Poisson regression) usually violate this condition since their densities are skewed. Breslow
and Lin (1995) and Lin and Breslow (1996) proposed a fourth-order Laplace approximation in order to reduce biases. Raudenbush, et al. (2000) provided a general framework for higher-order Laplace approximation. They showed that the sixth-order Laplace approximation is very competitive to AGQ where the number of abscissa is chosen empirically.

2.3.2 Pseudo-Likelihood

In the GLMM literature, pseudo-likelihood (PL) method, Wolfinger and O’Connell (1993), is a popular alternative to PQL. We quote Carroll and Ruppert (1988) to explain the PL idea:

Pseudo-likelihood estimates \( \theta \) are based on pretending that the regression parameter \( \beta \) is known and equal to the current estimate \( \hat{\beta} \), and then estimating \( \theta \) by maximum likelihood assuming normality.

Wolfinger and O’Connell (1993) defined the pseudo-response as follows. Let \( E(Y|u) = \mu = \eta^{-1}(\nu) \) where \( \nu_{n \times 1} = X\beta + Zu \). Note that \( \eta^{-1}(\cdot) \) denotes the inverse function of \( \eta(\cdot) \), not the reciprocal. With relevant initial values \( (\hat{\beta}, \hat{u}) \) we further define \( \hat{\nu} = X\hat{\beta} + Z\hat{u} \). The first order Taylor expansion of \( \eta^{-1}(\nu) \) with respect to \( \hat{\nu} \) yields

\[
\eta^{-1}(\nu) \approx \eta^{-1}(\hat{\nu}) + \hat{\mathbf{B}}X(\beta - \hat{\beta}) + \hat{\mathbf{B}}Z(u - \hat{u}) \tag{2.6}
\]

where \( \hat{\mathbf{B}} = \partial\eta^{-1}/\partial\nu|_{\nu=\hat{\nu}} \) is an \( n \times n \) diagonal matrix. Rearranging (2.6) yields

\[
\hat{\mathbf{B}}^{-1}(\mu - \eta^{-1}(\hat{\nu})) + X\hat{\beta} + Z\hat{u} \approx X\beta + Zu
\]

where the left-hand side of above equation is the (conditional) expectation of

\[
P = \hat{\mathbf{B}}^{-1}(Y - \eta^{-1}(\hat{\nu})) + X\hat{\beta} + Z\hat{u}.
\]

Thus, the pseudo-response \( P \) approximates the conditional mean \( X\beta + Zu \) and variance \( \hat{\mathbf{B}}^{-1}\mathbf{A}\hat{\mathbf{B}}^{-T} \) where \( A_{n \times n} = \text{Var}(Y|u) \). Assuming both \( P \) and \( u \) follow normal distributions with certain variance-covariance structures, we come back to the LMM framework.
2.3.3 Logistic-Normal Approximation

Denote $G$ and $g$ as the distribution and density function of the logistic distribution. Andrews and Mallows (1974) and Stefanski (1990) showed that

$$G(z) = \int_0^\infty \Phi(z/\sigma)q(\sigma)d\sigma$$

where $q(\sigma) = dL(\sigma/2)/d\sigma$ and

$$L(\sigma) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j+1} \exp\{-2j^2\sigma^2\}$$

is the Kolmogorov-Smirnov distribution. In words, a logistic distribution is a mixture of scaled normals. This motivates using a finite sum of weighted scaled-normal densities to approximate a logistic density. Monahan and Stefanski (1992) introduced this method to GLMM and measurement error model.

Similar to GQ, they approximated $G(z)$ by

$$G_k^*(z) = \sum_{j=1}^k w_j \Phi(s_j z), \ k = 1, 2, ...$$

where, for a fixed $k$, the pairs $(w_j, s_j)$ are chosen so that the error $\Delta_k = \sup_z |G_k^*(z) - G(z)|$ is minimized. Monahan and Stefanski (1992) provided a table for the pairs $(w_j, s_j)$ from $k = 1$ to $k = 7$ and also showed the corresponding errors. Consider the $i$th-observation marginal likelihood of logistic regression with normal random effect

$$L_i = \int_\mathbb{R} G(\mu_i)^{y_i}(1 - G(\mu_i))^{1-y_i}d\Phi(u; 0, D)$$

where $\mu_i = X_i\beta + Z_iu$. When $y_i = 1$ and replacing $G(\cdot)$ by $G^*(\cdot)$, we have

$$L_i^* = \int_\mathbb{R} \sum_{j=1}^k w_j \Phi(s_j \beta; 0, D) = \int_\mathbb{R} \sum_{j=1}^k w_j \Phi(X_i\beta; 0, s_j^{-2} + Z_iDZ_i^T)$$

where the second equality is due to Lemma 2.1 of Gupta et al. (2004). Consequently, the $k$th-order logistic-normal approximate marginal likelihood becomes

$$L_k^* = \prod_{i=1}^n \left[ \sum_{j=1}^k w_j \Phi(X_i\beta; 0, s_j^{-2} + Z_iDZ_i^T) \right]^{y_i} \left[ 1 - \sum_{j=1}^k w_j \Phi(X_i\beta; 0, s_j^{-2} + Z_iDZ_i^T) \right]^{1-y_i},$$

exactly the form of a generalized linear model.
2.4 GLMM: EM Algorithm

The EM algorithm of Dempster, Laird, and Rubin (1977) is a powerful tool to solve missing value problems. In the context of the GLMM, \( u \) is unobserved, i.e., missing, thus applying the EM algorithm is appropriate, see also Searle et al. (1991). In our notation, the EM algorithm can be described as follows. Define the complete data log-likelihood as

\[
l_C(\gamma, D) = \sum_{i=1}^{n} \log p(y_i|W_i, u; \gamma) + \log \phi(u; 0, D).
\]

Further, suppose \( \theta = (\gamma, \text{vec}(D))^T \) is the vector of unknown parameters and \( \theta^{(t)} \) denotes the current estimate. Then the EM algorithm is

- **Expectation (E-step):** calculate
  \[
  Q(\theta|\theta_t) = \int l_C(\theta) \times f_{U|Y}(u|y; \theta_t) du
  \]  
  where
  \[
  f_{U|Y}(u|y; \theta) = \frac{\phi(u; 0, D)}{\int \prod_{i=1}^{n} p(y_i; W_i, u, \gamma) \phi(u; 0, D) du}.
  \]

- **Maximization (M-step):** evaluate
  \[
  \theta_{t+1} = \arg \max_{\theta} Q(\theta|\theta_t)
  \]

- **Updating:** update \( \theta_t \) by \( \theta_{t+1} \) until certain stopping criteria are met.

The EM algorithm has many merits especially that it guarantees the improvement of the likelihood after each iteration. For regularity conditions for convergence, see Dempster et al. (1977) and Wu (1983).

In many cases, the EM algorithm is convenient when (2.7) and (2.8) have closed-form representations. However, this is not true in the GLMM case due to the integration in (2.7). As we mentioned previously, there are several MC methods to overcome this problem. Here, we would like to address the MC sample size issue in the EM
algorithm. Booth and Hobert (1999) pointed out that MC integration introduces extra error. This error decreases when MC sample size increases and conversely, increases when MC sample size decreases. From another viewpoint, the current value $\theta_t$ from MC or importance sampling is actually a random variable and thus, its variance suffices to be a measure of the error. Since the MC sample size directly affects the magnitude of the variance, researchers have proposed various ways to empirically adjust the MC sample size, e.g. McCulloch (1994, 1997) and Chan and Kuk (1997).

The solution proposed by Booth and Hobert (1999) is very straightforward. Suppose $\theta_{t+1}$ follows a normal distribution with mean $\theta^*_{t+1}$ and variance $D^*_{t+1}$. If the previous step solution $\theta_t$ is very close to $\theta^*_{t+1}$, say within its 75% confidence interval (C.I.), then the $(t + 1)$th step is meaningless since the previous solution is not significantly different from the current solution. A better way to make the update meaningful is to increase the MC sample size so that the variance is decreased. Booth and Hobert (1999) use the delta method to construct the variance estimator and extend it to the importance sampling case. With the help of this variance estimate, the updating procedure becomes

- Evaluate the variance estimate $\hat{\text{Var}}_m(\theta_{t+1})$ where the subscript $m$ denotes the current MC sample size.
- Check if $\theta_t$ locates within the 75% C.I. of normal distribution with mean $\theta_{t+1}$ and variance $\hat{\text{Var}}_m(\theta_{t+1})$. If yes, then increase sample size $m$ to $m'$ and then go back to E-step and recalculate $\theta_{t+1}$. If no, then keep the MC sample size as $m$ and calculate $\theta_{t+2}$.

Booth and Hobert (1999) suggested taking $m' = m + m/k$, for $k = 3, 4, 5$.

Another important issue in the EM algorithm is the stopping rule. Denote $\theta_t^{(j)}$ as the $j$th element of the vector $\theta_t$. Then, a popular stopping rule is

$$\max_j \left\{ \frac{|\theta_t^{(j)} - \theta_{t+1}^{(j)}|}{|\theta_t^{(j)}| + \delta_1} \right\}_{j=1,\ldots,d} < \delta_2$$

(2.9)
where $\delta_1$ and $\delta_2$ is known, e.g. $\delta_1 = 0.001$ and $\delta_2 = 0.0001$ (Searle et al., 1992). Note that $\delta_1$ is a small positive value so that the denominator can always be far away from zero. Booth and Hobert (1999) noticed that this criterion is good for GLMMs as long as there is no error in each step. Instead, they suggested using

$$
\max_j \left\{ \frac{|\theta_{t+1}^{(j)} - \theta_t^{(j)}|}{\sqrt{\text{Var} \left( \theta_t^{(j+1)} + \delta_1 \right)}} \right\}_{j=1,...,d} < \delta_2
$$

(2.10)

so that the error is taken into account. Since $\theta^{(t)}$ always contains certain error, Booth and Hobert (1999) suggested $\delta_2 = 0.02$; empirically, this approach performs very well. The ground-breaking idea in their paper is that given the data $(y, W)$ the EM solution $\theta_t$ aided by MC is a random variable rather than a fixed value. One can and should take advantage of this point to improve the algorithm analogous to the EM problem with deterministic objective function.

Last, we comment on solving GLMM problem using EM algorithm. We quote words from Boos and Stefanski (2007):

The EM Algorithm is often useful when the complete data likelihood has the form of an exponential family....the M step is straightforward and basically inherited from the exponential family, but the E step is often challenging and not necessarily aided by the exponential family...

It is worth emphasizing that in GLMM problems, except conjugate cases, both E and M steps are challenging. The benefit from using the EM algorithm is thus limited. It turns out that if we have a good initial value $\theta_0$ then applying a MC Newton-Raphson search (Chapter 7, McCulloch, 2003) on the log-likelihood provides a faster convergence rate.
2.5 GLMM: Semiparametric, Nonparametric, and Bayesian

Beyond those already discussed, there are still other ways to pursue a good solution to the GLMM problem. In this section, we briefly describe nonparametric methods, semi-parametric methods, and some recent-developed Bayesian methods. For brevity, all methods will be explained in the one-dimensional case, although they can be easily extended to the multivariate case.

Nonparametric methods focus on the freedom of assigning the distribution of random effect. Analogous to (1.1), we define the marginal likelihood as

\[ \int p_i(y_i|W_i, u; \gamma) dG(u; \tau) \]

where \( G \) and \( g \) denote the distribution and density function of random variable \( u \) with parameter vector \( \tau \). For the semi-nonparametric (SNP) model, Zhang and Davidian (2001) defined \( g(\cdot) \) as a normal density multiplied a polynomial and divided by a norming constant. For example, a second-order SNP model can be written as

\[ g_{SNP}(u) = (\alpha_0 + \alpha_1 u + \alpha_2 u^2)^2 \phi(u; \mu, \sigma^2)/C \]

(2.11)

where \( C \) is the norming constant. A more general class is assuming \( G(\cdot) \) as a mixture of normals, i.e.,

\[ g_{NP}(u) = \int \phi(u; \mu, \sigma^2) dH(\mu, \sigma^2) \]

(2.12)

for some distribution function \( H \in \mathcal{H} \) where \( \mathcal{H} \) is the collection of all possible distribution functions of \((\mu, \sigma^2)\). If the nonparametric maximum likelihood estimate is applied, the solution exists but \( G \) must be a step function (Lindsay, 1983). Further, Magder and Zeger (1996) defined another space \( \mathcal{H}_S \subset \mathcal{H} \) so that the estimated \( G_0 \) \((\in \mathcal{H}_S)\) yielded from (2.12) is smooth.

The semiparametric model is also known as the generalized estimating equation (GEE) method due to Liang and Zeger (1986). Unlike the GLMM model focusing
on subject-specific level, GEE method concentrates on the population-average level. Therefore, the GEE may be inefficient to estimate the variance components (McCulloch, 2003). GEE begins by assigning a marginal model for the response. The only assumption on \( u \) is that it has zero mean. Thus, the population average \( \text{E}(Y_i|W_i) \) is only the function of \( \beta \) but not of \( u \). For example, in logistic regression only specify

\[
\text{E}(Y_i|W_i) = p_i \quad \text{and} \quad \text{logit}(p_i) = X_i^T \beta.
\]

where \( Y_i \) is an \( r \)-dimensional vector with correlated element \( Y_{i1}, \ldots, Y_{ir} \). The correlation is caused by sharing random effects. Suppose that the correlation structure of \( Y_i \) is denoted as \( V(y_i) \). Then, by theory, the estimating equation is

\[
\sum_i \frac{\partial p_i}{\partial \beta^T} V^{-1}(y_i)(y_i - p_i).
\]

This approach is extended to GEE2 for estimating the variance components. For details, we refer to Prentice and Zhao (1991), Qu, et al. (2000) and references therein.

Concerning the frequentist GLMM model like (1.1), a Bayesian model relies on specifying relevant priors to \( \gamma \) and \( D \). A conjugate choice of \( D \) is the Wishart distribution. As Natarajan and McCulloch (1998) pointed out it is difficult to choose the hyper parameters for the Wishart and with some poor choices, the convergence rate of Gibbs sampling is affected. On the other extreme, if there is no prior information on \( \theta \) an intuitive choice is uniform for regression coefficients and Jeffreys prior for \( D \) (Tiao and Tan, 1965; Zeger and Karim, 1991). However, this yields an improper posterior (Natarajan and Kass, 2000). As a result, many alternatives between fully noninformative and fully subjective priors are introduced such as conditional-conjugate prior (Box and Tiao, 1973), reference prior (Natarajan and Kass, 2000), diffuse prior (Browne and Draper, 2006), and weakly-informative prior (Gelman, 2006). Other than picking a good prior, there are still many problems in the Bayesian GLMM: computational issues, centering models, or even applying frequentist large sample theory to prevent computational burden (see Su and Johnson, 2006.) Finally, we re-
fer to Conqdon (2005) for a more complete survey of Bayesian inference on categorical data.
Chapter 3

A Review of Stochastic Approximation

Estimation methods are often defined as the result of an optimization problem. Least squares estimation arises from minimizing the sum of squared errors. Maximum likelihood (ML) is another: Parameters are chosen so that the corresponding likelihood is maximized. When the objective function, that is, the function that one want to optimize, is smooth, basic calculus theory applies: If the function $f : \mathbb{R}^d \to \mathbb{R}$ has a maximum or minimum then there exists a vector $\theta^* \in \mathbb{R}^d$ such that

$$\frac{df(\theta)}{d\theta}_{\theta=\theta^*} = g(\theta^*) = 0$$

(3.1)

That is, if the objective function is well behaved so that the optimal point is unique, i.e. the objective function is strictly convex or concave in $\mathbb{R}^d$, then (3.1) is helpful to find the optimal point and, hence, the optimization problem is converted to a root-finding problem.

A popular, basic, and easy to implement optimization algorithm is Newton-Raphson (NR) which approximates the objective function with a quadratic function. The update step of NR either for maximizing/minimizing $f(\theta)$ or for finding a root
of $g(\theta)$ has the form:

$$\theta_{t+1} = \theta_t - \left[\nabla^2 f(\theta_t)\right]^{-1} \nabla f(\theta_t) = \theta_t - [J_g(\theta_t)]^{-1} g(\theta_t).$$

(3.2)

Here $\nabla f$ is the gradient, $\nabla^2 f$ is the Hessian matrix, and $J_g = \partial g(\theta)/\partial \theta$. It is well-known that when $f(\cdot)$ is quadratic the solution can be reached in one step from any initial value $\theta_0$; otherwise, a good initial value is crucial for convergence. Many other methods have similar form of (3.2), e.g. steepest descent, quasi-Newton, and others (see Monahan, 2001, and Spall, 2003, for more insightful discussion.) In multivariate case, $g$ is explained to be the best direction toward the optimal and $J_g$ rotates and scales the direction so that the move $[J_g(\theta_t)]^{-1} g(\theta_t)$ is optimized. However, the computation of $[J_g(\theta_t)]^{-1}$ is often costly so many other remedies have been proposed in literature. A common belief is that if the matrix $[J_g(\theta_t)]^{-1}$ is not exact then more iterations are required.

Above, we dealt with a deterministic function $f$ and its derivatives $g$ and $J_g$. However, in some cases, we can not evaluate them analytically but only can get unbiased evaluations, e.g. the MC and importance sampling integration are unbiased to its deterministic integral. Under this circumstance, Robins and Monro (1951) proposed the stochastic approximation (SA) to find the root of $g(\theta)$. Later, Kiefer and Wolfowitz (1952) defined the optimization version of SA; we called it KWSA. While both SA and KWSA originally designed for one-dimensional problems, Blum (1954) extended them to multi-dimensional problems. Recently, Spall (1992) proposed simultaneous perturbation SA (SPSA) which reduces computational burden dramatically and provides some elegant large sample properties. Next, we are going to introduce these methods followed by a section about implementing SPSA on GLMM.

### 3.1 SA – Root Finding

The milestone paper Robins and Monro (1951) considers the following problem. They targeted the root of the function $g : \mathbb{R} \rightarrow \mathbb{R}$ but $g$ can not be observed without
error; only \( y(\theta) \), a contaminated version of \( g(\theta) \), is available, for every \( \theta \).

More formally, the objective function for SA has the form

\[
g(\theta) = E(Y(\theta)) = \int ydH(y|\theta)
\]

where \( H(\cdot|\theta) : [-C, C] \rightarrow [0, 1] \) is a valid distribution function for any \( \theta \) and for some \( C \in \mathbb{R}^+ \). Moreover, assume that the derivative of \( g \) around the true value \( \theta \) is positive.

The proposed algorithm is

\[
\theta_{t+1} = \theta_t - s_t y_t \tag{3.3}
\]

where \( y_t = y(\theta_t) \) for shorthand and \( s_t \rightarrow 0, \sum_{t=1}^\infty s_t = \infty \) and \( \sum_{t=1}^\infty s_t^2 < \infty \). If the derivative is negative then we change the sign in (3.3) from “−” to “+.” Note that \( s_t \), analogous to \( J_g(\theta_t) \) in (3.2), controls the gain/step size of each iteration. Conditions on \( s_t \) determine the convergence of \( \theta_t \). The sequence \( s_t \) can not decrease too fast, otherwise the algorithm cannot search the whole support of \( \theta \). At the same time, it can not be too slow or the algorithm never converges. Robins and Monro (1951) also showed that under conditions mentioned above, \( \theta_t \) converges to the true value \( \theta_{True} \) in probability. In the case of a linear response function \( g(\theta) \) which corresponds to simple linear regression, Chung (1954) proved that the optimal gain sequence follows

\[
s_t = \frac{s_0}{t_0 + t} \times \frac{1}{b}
\]

where \( s_0 \) and \( t_0 \) are two positive numbers and \( b \) is the slope of regression line.

The motivating example in Robins and Monro (1951) paper is very helpful for understanding the usefulness of SA algorithm. In this paragraph, we restate their example in our words. Suppose people are interested in simple linear regression problem

\[
M(x) = \beta_0 + \beta_1 x
\]

where \( M(x) \) denotes the mean response of given \( x \). Rather than knowing \( \beta_0 \) and \( \beta_1 \), interest may lie in finding \( x_\alpha \) such that \( M(x_\alpha) = \alpha \). Note that, for a particular \( x_i \), one can not observe the mean \( M(x_i) \) but can observe a single realization \( w(x_i) = \beta_0 + \)
\[ \beta_1 x_i + e_i \] where \( e_i \) is not observable and has zero mean. Then, the SA algorithm with the objective function \( g(x) = M(x) - \alpha \) and its contaminated version \( y(x) = w(x) - \alpha \) provides a consistent estimate of \( x_\alpha \) if all conditions are met. It is worth emphasizing that this method is nonparametric since we only require the existence of a valid \( H(\cdot | \theta) \) but do not specify a particular form for it.

### 3.2 KWSA – Optimization

It is more common to have an optimization problem than to have a root finding problem in statistical modeling. Kiefer and Wolfowitz (1952) designed the optimization version of SA denoted as KWSA. The objective function is

\[ f(\theta) = \int y dH(y|\theta) \]

where \( H(\cdot | \theta) : \mathbb{R} \to [0, 1] \) is a valid distribution function for any \( \theta \). Moreover, if \( f(\theta) \) is Lipschitz and

\[ \int_{-\infty}^{\infty} (y - f(\theta))^2 dH(y|\theta) < \infty, \]

the algorithm following the iteration step

\[ \theta_{t+1} = \theta_t - s_t \frac{y_t^+ - y_t^-}{2c_t} = \theta_t - s_t \psi_t \]

with \( y_t^+ \sim H(y|\theta_t + c_t) \) and \( y_t^- \sim H(y|\theta_t - c_t) \) guarantees that \( \theta_t \) converges to \( \theta_{\text{True}} \) in probability as long as following conditions are satisfied: 1) \( s_t, c_t > 0 \) and \( t \to 0 \); 2) \( \sum_{t=1}^{\infty} s_t = \infty \); 3) \( \sum_{t=1}^{\infty} s_t c_t < \infty \); and 4) \( \sum_{t=1}^{\infty} s_t^2 c_t^{-2} < \infty \).

The strategy of KWSA is using \( \psi_t = (y_t^+ - y_t^-)/2c_t \) to approximate \( df(\theta_t)/d\theta_t \). Obviously, if \( \psi_t \) is unbiased to \( df(\theta_t)/d\theta_t \) then using SA is sufficient. Since \( \psi_t \) is not necessarily to be unbiased the Lipschitz condition is required to ensure the convergence. The multi-dimensional versions of SA and KWSA were introduced by Blum (1954). Although more sophisticated assumptions were made in his paper, the implementation is straightforward. Let \( f : \mathbb{R}^d \to \mathbb{R} \) and \( e_j \) denotes a \( d \)-dimensional zero
vector except the $j$th element of it equals to 1. Then, the update formula becomes

$$\theta_{t+1} = \theta_t - s_t \psi_t$$  \hspace{1cm} (3.4)$$

where the $s_t$ is either a scalar or a $d \times d$ matrix, $\theta_t$'s and $\psi_t$'s are $d$-dimensional vector and the $j$th element of $\psi_t$ is

$$\psi_{t(j)} = \frac{y_{t(j)}^+ - y_{t(j)}^-}{2c_t}$$  \hspace{1cm} (3.5)$$

where $y_{t(j)}^+$ $\sim$ $H(y|\theta_t + e_j c_t)$ and $y_{t(j)}^-$ $\sim$ $H(y|\theta_t - e_j c_t)$. Note that if the evaluation of $y_{t(j)}^+$'s and $y_{t(j)}^-$'s are computationally costly then this method is time consuming since it requires $2d$ evaluations for each step. Use of an one-sided differences

$$\frac{y_{t(j)}^+ - y_{t}^0}{c_t}$$

where $y_{t}^0 \sim H(y|\theta_t)$ cuts the computational cost in half, but does not remove its dependence on the dimension $d$.

### 3.3 Simultaneous Perturbation SA

Spall (1992) developed SPSA so that the computation of (3.5) is less demanding. The updating formula is exactly the same as (3.4) but replacing $\psi_t$ by

$$\psi_t^{SP} = \Delta_t \frac{y(\theta_t + \Delta t c_t) - y(\theta_t - \Delta t c_t)}{2c_t ||\Delta_t||}$$

where $||\cdot||$ denotes the $L_2$ norm of its argument, $y(\cdot) \sim H(y|\cdot)$ and $\Delta_t = (\Delta_{t1}, ..., \Delta_{td})^T$, $\Delta_{tj}$'s are $i.i.d.$ random variable which has equal probabilities to be 1 or -1. Actually, the choice of the distribution of $\Delta_{tj}$ can be very flexible but requires: $E(\Delta_{tj}) = 0$, $E(\Delta_{tj}^2) < \infty$, $E(\Delta_{tj}/\Delta_{tj'}) < \infty$ for any $j \neq j'$. For instance, $\Delta_{tj}$’s can be also be uniformly sampled from a unit $d$-dimensional sphere. In addition, conditions on $\{s_t\}$ and $\{c_t\}$ are identical to KWSA’s.
An attractive property of SPSA is its asymptotic normality, see Spall (1992). With certain regularity conditions,

\[ s_t = \frac{s}{t^\alpha} \quad \text{and} \quad c_t = \frac{c}{t^r} \]

and \( \beta = \alpha - 2r > 3, \ 3r - \alpha/2 > 0 \) where \( s, c, \alpha \) and \( r \) are some positive numbers, then, for sufficiently large \( t \),

\[ t^{\beta/2}(\theta - \theta_{True}) \overset{D}{\longrightarrow} N(0, \omega \mathbf{H}^{-1}(\theta_{True})) \]

where \( \mathbf{H}(\theta_{True}) \) is the Hessian matrix of the objective function and \( \omega \) is a known constant dictated by the rate parameter.

Investigating multivariate KWSA, one can find that the vector \( \psi_t \) is the numerical gradient of \( f(\theta_t) \). In other words, for each \( j \), \( \psi_{tj} \) is an approximation of the partial derivative. Similarly, \( \psi^{SP}_t \) is the numerical directional derivative of \( f(\theta_t) \) on direction \( \Delta_t \). SPSA attempts to reduce the computational burden by replacing the exact derivative by random directional derivatives. However, this approach may require more steps to reach the convergence. On the other hand, \( \psi^{SP}_t \) only evaluates \( y(\cdot) \) twice for each iteration whereas \( \psi_t \) does it \( 2d \) times. Therefore, SPSA is preferred as long as the saving from evaluating fewer \( y(\cdot) \)’s is greater than the cost of extra iterations induced due to randomly chosen directions.

### 3.4 Implementation of SPSA on GLMM

Consider the GLMM problem in a smaller class: logistic regression with linear mixed effects. Suppose we have \( n \) observations \((Y_i, X_i, Z_i), \ i = 1, ..., n\), where \( X_i \) is a \( p \times 1 \) row vector and \( Z_i \) is a \( q \times 1 \) row vector. Further, let \( \text{logit}(p_i) = X_i^T \beta + Z_i \mathbf{u} \) where \( \mathbf{u} \sim N(0, \mathbf{D}_{q \times q}) \) and \( Y_i | \mathbf{u} \sim \text{Bernoulli}(p_i) \). This leads to a marginal likelihood

\[
L = \prod_{i=1}^{n} \left\{ \int_{\mathbb{R}^q} p_i^{y_i} (1 - p_i)^{1-y_i} \phi(\mathbf{u}; 0, \mathbf{D}) d\mathbf{u} \right\}.
\]
The integration is not easy but an unbiased estimate is straightforward

\[ L_n = \prod_{i=1}^{n} \left\{ \frac{1}{N_{IS}} \sum_{j=1}^{N_{IS}} p_{ji}^{y_i} (1 - p_{ji})^{1-y_i} \right\} \]  

where \( \text{logit}(p_{ji}) = X_i^T \beta + Z_i u_j \) and \( u_j \sim N(0, D) \), \( j = 1, ..., N_{IS} \). Notice we need that,

\[ EL_n = \mathbb{E} \left\{ \prod_{i=1}^{n} \left[ \frac{1}{N_{IS}} \sum_{j=1}^{N_{IS}} p_{ji}^{y_i} (1 - p_{ji})^{1-y_i} \right] \right\} = \prod_{i=1}^{n} \left\{ \frac{1}{N_{IS}} \sum_{j=1}^{N_{IS}} \mathbb{E}_{u_j} [p_{ji}^{y_i} (1 - p_{ji})^{1-y_i}] \right\} \]

\[ = \prod_{i=1}^{n} \left\{ \mathbb{E}_{u_i} [p_{1i}^{y_i} (1 - p_{1i})^{1-y_i}] \right\} = \prod_{i=1}^{n} \left\{ \int_{\mathbb{R}^q} p_{1i}^{y_i} (1 - p_{1i})^{1-y_i} \phi(u; 0, D) du \right\} = L, \]

\( i.e. \) \( L_n \) is unbiased for \( L \). In this sequel, we can say that \( L_n \) is a contaminated version of \( L \) and hence, SA methods are relevant. Note that \( L_n \) is the likelihood but not log-likelihood. Since \( L_n \) may become extremely large or small, scaling will be needed to avoid underflow or overflow.

As mentioned in section 7.5 of Spall (2003), the SPSA algorithm is as follows:

1. Begin with starting value \( \theta_0 \) and sequences

   \[ s_t = \frac{s}{(t + 1 + A)^a} \quad \text{and} \quad c_t = \frac{c}{(t + 1)^r} \]

   where \( a, A, \alpha, c \) and \( r \) are all positive and \( \alpha - 2r > 0 \) and \( 3r - \alpha/2 > 0 \).

2. Generate the random perturbation vector \( \Delta_t \) where each component is either 1 or -1 with equal probability. Evaluate \( L_n(\theta_t + c_t \Delta_t), L_n(\theta_t - c_t \Delta_t) \) and hence

   \[ \psi_t^{SP} = \Delta_t \left[ \frac{L_n(\theta_t + c_t \Delta_t) - L_n(\theta_t - c_t \Delta_t)}{2c_t ||\Delta_t||} \right] \]  

3. Update \( \theta_t \) by \( \theta_{t+1} = \theta_t - s_t \psi_t^{SP} \).

4. Return to step 2 until convergence criteria are met.

Issues about implementation are addressed in following subsections: choosing differencing sequence \( c_t \), pairing, parameter scaling, and importance sampling.
3.4.1 Choosing Differencing Sequence

The basic theory of calculus defines the derivative of a function \( f \) as

\[
\lim_{c \to 0} \frac{f(x + c) - f(x)}{c}
\]

which may suggest in SPSA that we should set \( c_t \) as small as possible. However, computationally, (3.7) may fail if \( c_t \) is too small, due to error in the numerator.

To be convenient, suppose we are interesting in the finite difference (the numerical derivative) of function \( f(x) : \mathbb{R} \to \mathbb{R} \). The logic of choosing \( c_t \) can be accessed as follows. Let \( f(\cdot) \) denote the true function we want to evaluate. Due to the limitation of computer, it can be recorded as \( f^*(\cdot) \). Thus, the numerical derivative can be expressed as

\[
\frac{f^*(x + h) - f^*(x)}{h} = \frac{f(x + h) + \epsilon_{x+h} - f(x) - \epsilon_x}{h} \approx f'(x) + \frac{h}{2} f''(x) + O(U) \tag{3.8}
\]

where \( f^*(x + h) = f(x + h) + \epsilon_{x+h} \) and \( f^*(x) = f(x) + \epsilon_x \). Note that \( \epsilon_{x+h} - \epsilon_x \) is \( O(U) \) and \( U \) denotes the machine unit. Replacing the numerator on the second line of (3.8) by its second-order Taylor expansion at \( h = 0 \) yields

\[
\frac{f^*(x + h) - f^*(x)}{h} \approx f'(x) + \frac{h}{2} f''(x) + \frac{O(U)}{h}.
\]

In order to minimize the bias (the last two terms above), we have

\[
h = \sqrt{\frac{2}{f''(x)}} O(U) = O(U^{1/2})
\]

providing \( f''(x) \) is finite and far from zero. Following a similar analysis, Gill, Murray and Wright (1981) suggest \( h = U^{1/3} \) for centered difference,

\[
\nabla f(x) \approx \frac{f(x + h) - f(x - h)}{2h}.
\]
3.4.2 Pairing

The central difference form of (3.7) suggests the use of pairing to reduce the variance in $\psi_{t}^{SP}$. Consider the relationship

$$g(\theta; h) = \int_{\mathbb{R}} \frac{y(x; \theta + h) - y(x; \theta - h)}{2h} dH(x) \rightarrow \int_{\mathbb{R}} \frac{\partial}{\partial \theta} y(x; \theta) dH(x)$$

when $h \rightarrow 0$, assuming the limit and the integration are exchangeable. Suppose $X_1$ and $X_2$ are two independent random variable with common distribution function $H(x)$ then the left-hand side of (3.9) can be estimated by either the pairing numerical derivative

$$\hat{g}_P(h) = \frac{y(X_1; \theta + h) - y(X_1; \theta - h)}{2c}$$

or the regular numerical derivative

$$\hat{g}_R(h) = \frac{y(X_1; \theta + h) - y(X_2; \theta - h)}{2h}.$$  

Note that $E\hat{g}_P(h) = E\hat{g}_R(h) = g(\theta; h)$. Conceptually, both (3.10) and (3.11) yield a valid estimation of the right-hand side of (3.9). However, when considering their variance we have

$$\text{Var}(\hat{g}_P(h)) \approx \text{Var}\left(\frac{\partial}{\partial \theta} y(X; \theta)\right)$$

whereas

$$\text{Var}(\hat{g}_R(h)) = \frac{1}{4h^2} \{\text{Var}(y(X; \theta + h)) + \text{Var}(y(X; \theta - h))\} \approx \frac{1}{2h^2} \text{Var}(y(X; \theta))$$

where $X \sim H(x)$. Notice that $H$ should be free of $\theta$ so that pairing is possible.

Recall that one of the convergence conditions of SPSA is $c_t \rightarrow 0$ but, in Section 3.4.1, we claim that $c_t$ should be $O(U^{1/2})$ or $O(U^{1/3})$. According to their variance formulae, we know that when $c_t$ becomes very small, $\hat{g}_P(h)$ is superior to $\hat{g}_R(h)$. This motivates us taking $c_t$ fixed and as small as $O(U^{1/2})$ or $O(U^{1/3})$. 
3.4.3 Parameter Scaling

Unlike least squares or Gauss-Newton, stochastic approximation algorithms do not scale the parameter vector automatically. Consider the following example. Suppose $\theta_t = (0.01, 100)^T$, $\Delta_t = (+1, +1)^T$ and $\psi_t^{SP} = \Delta_t \times 1$ then the update formula of SPSA is

$$
\theta_{t+1} = \begin{pmatrix} 0.01 \\ 100 \end{pmatrix} + s_t \times \begin{pmatrix} +1 \\ +1 \end{pmatrix}.
$$

If $s_t = 1$ then, from $\theta_t$ to $\theta_{t+1}$, the first component of $\theta$ moves too large a step. If set $s_t = 0.00001$ to compromise the first component of $\theta$ then the second component barely moves after many iterations. Multiplying a scaling matrix, say $S$, in front of $\Delta_t$ may help. Continue above example. Let $S = diag\{0.001, 100\}$ then the new update formula becomes

$$
\theta_{t+1} = \begin{pmatrix} 0.01 \\ 100 \end{pmatrix} + s_t \times S \Delta_t = \begin{pmatrix} 0.01 \\ 100 \end{pmatrix} + s_t \times \begin{pmatrix} 0.01 \\ 100 \end{pmatrix}
$$

which moves each component of $\theta$ evenly toward the solution with a relevant $s_t$. In practice, if a good initial value, say $\theta_0$, is tenable, then we suggest using $S = diag\{\theta_0\}$.

3.4.4 Importance Distribution

Consider again the likelihood expression given by (3.6). There are two issues in the normal density: 1) the density contains unknown parameters $D$ and 2) centering at $\mathbf{0}$ may not be the best choice. These problems can be improved by using importance sampling from a normal distribution centered at a good prediction $\hat{\mathbf{u}}$ and scaled by Hessian $\hat{H}$, that is, we can rewrite (3.6) as

$$
L_n^{IS} = \prod_{i=1}^n \left\{ \frac{1}{N_{IS}} \sum_{j=1}^{N_{SN}} p_{ji}^y (1 - p_{ji})^{1-y_i} \frac{\phi(u_j; 0, \Sigma)}{\phi(u_j; \hat{\mathbf{u}}, \hat{H}^{-1})} \right\}
$$

where $u_j$'s are i.i.d. samples from $N(\hat{\mathbf{u}}, \hat{H}^{-1})$ and remembering that $p_{ji}$ is a function of $u_j$. The question is how much effort should we pay for computing $\hat{\mathbf{u}}$ and $\hat{H}$. Note
that both of them are functions of $\theta_t$ so it can be costly if we update them in each step of SPSA.

The importance sampling sample size $N_{IS}$ has been shown to be a mild factor in the performance of SPSA algorithm, Wang, J (2005). The reason is that the theory works as long as $L_n$ is unbiased to $L$. In fact, $L_n$ is always unbiased to $L$ even through $N_{IS}$ may be as small as 1. The tradeoff of small $N_{IS}$ is the increase in the variance of the response $y$ which will reflect on the accuracy of SPSA estimate.

### 3.5 GLMM: using SA’s

In the literature, there are a couple of papers using SA as a solution to GLMM. First, they consider EM’s $Q$ function (2.7) as the objective function. Since this $Q$ function has no closed form, some numerical or sampling method is required. So, second, some MC techniques are applied, e.g. MCMC and importance sampling, in order to provide a contaminated objective function. Finally, since the objective function is contaminated and is unbiased to the original objective function SA methods may be used.

Zhu and Lee (2002) proposed SA-MCMC method to solve GLMM problem. The main idea is using a MCMC procedure to approximate the integral in (2.7). As we can see from the next equation of (2.7), the conditional distribution $f_{U|Y}$ has no closed form, too. Since MC integration is not available, MCMC becomes an option. Analogous to Zhu and Lee (2002), Jank (2006) proposed an importance sampling method for the integration. They use different approximations as objective functions and rely SA for optimization.

A difficult issue in SA is the stopping criterion. As Jank (2006) showed, a stopping criterion like (2.9) is bad since it may cause premature stopping due to sharply decreasing $s_t$’s. A stopping criterion like (2.10) is preferred. The question is how to get the variance estimate. Fortunately, EM algorithm theory provides a ready-to-use
variance-covariance formula. Zhu and Lee (2002) took advantage of this point and adopted (2.10) as the stopping criterion. Similarly, Jank (2006) derived the approximated variance-covariance matrix and bias so that the convergence of the sequence of SA updates can be visualized and hence monitored.

We close this section with a short comment on EM based SA algorithm. As we commented at the end of section 2.4, there is no real benefit on using EM since EM is designed for deterministic objective function and the result of its M step has a closed form. As we can see from Section 2.4, neither of them are satisfied in GLMM problem. We suggest directly applying SA on marginal likelihood function with MC or importance sampling.
Chapter 4

Stationary Simultaneous Perturbation Stochastic Approximation

In this chapter, we define and provide theoretical support for the stationary simultaneous perturbation stochastic approximation algorithm. The algorithm is applied to the objective function which is quadratic around its maximum. We prove that under some mild conditions the proposed algorithm yields a strictly stationary sequence. This motivates us treating the optimization problem as an estimation problem. An estimator of the optimum and its variance are suggested in this chapter. We also provide conditions for the SSPSA sequence to be ergodic. In order to make the proofs compact, some crucial matrix properties (but not usually seen in basic linear algebra texts) are extracted and summarized in Appendix A including matrix norm, von Neuman series, and Kronecker product.
4.1 Motivating Example

The toy example proposed by Jank and Booth (2003) is well describes the behavior of SPSA. Define

\[ y_i = \mu + u_i + \epsilon_i \]

with \( u_i \sim N(0, 1) \) and \( \epsilon_i \sim N(0, 1) \) for \( i = 1, \ldots, n \) and jointly independent. The corresponding MLE of the only parameter \( \mu \) is \( \bar{y} \), the sample mean. To be convenient, let \( n = 2 \) and \((y_1, y_2) = (1.5, 0.5)\). Therefore, the MLE of \( \mu \) is 1. In addition, as mentioned in Chapter 3, the step size can be \( s_t = 1/t^r \). Choosing \( r = 0, 0.51, 1 \) and 2, Figure 4.1 shows the path of these 4 SPSA sequences. Although \( r = 0 \) does not satisfy the convergence condition (neither does \( r = 2 \)), from Figure 4.1, we have the clue that the sequence may move around the solution when the step size is a constant. Clearly, as the decaying rate \( r \) is large, the path is smooth and moves slowly toward the answer. On the contrary, if the rate is small, the path is wiggly and moves quickly. Even when the step size is a fixed number \( r = 0 \) the path lingers around (but does not converge to) the answer after some “burn-in time”. This fact reflects the importance of choosing a good step size. If the step size decays too fast, say \( s_t = 1/t^2 \), the sequence barely moves after \( t = 100 \) and causes “premature” convergence. If the step size decays too slowly, the sequence converges slowly. Currently, trial-and-error may be the most reliable way to figure out a good step size. This motivates us choose a constant step size. Although it does not lead to a convergence sequence if we can show the stationarity of this sequence then we can estimate the solution. Following the idea of Liddle and Monahan (1988), we replace the \( s_t \) in SPSA by a constant \( s \) and call it stationary SPSA (SSPSA). The rest of this chapter shows properties of the stationarity of SSPSA sequence.
4.2 SSPSA with Quadratic Objective Function

It is reasonable to assume that the objective function is quadratic around the true answer. This motivates the investigation of the properties of SSPSA using such an objective function. Let the objective function (which may be a likelihood function) be \( g(\theta) \) and its observable version as

\[
y(\theta) = g(\theta) + \epsilon = a + b^T \theta + \frac{1}{2} \theta^T B \theta + \epsilon
\]  

(4.1)

where \( \theta \in \mathbb{R}^d, a \in \mathbb{R}, b \in \mathbb{R}^d \) and \( B \in \mathbb{R}^{d \times d} \). In other words, \( y(\cdot) \) is observed with error or a contaminated version of \( g(\cdot) \). We further assume \( B \) is symmetric and positive definite so that \( g(\theta) \) is a convex function with respect to \( \theta \). The SSPSA
algorithm is defined as

$$\theta_{t+1} = \theta_t + s \times \Delta_t \times \frac{y(\theta_t + c\Delta_t) - y(\theta_t - c\Delta_t)}{2c||\Delta_t||}$$  \hfill (4.2)$$

where $s$ and $c$ are small positive numbers and $\Delta_t = (\Delta_{1t}, \ldots, \Delta_{dt})^T$ where $\Delta_i$'s are i.i.d. random direction with $E\Delta_t = 0$ and $E\Delta_t\Delta_t^T = \xi^2 I$ for some constant $\xi^2$. Note that $||\cdot||$ denotes the vector Euclidean-norm and we assume that $\Delta_t$ is chosen so that $||\Delta_t|| = \nu$ is a fixed number. Valid $\Delta = (\Delta_1, \ldots, \Delta_d)$ can be, for example, $\Delta$'s are i.i.d. Bernoulli taking value 1 or -1 with equal probability, $(\xi^2, \nu) = (1, \sqrt{d})$, or $\Delta$ is a random vector on the $d$-dimensional unit sphere, $(\xi^2, \nu) = (\xi^2_s, 1)$, see Appendix B. For $d \geq 2$, $\xi^2_s$'s are less than 1, see Figure B.1 for demonstration.

Define $\epsilon_t^+ = y(\theta_t + c\Delta_t) - g(\theta_t + c\Delta_t)$ and $\epsilon_t^- = y(\theta_t - c\Delta_t) - g(\theta_t - c\Delta_t)$ and

$$G_t = \Delta_t\Delta_t^T \times \frac{s}{\nu^2}.$$

We also assume that $\epsilon_t^+ - \epsilon_t^-$ has mean 0 and finite variance for all $t$ and, $\text{Var}(\epsilon_t^+ - \epsilon_t^-) = O(\epsilon_t^2)$. Further, denote $\epsilon_t^* = \Delta_t(\epsilon_t^+ - \epsilon_t^-)/2cv$ and $\tilde{\epsilon}_t = (\epsilon_t^+ - \epsilon_t^-)/2c$. After some algebra, we have

$$\theta_{t+1} = \theta_t - G_t\mathbf{b} - G_t\mathbf{B}\theta_t + \epsilon_t^* = (I - G_t\mathbf{B})\theta_t - G_t\mathbf{b} + \epsilon_t^*.$$  \hfill (4.3)$$

For a convergence sequence, we expect that the norm $||\theta_{t+1} - \theta_t||$ goes to zero at a certain rate. In addition, with the quadratic model (4.1), if $\theta_{\text{True}}$ represents the true solution then $\mathbf{b} + \mathbf{B}\theta_{\text{True}} = 0$ and

$$||\theta_{t+1} - \theta_t||^2 = (-G_t\mathbf{b} - G_t\mathbf{B}\theta_t + \epsilon_t^*)^T(-G_t\mathbf{b} - G_t\mathbf{B}\theta_t + \epsilon_t^*)$$

$$= \left[\bar{\epsilon}_t - \frac{s}{\nu^2} \Delta_t^T (\mathbf{b} + \mathbf{B}\theta_t)\right]^2.$$  \hfill (4.4)$$

Equation (4.4) says that even though $\theta_t$ is sufficiently close to $\theta_{\text{True}}$ the norm $||\theta_{t+1} - \theta_t||_2$ can be large by chance due to the randomness of $\bar{\epsilon}_t$. If the variance of $\bar{\epsilon}_t$ can be controlled (in GLMM, one can control this variation by tuning the importance sampling sample size) then the convergence relies on the second term in the square
brackets. In SSPSA case, it depends on the rate at which $b + B\theta_t$ converges to 0. On the other hand, when considering SPSA, replacing $s$ by $s_t$, the convergence depends on the rate of $s_t$ as well. Immediately, we realize that if $s_t$ goes to 0 in a faster rate than $b + B\theta_t$ then the square-norm shows the sign of convergence erroneously. Although these facts may only reflect the irrelevancy of using square-norm as the stopping criterion, equation (4.4) also reveals some clues of the drawbacks of SPSA.

4.3 The Stationarity of SSPSA

In this section, we show that the SSPSA algorithm yields a strictly stationary sequence. We begin with fixing the notation of eigenvalues of a symmetric matrix, say $A \in \mathbb{R}^{d \times d}$. Denote the eigenvalues of matrix $A$ as $\lambda_1^A \geq \cdots \lambda_d^A$ or $\lambda_1(A) \geq \cdots \lambda_d(A)$. If $A$ is symmetric it can be shown that all $\lambda_j(A)$'s are real.

Following the construction of Bougerol and Picard (1992), consider the random coefficient Autoregressive (RCA) model

$$X_{t+1} = A_tX_t + C_t, \ t \in \mathbb{Z},$$

(4.5)

where $X_t$ and $C_t$ are $d$-dimensional vectors and $A_t$ is a $d \times d$ matrix. In addition, $(A_t, C_t)$'s are i.i.d. random variables. Therefore, we have the following theorem

**Theorem 1.** Theorem 2.5 of Bougerol and Picard (1992)

Suppose that the generalized autoregressive model (4.5) with i.i.d. coefficients is irreducible and that both $E(\log^+ ||A_0||)$ and $E(\log^+ ||C_0||)$ are finite. Then (4.5) has a nonanticipative strictly stationary solution if and only if the top Lyapounov exponent (TLE)

$$\gamma = \inf_{t \in \mathbb{N}} \left\{ E \left( \frac{1}{t+1} \log ||A_tA_{t-1} \cdots A_0|| \right) \right\}$$

(4.6)

is strictly negative.

In the sense of “all norms are equivalent”, Theorem 2.4-5 (Kreyszig, 1989), the norm used in (4.6) is not restricted since the matrix $A_t$ is finite dimensional. For
equivalence, we mean that, for a real matrix $B$, there exists a finite real number $M$ such that $||B||_p/||B||_q \leq M$ where $p$ and $q$ indicates two different norms. However, not all norms have closed form representation, e.g. $p$-norm with $p = 3$, while some do, e.g. Frobenius norm and 2-norm among others, see Appendix A.1. With equivalence, if one norm converges to 0 then the rest converge to 0 as well. In the following, we use the Frobenius norm and the result can be extended to all other norms automatically.

Before relating TLE with the SSPSA algorithm, we describe a rather general result. Let $A_i$'s be i.i.d. random matrices and define $D_t = A_{t-1}A_{t-2} \cdots A_1A_0$. Then we have the following theorem.

**Theorem 2.** Let $H = EA_t^T A_t$ and assume that $\lambda_1(H)$ falls in the open interval $(0, 1)$. Then, the sequence $\{|D_t||\}$ converges to 0 in probability.

**Proof.** Notice that $||D_t||_2^2 = \sum_{j=1}^d ||D_t e_j||_2^2$. Then,

$$
E||D_{t+1}e_i||_2^2 = E \{ E \left[ ||D_{t+1}e_i||_2^2 \left| A_0, ..., A_{t-1} \right. \right] \} \\
= E \left\{ E \left[ e_i^T D_t^T A_t D_t e_i \left| A_0, ..., A_{t-1} \right. \right] \right\} = E \{ e_i^T D_t^THD_t e_i \} \\
\leq \lambda_1(H)E \{ e_i^T D_t^T D_t e_i \} = \lambda_1(H)E||D_t e_i||_2^2 \leq \cdots \\
\leq \lambda_t^{t+1}(H)||e_i||_2^2 = \lambda_t^{t+1}(H). 
$$

Thus, by Markov inequality,

$$
Pr(||D_t||_2^2 > \varepsilon) = Pr \left( \sum_{i=1}^d ||D_t e_i||_2^2 > \varepsilon \right) \leq \frac{1}{\varepsilon} \sum_{i=1}^d E||D_t e_i||_2^2 \leq \frac{d}{\varepsilon} \lambda_t^{t+1}(H).
$$

Sine $\lambda_1(H) \in (0, 1)$, $E||D_t e_i||^2 \rightarrow 0$ as $t$ goes to infinity. So we proved. \qed

Considering SSPSA with quadratic objective function (4.1). Substituting $y(\cdot)$'s in (4.1) by (4.2) yields

$$
\theta_{t+1} = (I - s\rho_t^T \rho_t B)\theta_t + (-s\rho_t^T \rho_t B + \epsilon_t^*) \tag{4.8}
$$

where $\rho_t = \Delta_t/\nu$. Letting $X_t = \theta_t$, $A_t = I - s\rho_t^T \rho_t B$, and $C_t = -G_t b + \epsilon_t^*$ turns out to be exactly the same form of RCA model (4.5). Note that $E\rho \rho^T = I \times \xi^2/\nu^2$, see
Appendix B. Immediately, we have

\[ EA_t^T A_t = E(I - s \rho_t \rho_t^T B - s B \rho_t \rho_t^T + s^2 B \rho_t \rho_t^T B) = I - 2s \xi^2 B / \nu^2 + s^2 \xi^2 B^2 / \nu^2 \]

and define the function

\[ \varrho(s, \lambda_j^B) = 1 - \frac{2s \xi^2 \lambda_j^B / \nu^2 + s^2 \xi^2 (\lambda_j^B)^2 / \nu^2}{2} = 1 - \frac{\xi^2}{\nu^2} \left(1 - s \lambda_j^B \right)^2 \]

for any \( i = 1, \ldots, d \). After some algebra, we know that as long as

Condition S: \( s \in (0, 2 / \lambda_1^B) \) and

Condition R: \( \xi^2 / \nu^2 \in (0, 1] \)

are satisfied, the largest eigenvalue of \( E A_t^T A_t \) will be strictly less than 1. For instance, if \( \lambda_1^B = \cdots = \lambda_d^B = 1 \) then the largest eigenvalue of \( E A_t^T A_t \) is \( \varrho_1(s) = 1 - \xi^2 / \nu^2 (2s - s^2) \). If not all eigenvalues are equal we can show that, given \( s \in (0, 2 / \lambda_1^B) \), the largest eigenvalue of \( E A_t^T A_t \) is

\[ \varrho_1(s) = \max_{j=1, \ldots, d} \varrho(s, \lambda_j^B) \in (0, 1). \]

Note that Condition R is satisfied by using Bernoulli or unit sphere random direction, see Appendix B.

**Theorem 3.** With Condition S and R, the TLE of the model (4.8) is strictly negative.

Proof. Let \( \gamma_t = E \{ \frac{1}{t} \log ||D_t||_F \} \). Then

\[ \gamma_t \leq \frac{1}{t} \log \left( E ||D_t||_F \right) \text{ by Jensen’s Inequality} \]
\[ \leq \frac{1}{t} \log \left( E ||D_t||_F^2 \right)^{1/2} \text{ by Hölder’s Inequality} \]
\[ \leq \frac{1}{2t} \log \left\{ d \times (\varrho_1(s))^t \right\} \text{ by (4.7)} \]
\[ = \frac{\log d}{2t} + \frac{1}{2} \log \varrho_1(s) = f(t) \]

which suffices to say that the TLE \( \gamma = \inf_t \gamma_t \) is strictly less than 0 and has an upper bound \( \log \varrho_1(s) / 2 \) because \( \inf \gamma_t \leq \lim \sup \gamma_t \leq \lim \sup f(t) = \lim f(t) \) since \( f(t) \) is a
monotone decreasing function with respect to $t$ for any $s \in (0, 2/\lambda_1^B)$. With the fact that all norms are equivalent, we have

$$\gamma_t \leq \frac{1}{t} \log (E||D_t||) \leq \frac{\log k}{2t} + \frac{1}{2} \log \varphi_1(s)$$

for some finite and positive $k$. So we proved.

Together with Theorem 1 and 3, the sequence $\{\theta_t\}$ is strictly stationary as long as Conditions S and R are satisfied. We close this section with following remarks.

**Remark 1.** A straightforward way to show $\gamma < 0$ is to show $||A_0|| \in (0, 1)$ and then applied the relationship $||A_1A_0|| \leq ||A_1||||A_0||$ and the theory that

$$\gamma = \lim_{t \to \infty} \frac{1}{t+1} \log ||A_t \cdots A_0||$$

almost surely, Furstenberg and Kesten(1960) and Kingman(1973). Unfortunately, if the square matrix norm is considered the largest eigenvalue of $A_0^T A_0$ is always larger than 1, see Appendix C. This also tells that $E\lambda_1(A^T A) > 1 > \lambda_1(EA^T A)$.

**Remark 2.** In reality, $B$ may not be symmetric. However, from (4.1), we know $\theta^T B \theta = \theta^T B^T \theta$. Thus, $B^* = (B + B^T)/2$ such that $\theta^T B^* \theta = \theta^T B \theta$, i.e. the value of objective function is not changed by replacing $B$ by $B^*$.

### 4.4 Mean and Covariance of SSPSA

In this section, we consider the mean and variance of $\theta_t$. From Theorem 1 and 3, we know that the sequence $\{\theta_t\}$ is stationary but do not know if the sequence has the desired mean. So, first of all, we derive the mean of $\theta_t$ which is asymptotically unbiased to the true solution. Second, we derive the variance of $\theta_t$ and the matrix autocorrelation function (ACF), analogous to the one-dimensional AR model. In addition, note the correspondence between $\theta_t$ and $X_t$ since model (4.8) and model (4.5) are identical. Below, the mean, variance and ACF of $X_t$ are shown in Proposition 1 and 2.
Proposition 1. With Condition S, R and model (4.8), \( \lim_{t \to \infty} E(\theta_t) = -B^{-1}b \), the true solution.

Proof. Immediately, we have

\[
X_{t+1} = A_tX_t + C_t = A_t(A_{t-1}X_{t-1} + C_{t-1}) + C_t \\
= A_tA_{t-1}X_{t-1} + A_tC_{t-1} + C_t \\
= \ldots \\
= A_t \ldots A_0X_0 + C_t + \sum_{j=0}^{t-1} \left( \prod_{i=0}^{j} A_{t-i} \right) C_{t-j-1}.
\]

(4.9)

Note that \( EA_t = I - s^*B \) with \( s^* = s \times \xi^2/\nu^2 \) and, under Condition S, the eigenvalue of \( EA_t \) is strictly less than 1, too. So, the expectation of the first term above becomes \((I - s^*B)^tX_0\) which converges to 0; of the second term is \(-s^*b\); and, consequently, we have

\[
EX_{t+1} = (I - s^*B)^{t+1}X_0 + \sum_{j=0}^{t} (I - s^*B)^j(-s^*b) \longrightarrow -B^{-1}b
\]

as \( t \) goes to infinity by von Neumann series, see Appendix A.2.

Without loss of generality, we can derive the variance with \( b = 0 \). Let \( \theta^* = \theta - B^{-1}b \) so that \( y(\theta^*) = a^* + \theta^TB\theta/2 + \epsilon \) where \( a^* = a - b^TB^{-1}b/2 \) which is known, too. Clearly, after centering to its mean, \(-B^{-1}b\), the centered random variable \( \theta^* \) has the same variance as the uncentered one \( \theta \). So letting \( b = 0 \) does not change the variance but provides a great convenience in inference especially in our case. Hereafter, we have \( EC_t = 0 \) and \( EX_t = 0 \).

Proposition 2. With Condition S and model (4.8), the variance of \( X_{t+1} \) is

\[
V_{X,t+1} = \sigma^2 \times \frac{\xi^2}{\nu^2} \times \sum_{j=0}^{t} ED_jD_j^T + Var(D_{t+1}X_0).
\]

In addition, for any integer \( h \geq 0 \), the autocorrelation function, the covariance of \( X_{t+1} \) and \( X_{t+1+h} \), is approximately

\[
R_{X,t+1}(h) = V_{X,t+1}(I - s^*B)^h.
\]
Proof. Define $D_0 = I$. Then, for the variance,

$$EX_{t+1}X_{t+1}^T = E\{C_tC_t^T\} + E\left(\sum_{j=0}^{t-1} \left[ \prod_{i=0}^{j} A_{t-i} \right] C_{t-j-1} + C_t \right)^T \left(\sum_{j=0}^{t-1} \left[ \prod_{i=0}^{j} A_{t-i} \right] C_{t-j-1} + C_t \right)^T + E(D_{t+1}X_0X_0^T D_{t+1}^T)
$$

$$= I \times \sigma^2 \times \frac{\xi^2}{\nu^2} + \sum_{j=1}^{t} E \left\{ \left[ \prod_{i=0}^{j} A_{t-i} \right] C_{t-j-1} C_{t-j-1}^T \left[ \prod_{i=0}^{j} A_{t-i} \right]^T \right\} + \text{Var}(D_{t+1}X_0)
$$

$$= I \times \sigma^2 \times \frac{\xi^2}{\nu^2} + \sigma^2 \times \frac{\xi^2}{\nu^2} + \sum_{j=1}^{t} E \left\{ \left[ \prod_{i=0}^{j} A_{t-i} \right] \left[ \prod_{i=0}^{j} A_{t-i} \right]^T \right\} + \text{Var}(D_{t+1}X_0)
$$

Note that $E \left[ \prod_{i=0}^{j} A_{t-i} \right] \left[ \prod_{i=0}^{j} A_{t-i} \right]^T$ is equal to $E \left[ \prod_{i=0}^{j} A_{j-i} \right] \left[ \prod_{i=0}^{j} A_{j-i} \right]^T$ since $A_i$'s are i.i.d. In this sequel,

$$EX_{t+1}X_{t+1}^T = \sigma^2 \times \frac{\xi^2}{\nu^2} \times \sum_{j=0}^{t} E \{D_j D_j^T\} + \text{Var}(D_{t+1}X_0).
$$

For shorthand, let $R_{X,t+1}(h) = \text{Cov}(X_{t+1}, X_{t+1+h})$. Then,

$$R_{X,t+1}(h) = EX_{t+1}X_{t+1+h}^T
$$

$$= E\left(\sum_{j=0}^{t-1} \left[ \prod_{i=0}^{j} A_{t-i} \right] C_{t-j-1} + C_t \right)^T \left(\sum_{j=0}^{t-1} \left[ \prod_{i=0}^{j} A_{t+i} \right] C_{t-j-1} + C_t \right)^T
$$

$$= E\left(\sum_{j=0}^{t-1} \left[ \prod_{i=0}^{j} A_{t-i} \right] C_{t-j-1} + C_t \right)^T \left(\sum_{j=0}^{t-1} \left[ \prod_{i=0}^{j} A_{t-i} \right] C_{t-j-1} + C_t \right)^T
$$

$$= E\left(\sum_{j=0}^{t-1} \left[ \prod_{i=0}^{j} A_{t-i} \right] C_{t-j-1} + C_t \right)^T \left(\sum_{j=0}^{t-1} \left[ \prod_{i=0}^{j} A_{t-i} \right] C_{t-j-1} + C_t \right)^T
$$

$$= V_X(I - s^*B)^h
$$

for any positive integer $h$. So we showed. 

Remark 3. It is worthwhile to notice that

$$EX_{t+1+h}X_{t+1}^T = (I - s^*B)^h V_{X,t+1}.
$$
Remark 4. According to (4.7), we have
\[ \text{tr}(V_{X,t}) \leq \sigma^2 \times \frac{\xi^2}{\nu^2} \frac{1}{1 - \lambda_1(H)} < \infty \]
for all \( t = 1, 2, \ldots \). Moreover, \( V_{X,t} \) is a variance-covariance matrix so \( V_{X,t} \) should be nonnegative definite. These facts suffice for implying \( 0 < \| V_{X,t} \| < \infty \).

In practice, \( t \) can be very large so it is helpful that the limit of \( V_{X,t} \) is known to us. Let \( V_X = \lim_{t \to \infty} V_{X,t} \). Evaluating \( V_X \) is not trivial. In this paragraph, we describe how to derive it by taking advantage of Kronecker operator, see Appendix A.3. Note that \( D_tD_t^T = A_{t-1} \cdots A_0 I A_0^T \cdots A_{t-1}^T \) and let \( K = E(A_t \otimes A_0) \). Then
\[ \text{vec} \left( D_tD_t^T \right) = (A_{t-1} \cdots A_0) \otimes (A_{t-1} \cdots A_0) \text{vec}(I) \]
and thus,
\[ \text{vec}(V_X) = \frac{\alpha^2 \xi^2}{\nu^2} \sum_{j=0}^{\infty} K^j \text{vec}(I_d) = \frac{\alpha^2 \xi^2}{\nu^2} (I_d^2 - K)^{-1} \text{vec}(I_d). \]
Notice that \( \text{vec}(V) \) is a function that piles up the matrix \( V \) one column by another into a vector. So if \( V_X \) is a \( d \times d \) matrix then \( \text{vec}(V_X) \) is a \( d^2 \times 1 \) vector. In addition, \( K \) is a \( d^2 \times d^2 \) matrix. Now, the last piece of deriving \( V_X \) is evaluating \( K \). Denote \( \alpha = \xi^2/\nu^2 \). Then,
\[ K = E(A_t \otimes A_t) = E((I - s \rho_t \rho_t^T B) \otimes (I - s \rho_t \rho_t^T B)) \]
\[ = E(I \otimes I - s \rho_t \rho_t^T B \otimes I - s I \otimes \rho_t \rho_t^T B + s^2(\rho_t \rho_t^T B) \otimes (\rho_t \rho_t^T B)) \]
\[ = I \otimes I - s \alpha s(B \otimes I) - s \alpha s(I \otimes B) + s^2 E[(\rho_t \otimes \rho_t)(\rho_t^T \otimes \rho_t^T)](B \otimes B). \]
Let \( U_{ij} = e_i e_j^T \), \( m_{22} = E \rho_{i1}^2 \rho_{i2}^2 < \infty \) and \( m_4 = E \rho_{i1}^4 < \infty \). The last expectation above becomes
\[ E[(\rho_t \otimes \rho_t)(\rho_t^T \otimes \rho_t^T)] = E[((\rho_t \rho_t^T) \otimes (\rho_t \rho_t^T))] = \sum_{i,j,k,l} w_{ijkl} U_{ij} \otimes U_{kl} \]
\[ = \sum_{i=1}^{d} U_{ii} \otimes U_{ii} + m_{22} \sum_{i \neq j} U_{ii} \otimes U_{jj} + m_{22} \sum_{i \neq j} U_{ij} \otimes U_{ij} + m_{22} \sum_{i \neq j} U_{ij} \otimes U_{ji}. \]
where \( w_{ijkl} = E(\rho_i \rho_j \rho_k \rho_l) \). As mentioned in Appendix B, 

\[
 w_{ijkl} = \begin{cases} 
 m_4 & \text{if } i = j = k = l \\
 m_{22} & \text{if } (i = j, k = l) \text{ or } (i = k, j = l) \text{ or } (i = l, k = j) \\
 0 & \text{otherwise}
\end{cases}
\]

For instance, if \( d = 3 \) then \( E[(\rho_i \rho_i^T) \otimes (\rho_i \rho_i^T)] \) should be

\[
\begin{bmatrix}
 m_4 & 0 & 0 & 0 & m_{22} & 0 & 0 & 0 & m_{22} \\
 0 & m_{22} & 0 & m_{22} & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & m_{22} & 0 & 0 & 0 & m_{22} & 0 & 0 \\
 0 & m_{22} & 0 & m_{22} & 0 & 0 & 0 & 0 & 0 \\
 m_{22} & 0 & 0 & 0 & m_4 & 0 & 0 & 0 & m_{22} \\
 0 & 0 & 0 & 0 & m_{22} & 0 & m_{22} & 0 & 0 \\
 0 & 0 & m_{22} & 0 & 0 & 0 & m_{22} & 0 & 0 \\
 0 & 0 & m_{22} & 0 & 0 & 0 & m_{22} & 0 & 0 \\
 m_{22} & 0 & 0 & 0 & m_{22} & 0 & 0 & 0 & m_4
\end{bmatrix}
\]

which is symmetric, sparse and singular. Let \( B = \text{diag}\{3, 2, 1\} \) and \( \Delta_i \)'s are sampled independently from Bernoulli taking value 1 or \(-1\) with equal probability. In what follows, we have \((\nu^2, \xi^2) = (3, 1)\) and \((m_{22}, m_4) = (1/9, 1/9)\) and thus, when \( s = 0.5 \in (0, 2/3)\)

\[
 V_X = \sigma^2 \times \frac{1}{3} \times \begin{bmatrix}
 2/3 & 0 & 0 \\
 0 & 1 & 0 \\
 0 & 0 & 2
\end{bmatrix}
\]

whereas when \( s = 0.25 \)

\[
 V_X = \sigma^2 \times \frac{1}{3} \times \begin{bmatrix}
 8/9 & 0 & 0 \\
 0 & 4/3 & 0 \\
 0 & 0 & 7/3
\end{bmatrix}.
\]

Up to now, the asymptotic mean and variance of \( X_t \) is available. However, in reality, \( X_0, ..., X_T \) is observed. It is not efficient to use \( X_T \) as our estimator since
its variance is a constant even when $T$ is large. On the other hand, we may use the same initial value $X_0$ but different sequences of random direction in order to get replicates of $X_T$, say $X_{T1}, \ldots, X_{Tn}$. It is natural to use sample mean of $X_{Ti}$'s as an estimate. However, the computing time increases inevitably. Taking advantage of the stationarity of $\{X_t\}$, the sample mean of $X_0, \ldots, X_{n-1}$ denoted as $\tilde{X}_{n-1}$ can be a relevant estimator and its variance is

$$\text{Var}(\tilde{X}_{n-1}) = \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} E X_iX_j^T - \frac{1}{n^2} E \left( \sum_{i=0}^{n-1} X_i \right) E \left( \sum_{j=0}^{n-1} X_j^T \right)$$

$$\approx \frac{1}{n^2} \left\{ n V_X + \sum_{i<j} V_X(I - L)^{j-i} + \sum_{i>j} (I - L)^{i-j} V_X \right\}$$

$$= \frac{1}{n^2} \left\{ -n V_X + \sum_{j=0}^{n-1} (n - j)V_X(I - L)^j + \sum_{j=0}^{n-1} (n - j)(I - L)^j V_X \right\}$$

$$\approx \frac{1}{n^2} \left\{ -n V_X + V_X((n + 1)L^{-1} - L^{-2}) + ((n + 1)L^{-1} - L^{-2})V_X \right\}$$

$$\approx \frac{1}{n} \left( -V_X + V_X L^{-1} + L^{-1}V_X \right)$$

(4.10)

where $L = s^*B = s \times \xi^2/\nu^2B$ and $L^{-1} = B^{-1}/s^*$. For the case just discussed with $B = \text{diag}\{3, 2, 1\}$. When $s = 0.5$,

$$n \text{Var}(\tilde{X}_{n-1}) \approx \sigma^2 \times \begin{bmatrix} 0.11 & 0 & 0 \\ 0 & 0.67 & 0 \\ 0 & 0 & 3.33 \end{bmatrix}$$

and when $s = 0.25$,

$$n \text{Var}(\tilde{X}_{n-1}) \approx \sigma^2 \times \begin{bmatrix} 0.30 & 0 & 0 \\ 0 & 0.89 & 0 \\ 0 & 0 & 3.89 \end{bmatrix}.$$  

**Remark 5.** In our logic, the well-scaled SSPSA has the form

$$\theta_{t+1} = \theta_t - s \times S \Delta_t \frac{L_t^+ - L_t^-}{2c||S \Delta_t||_2}$$

(4.11)
where $L_t^+ = L(\theta_t + S\Delta_t \times c)$, $L_t^- = L(\theta_t - S\Delta_t \times c)$ and $S = \text{diag}\{\theta_0\}$ where $\theta_0$ is a vector of proper initial values. Consequently, $\theta_t$’s are actually evaluated in the same scale. To clarify this point, consider the marginal likelihood having a quadratic form as (4.1) then (4.11) yields

$$\theta_{t+1} = \theta_t - c^* \times S\Delta_t \Delta_t^T S B \theta_t + \epsilon_t^* = (I - c^* \times S\Delta_t \Delta_t^T S B) \theta_t + \epsilon_t^*$$

where $c^* = c/||A_0\Delta_t||$ and $\epsilon^* = S\Delta_t (\epsilon_t^+ - \epsilon_t^-)/2||S\Delta_t||$. After some algebra, the equation above becomes

$$\varphi_{t+1} = (I - c^* \Delta_t \Delta_t^T S B S) \varphi_t + S^{-1} \epsilon_t^*$$

where $\varphi_t = S^{-1} \theta_t$ which should be a vector with elements nearly 1, hopefully. Notice that, in previous chapter, the mean and variance are derived without considering parameter scaling $S$. Therefore, the sequence resulting from the algorithm (4.11) should be scaled by $S$ to match the standard deviation derived above.

### 4.5 The Ergodicity of SSPSA

As we have shown that SSPSA can be treated as a RCA model, the conditions for $X_t$ or $\theta_t$ to be ergodic have been given by Feigin and Tweedie (1985). In this section, we introduce the theory and prove the required conditions. First, let us take a closer look at (4.5). The random coefficient $A_t$ can be factored into two parts: the fixed part, $\Theta$, and the random part, $\Gamma_t$, i.e.

$$X_{t+1} = (\Lambda + \Gamma_t)X_t + C_t. \quad (4.12)$$

In addition, let $K = E\Gamma_t \otimes \Gamma_t$. The theory below postulates sufficient conditions for the ergodicity of SSPSA.

**Theorem 4.** *Theorem 3, Feigin and Tweedie (1985)*

*With the model (4.5), representation (4.13), and assumption*
1. Both \( \{\Gamma_t\} \) and \( \{C_t\} \) are i.i.d. sequences with zero mean and \( \mathbb{E}\Gamma_t C_t = 0 \).

2. The random vector \( C_t \) has a positive probability mass around the origin.

3. Eigenvalues of \( (\Lambda \otimes \Lambda + K) \) have moduli less than unity.

then the process \( \{X_t\} \) is geometrically ergodic.

The equation (4.5) can be rewritten as

\[
X_{t+1} = (I - s\rho_t\rho_T^T B)X_t + \rho_t \epsilon_t
\]

\[
= (I - s\alpha B + (-s\rho_t\rho_T^T B + s\alpha B))X_t + \rho_t \epsilon_t
\]

where \( \alpha = \xi^2/\nu^2 \). So, from now on, let \( \Lambda = I - s\alpha B, \Gamma_t = s\alpha B - s\rho_t\rho_T^T B, C_t = \rho_t \epsilon_t \) and \( K = \mathbb{E}\Gamma_t \otimes \Gamma_t \). With this representation, it is easy to show that \( \mathbb{E}\Gamma_t = 0 \in \mathbb{R}^{d \times d}, \mathbb{E}C_t = 0 \in \mathbb{R}^d \) and \( \mathbb{E}\Gamma_tC_t = 0 \in \mathbb{R}^d \). Assumption 2 can be fulfilled by a carefully chosen distribution of the random error \( \epsilon_t \) and the questionable one is assumption 3.

The matrix \( K \) is evaluated first. Define \( U_{ij} \in \mathbb{R}^{d \times d} \) with 1 in \((i, j)\) entry and 0 otherwise. Then \( K \) can be rewritten as below.

**Proposition 3.**

\[
K = s^2 \left( -\alpha^2 I_d \otimes I_d + \eta^2 \sum_{i=1}^{d} U_{ii} \otimes U_{ii} + \alpha^2 \sum_{i \neq j} U_{ii} \otimes U_{jj} + \alpha^2 \sum_{i \neq j} U_{ij} \otimes U_{ij} \right) (B \otimes B)
\]

where \( \alpha = \xi^2/\nu^2 \) and \( \eta^2 = \mathbb{E}\rho_t^4 \).

**Proof.**

\[
K = \mathbb{E}\Gamma_t \otimes \Gamma_t = s^2 \mathbb{E}\left\{ (\alpha I_d - \rho_t\rho_T^T B) \otimes (\alpha I_d - \rho_t\rho_T^T B) \right\}
\]

\[
= s^2 \mathbb{E}\left\{ (\alpha I_d - \rho_t\rho_T^T) \otimes (\alpha I_d - \rho_t\rho_T^T) \right\} \{B \otimes B\}
\]

\[
= s^2 \mathbb{E}\left\{ \alpha^2 I_d \otimes I_d - \alpha I_d \otimes \rho_t\rho_T^T - \alpha \rho_t\rho_T^T \otimes I_d + \rho_t\rho_T^T \otimes \rho_t\rho_T^T \right\} \{B \otimes B\}
\]

\[
= s^2 \mathbb{E}\left\{ \alpha^2 I_d^2 + \mathbb{E}[(\rho_T \rho_T^T) \otimes (\rho_t \rho_T^T)] \right\} \{B \otimes B\}
\]

\[
= s^2 \left\{ -\alpha^2 I_d^2 + \mathbb{E}[(\rho_t \otimes \rho_t)(\rho_T^T \otimes \rho_T^T)] \right\} \{B \otimes B\}
\]
where $E\rho_t \rho_t^T = \alpha$. In addition, since $E\rho_t = 0$, by Condition R, we have

$$E(\rho_t \otimes \rho_t)(\rho_t^T \otimes \rho_t^T) = \eta^2 \sum_{i=1}^d U_{ii} \otimes U_{ii} + \alpha^2 \sum_{i \neq j} U_{ii} \otimes U_{jj}$$

$$+ \alpha^2 \sum_{i \neq j} U_{ij} \otimes U_{ij} + \alpha^2 \sum_{i \neq j} U_{ij} \otimes U_{ji}$$

so we showed. $\square$

Following, we are going to describe several propositions about matrices which may be part of common sense. However, they are important to the future inference so we list them without proof.

**Proposition 4.** With the fact that $U_{ij} = e_i e_j^T$, we have

$$U_{i1i2}^T U_{j1j2} = \begin{cases} 0 & \text{if } i_1 \neq j_1 \\ U_{i2j2} & \text{if } i_1 = j_1 \end{cases}$$

**Proposition 5.** Let $U = \sum_{ij} w_{ij} U_{ij} \otimes U_{ij}$. Then, by Proposition 4,

$$U^T U = \sum_{j=1}^d \sum_{k=1}^d w_{jk}^* U_{jk}^T \otimes U_{jk}$$

where $w_{jk}^* = \sum_{i=1}^d w_{ij} w_{ik}$.

**Proposition 6.**

$$||K||_F \leq \alpha^2 s^2 \sqrt{d \times \sum_{j=1}^d (\lambda_j^B)^4}$$

**Proof.** First, we rewrite (4.14) as

$$K = s^2 U (B \otimes B)$$

where

$$U = -\alpha^2 I_d \otimes I_d + \eta^2 \sum_{i=1}^d U_{ii} \otimes U_{ii} + \alpha^2 \sum_{i \neq j} U_{ii} \otimes U_{jj}$$

$$+ \alpha^2 \sum_{i \neq j} U_{ij} \otimes U_{ij} + \alpha^2 \sum_{i \neq j} U_{ij} \otimes U_{ji}$$

$$= \sum_{i=1}^d \sum_{j=1}^d w_{ij} U_{ij} \otimes U_{ij}.$$
Note that the value of $w_{ij}$ belongs to the set $\{-\alpha^2, \eta^2 - \alpha^2, 0, \eta^2, \alpha^2\}$ and most of them are 0. Bear in mind that $|w_{ij}| \leq \alpha^2$. Next, by Proposition 5, we know that $U^T U$ is sum of $(U_{jk} \otimes U_{jk})$’s with weight $\tilde{w}_{jk} = \sum_i w_{ij} w_{jk} \leq d \alpha^2$. Therefore,

$$||K||_F^2 = \text{tr}(K^T K) = s^4 \text{tr}((B^T \otimes B^T)U^T U (B \otimes B))$$

$$= s^4 \sum_{j=1}^d \sum_{k=1}^d \tilde{w}_{jk} \text{tr}((B \otimes B)(U_{jk} \otimes U_{jk})(B \otimes B))$$

$$= s^4 \sum_{jk} \tilde{w}_{jk} \text{tr}((BU_{jk}B) \otimes (BU_{jk}B))$$

$$= s^4 \sum_{jk} \tilde{w}_{jk} \text{tr}((BU_{jk}B))^2 = s^4 \sum_{jk} \tilde{w}_{jk} \text{tr}(e_k^T B^2 e_j)^2$$

$$\leq d \alpha^4 s^4 \sum_{jk} ([B^2]_{jk})^2 = d \alpha^4 s^4 ||B^2||_F^2.$$

Recall that $||B^2||_F^2 = \text{tr}(B^4) = \sum_{j=1}^d (\lambda_{Bj}^d)^4$ so we proved. \hfill \qed

**Theorem 5.** With Condition B and R, if

$$\alpha s \in S'' = \left(0, \frac{2 \lambda_d^B}{(\lambda_d^B)^2 + \sqrt{d \times \sum_{j=1}^d (\lambda_{Bj}^d)^4}}\right)$$

then the sequence yielded by model (4.5) is ergodic.

**Proof.** By Theorem 4, the proof is completed by checking the largest eigenvalue of matrix $\Lambda \otimes \Lambda + K$. Thus

$$||\Lambda \otimes \Lambda + K||_2 \leq ||\Lambda \otimes \Lambda||_2 + ||K||_2 \leq \lambda_1(\Lambda \otimes \Lambda) + ||K||_F$$

$$= (1 - s \alpha \lambda_d^B)^2 + s^2 \alpha^2 p$$

where $p^2 = d \times \sum_{j=1}^d (\lambda_{Bj}^d)^4$. The region $S''$ is derived so that $(1 - s \alpha \lambda_d^B)^2 + s^2 \alpha^2 p \leq 1$. \hfill \qed

For instance, let $B = I_d$ and therefore $\lambda_1^B = \cdots = \lambda_d^B = 1$. The resulting ergodic region is thus $s \in (0, 2/(1 + d)]$. As the dimension increases, this region becomes smaller and smaller. The situation becomes worse when the smallest eigenvalue of $B$, $\lambda_d^B$ is tiny. In practice, a tiny $\lambda_d^B$ is not rare.
Chapter 5

Simulation and Case Studies

In this chapter, several simulations and real data analyses are presented, beginning with the Normal-Normal model. Since this model has a closed-form likelihood we can track the performance of SSPSA sequences because the exact solution and marginal likelihood are available. Then, a logistic regression model with random intercept is considered. This simulation study allows us to compare the performance among leading methods and SSPSA. Finally, the motivating examples discussed in Chapter 1 are revisited.

5.1 Simulation Studies

The simulation design follows that of Zeger and Karim (1991) consisting of 7 measurements on each of \( n \) individuals. The model is

\[
Y_{ij} | X_{1j}, X_{2i}, X_{3ij}, u_i \sim P(\mu_{ij})
\]

where the sampling distribution \( P \) is normal or binomial and

\[
\eta(\mu_{ij}) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2i} + \beta_3 X_{3ij} + u_i
\]

where \( \eta(\cdot) \) denotes the link function, \( X_{1i} \sim Bin(1, 0.5) \), \( i = 1, \ldots, n \), \( X_{2j} = j - 4 \), \( j = 1, \ldots, 7 \), \( X_{3ij} = X_{1j} \times X_{2i} \), \( u_i \sim N(0, \sigma^2) \). Moreover, \((\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2) = \)
(-2.5, 2.0, -1.0, -0.5, 1.0) and n = 100 or 30. As for the algorithm parameters, for AGQ, the number of abscissa equals to 5 and, for SSPSA, let \( c = 0.001, s = 1 \) since the objective function is already properly scaled, and the importance sampling sample size as 30. Estimates are averages of 4,000 iterations following a 1,000 burn-in period.

Other than mean / bias and standard error, we compare differences among three methods, AGQ, RSPL and SSPSA, by taking the dataset effect into account. The model for analysis is a two way crossed classified model with no interaction

\[
\text{Estimate}_{ij} = \mu + \text{METHOD}_i + \text{Dataset}_j + \epsilon_{ij}
\]

where \( i = 1, 2, 3, j = 1, \ldots, 200 \). In addition, the methods are indexed as \( i = 1 \) for AGQ, \( i = 2 \) for RSPL (Restricted Subject-specific Pseudo Likelihood) and \( i = 3 \) for SSPSA. Multiple comparisons among these three methods was done based on Bonferroni adjustment. (Of course, we assume the normality of \( \epsilon \)'s. Although it may not be true the result can still provide some idea about the difference / similarity among these methods.) Note that, among those methods provided by SAS procedure GLIMMIX, RSPL performs best according to our previous experiments. Therefore, RSPL is chosen to compete with SSPSA and AGQ.

Furthermore, the variance of the proposed estimator is of interest as well. In Section 4.4, we introduced the variance formula, Proposition 2 and (4.13), for SSPSA estimate of the MLE. Note that this variance expression only takes into account the randomness of random direction and of Monte Carlo integration. However, we want the variance considering the randomness from both the data and the SSPSA algorithm. Define the data as \( \mathbf{W} = (\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \) and the random components in SSPSA as \( \mathbf{R} = \{(\rho_i, \epsilon_i), i = 1, \ldots, n_s\} \). Then, the total variance of SSPSA estimate \( \hat{\theta}_{WR} \) should be decomposed as

\[
\text{Var}(\hat{\theta}_{WR}) = \text{Var}_W \left( E_R(\hat{\theta}_{WR}|\mathbf{W}) \right) + E_W \left( \text{Var}_R(\hat{\theta}_{WR}|\mathbf{W}) \right).
\]

As we have been shown that \( \hat{\theta}_{WR} \) is unbiased to the true MLE (based on data \( \mathbf{W} \)) and \( \text{Var}_R(\hat{\theta}_{WR}|\mathbf{W}) = O(n_s^{1/2}) \). Thus, the first term above can be replaced by the
inverse of Fisher's information matrix and the second term will be negligible if \( n_s \) is sufficiently large.

### 5.1.1 Normal Regression with Normal Random Intercept

The random intercept linear (N-N) model has the form

\[
Y_{ij} = u_i + x_{ij}^T \beta + \epsilon_{ij}
\]

where \( \epsilon_{ij} \sim i.i.d. \ N(0, 1) \), \( u_i \sim i.i.d. \ N(0, \sigma^2) \) and \( i = 1, ..., n \) (individuals), \( j = 1, ..., r \) (repeated measures.) Let \( \phi(\cdot; \mu, \sigma^2) \) denote the pdf of normal random variable with mean \( \mu \) and variance \( \sigma^2 \). The corresponding log marginal likelihood is

\[
\log(L_N(\theta)) = \sum_{i=1}^{n} \log \left\{ \int_{\mathbb{R}} \prod_{j=1}^{r} \phi(y_{ij}; u_i + x_{ij}^T \beta, 1) \phi(u_i; 0, \sigma^2) du_i \right\}
= -\frac{nr}{2} \log 2\pi - \frac{n}{2} \log(r\sigma^2 + 1) - \frac{1}{2} \sum_{ij} (y_{ij} - x_{ij}^T \beta)^2
+ \frac{1}{2} \frac{\sigma^2}{r\sigma^2 + 1} \sum_{i} \left( \sum_{j} (y_{ij} - x_{ij}^T \beta) \right)^2
\]

where \( \theta = (\beta^T, \sigma^2)^T \) and thus a closed-form solution of parameters is available.

The simulation considered model \( P(\mu_{ij}) = N(\mu_{ij}, 1) \) and identity link, namely, \( \eta(\mu_{ij}) = \mu_{ij} \). Both AGQ and RSPL method are expected to perform well because the underlying integrand has exactly the same form as its approximations. Table 5.1 displays the mean bias and standard error over 200 replications and Table 5.3 shows the \( p \)-values of testing “no difference between the approximation method and the exact solution.” The corresponding mean square errors are also listed in Table 5.2. Together, we conclude that, with N-N model, these methods provide good estimates to the true MLE, except for the estimation of variance component \( \sigma^2 \). Of course, SSPSA should not perform as well as the others (according to multiple comparison results) since it introduces more random errors.
Table 5.1: Mean Bias and Empirical Standard Error of Normal-Normal Model

<table>
<thead>
<tr>
<th></th>
<th>Sample Size $n = 100$</th>
<th>Sample Size $n = 30$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSPSA</td>
<td>AGQ</td>
</tr>
<tr>
<td>$\beta_0 = -2.5$</td>
<td>0.011</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.011)</td>
</tr>
<tr>
<td>$\beta_1 = 2.0$</td>
<td>-0.003</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>$\beta_2 = -1.0$</td>
<td>-0.013</td>
<td>-0.011</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>$\beta_3 = -0.5$</td>
<td>-0.003</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>$\sigma^2 = 1.0$</td>
<td><strong>0.029</strong></td>
<td><strong>-0.031</strong></td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
</tr>
</tbody>
</table>

Boldface numbers indicate a significant $t$-test at level 0.05.

Table 5.2: Mean Square Error ($\times 100$) of Normal-Normal Model

<table>
<thead>
<tr>
<th></th>
<th>Sample Size $n = 100$</th>
<th>Sample Size $n = 30$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSPSA</td>
<td>AGQ</td>
</tr>
<tr>
<td>$\beta_0 = -2.5$</td>
<td>2.87</td>
<td>2.46</td>
</tr>
<tr>
<td>$\beta_1 = 2.0$</td>
<td>0.21</td>
<td>0.07</td>
</tr>
<tr>
<td>$\beta_2 = -1.0$</td>
<td>5.19</td>
<td>5.19</td>
</tr>
<tr>
<td>$\beta_3 = -0.5$</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td>$\sigma^2 = 1.0$</td>
<td>2.98</td>
<td>2.88</td>
</tr>
</tbody>
</table>

Figure 5.1 and 5.2 show comparisons of these three methods versus the exact solution. Since it is well known that the MLE of $\sigma^2$ is not unbiased, both AGQ and SSPSA are not unbiased. On the other hand, RSPL applies REML and hence yields an unbiased estimate of $\sigma^2$. However, we are interested in the performance of SSPSA. Since the goal is finding the exact MLE, we expect that both SSPSA and AGQ are close to the exact MLE which is evident from the last row of Table 5.3. In conclusion, N-N model is good for testing the performance of SSPSA algorithm and the proposed algorithm behaves well. Moreover, we show the variation due to SSPSA in Figure 5.3. Two hundred SSPSA sequences were generated based on the same dataset but...
different random directions. Their empirical standard errors were calculated. Figure 5.3 tells that we can have 2 digits accuracy of the variance estimate since, after 4,000 steps, the extra variation due to SSPSA is less than 0.001.

Table 5.3: Multiple Comparisons with Control of N-N Model: $p$-values of Testing No Difference among Parameter Estimates

<table>
<thead>
<tr>
<th></th>
<th>Sample Size = 100</th>
<th>Sample Size = 30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-C</td>
<td>2-C</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1.00</td>
<td>0.01</td>
</tr>
</tbody>
</table>

1: AGQ; 2: RSPL; 3: SSPSA; C: control, the exact solution.

Figure 5.1: N-N model, sample size 100: Comparisons of three methods: SSPSA, AGQ and RSPL. Exact MLEs versus approximated MLEs are plotted.
5.1.2 Logistic Regression with Normal Random Intercept

Next, we examined the performance of SSPSA, AGQ and RSPL via logistic regression with random intercept (L-N) model. Accordingly, let $P(\mu_{ij}) = Bin(1, \mu_{ij})$ with logit link, $\eta(\mu_{ij}) = 1/(1 + \exp(-\mu_{ij}))$. Note that, under this model, the integrand is no longer symmetric so the approximation in AGQ becomes worse. As shown in Table 5.4, SSPSA outperformed AGQ in general. Surprisingly, it seems that RSPL beats the other two methods in terms of unbiasness or smallest mean square error. However, unlike SSPSA and AGQ, RSPL had smaller bias when the sample size, $n$, was smaller. This performance is suspicious but we have no explanation currently. Table 5.5 contains corresponding mean square errors. Similar to the conclusion from Table 5.4, RSPL is superior in general except the variance component $\sigma^2$ estimate.
Mean square errors show this tendency as well. Table 5.6 shows the $p$-values of multiple comparisons for testing parameter estimates. Surprisingly, none of these methods yields similar estimates in this model except some parameters. The natural variation of SSPSA can be seen in Figure 5.4.

### 5.2 Case Studies

SSPSA was applied to the following studies. We set $c = 0.001$ and step size $s = 1$ providing that the objective function is properly scaled. In addition, the importance sampling sample size was set case-by-case. For epilepsy seizure data, we used importance sampling sample size $n_{IS} = 30$ whereas, for lung cancer data and salamander data, we used $n_{IS} = 1,000$. The increase of sample size is due to that the corresponding surface of marginal likelihoods are extremely flat. According to our experience, such a flat surface makes the precision of the numerical derivative low and hence the SSPSA sequence becomes unstable. So we tried 1,000 to increase the
### Table 5.4: Mean Bias and Empirical Standard Error of Logistic-Normal Model

<table>
<thead>
<tr>
<th></th>
<th>Sample Size = 100</th>
<th>Sample Size = 30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSPSA</td>
<td>AGQ</td>
</tr>
<tr>
<td>$\beta_0 = -2.5$</td>
<td>-0.01</td>
<td>-0.07</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>$\beta_1 = 2.0$</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>$\beta_2 = -1.0$</td>
<td>-0.04</td>
<td>-0.09</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>$\beta_3 = -0.5$</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>$\sigma^2 = 1.0$</td>
<td>-0.17</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.05)</td>
</tr>
</tbody>
</table>

Boldface numbers indicate a significant $t$-test at level 0.05.

### Table 5.5: Mean Square Error ($\times 100$) of Logistic-Normal Model

<table>
<thead>
<tr>
<th></th>
<th>Sample Size $n = 100$</th>
<th>Sample Size $n = 30$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSPSA</td>
<td>AGQ</td>
</tr>
<tr>
<td>$\beta_0 = -2.5$</td>
<td>21.99</td>
<td>23.41</td>
</tr>
<tr>
<td>$\beta_1 = 2.0$</td>
<td>9.57</td>
<td>10.55</td>
</tr>
<tr>
<td>$\beta_2 = -1.0$</td>
<td>40.92</td>
<td>41.57</td>
</tr>
<tr>
<td>$\beta_3 = -0.5$</td>
<td>11.59</td>
<td>11.54</td>
</tr>
<tr>
<td>$\sigma^2 = 1.0$</td>
<td>18.02</td>
<td>41.76</td>
</tr>
</tbody>
</table>

Comparing to RSPL estimates, the bold face numbers indicate a non-significant paired-$t$ test at level 0.05.

5.2.1 Epilepsy Seizure Data

In this dataset, 59 epileptic patients were recruited. Seizure counts occurring during the first 8 weeks were recorded as the baseline. After patients were randomly assigned to treatment or control groups, seizure counts were recorded for the following 4 consecutive 2-week-long periods. Other than the treatment, covariates including (log) baseline counts, (log) age and an indicator for the fourth clinic visit. Only the
Table 5.6: Pairwise Multiple Comparisons of L-N Model: \( p \)-values of Testing No Difference of Parameter Estimates

<table>
<thead>
<tr>
<th>Sample Size = 100</th>
<th>Sample Size = 30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-2</td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>0.00</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>0.00</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>0.00</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>0.00</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>0.00</td>
</tr>
</tbody>
</table>

1: AGQ; 2: RSPL; 3: SSPSA.

Figure 5.4: L-N model, \( n = 100 \) (Left) and \( n = 30 \) (Right): Variation due to SSPSA with different numbers of iterations

treatment-baseline interaction is considered. The model proposed by Breslow and Clayton (1993) is

\[
Y_{ij}|X_{ij}, u_i \sim \text{Poisson}(\mu_{ij}) \quad \text{and} \quad \log(\mu_{ij}) = X_{ij}\beta + u_i
\]

where \((i,j)\) denotes the \( j \)th clinic visit of the \( i \)th individual, \( X_{ij} \) denotes the collection of covariates and interaction and \( u_i \sim N(0, \sigma^2) \). Results are summarized in Table 5.7.
As we can see, there is no significant difference among PQL, GQ and SSPSA solutions.

Table 5.7: Summary Results, Parameter Estimates (Standard Errors), for Epilepsy Seizure Data

<table>
<thead>
<tr>
<th></th>
<th>GLM*</th>
<th>PQL*</th>
<th>GQ*</th>
<th>SSPSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.90(.04)</td>
<td>1.87(.11)</td>
<td>1.88(.28)</td>
<td>1.89(.05)</td>
</tr>
<tr>
<td>lbase</td>
<td>0.95(.04)</td>
<td>0.88(.13)</td>
<td>0.89(.14)</td>
<td>0.95(.07)</td>
</tr>
<tr>
<td>trt</td>
<td>-0.34(.06)</td>
<td>-0.31(.15)</td>
<td>-0.34(.17)</td>
<td>-0.33(.08)</td>
</tr>
<tr>
<td>lage</td>
<td>0.89(.12)</td>
<td>0.53(.35)</td>
<td>0.47(.37)</td>
<td>0.57(.53)</td>
</tr>
<tr>
<td>V4</td>
<td>-0.16(.15)</td>
<td>-0.16(.08)</td>
<td>-0.16(.05)</td>
<td>-0.15(.01)</td>
</tr>
<tr>
<td>lbase x trt</td>
<td>0.56(.06)</td>
<td>0.34(.20)</td>
<td>0.34(.22)</td>
<td>0.31(.18)</td>
</tr>
<tr>
<td>σ</td>
<td></td>
<td>0.44</td>
<td>0.55(.07)</td>
<td>0.56(.03)</td>
</tr>
</tbody>
</table>

*: Venables and Ripley, pp. 297.

5.2.2 Lung Cancer Data

These Lung cancer data have been analyzed by many methods, including Cox and Snell, 1988, Booth and Hobert, 1999, among others. These data contain 14 retrospective studies regarding the relationship between smoking and lung cancer. As described in Chapter 1, Booth and Hobert (1999) used the model

\[ Y_{ij} \sim Bin(n_{ij}, p_{ij}) \]

and

\[ \eta_{ij} = \log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta_0 x_{ij} + \beta_1 x_{ij} + u_i + v_{ij} \]

where \( x_{ij} \) indicates the smoking or non-smoking of the \( i \)th study, \( u_i \)'s are \( i.i.d. \) samples from \( N(0, \sigma_u^2) \), \( v_{ij} \)'s are \( i.i.d. \) samples from \( N(0, \sigma_v^2) \), and \( u_i \) and \( v_{ij} \) are independent. Further, assume that conditional on \((x_{ij}, u_i, v_{ij})\)'s, \( Y_{ij} \)'s are independent. Note that the second random effect \( v_{ij} \) results in varying odds ratio among studies. That is to say, we have log odds ratio

\[ \eta_{i1} - \eta_{i2} = \log \left( \frac{p_{i1}(1 - p_{i2})}{(1 - p_{i1})p_{i2}} \right) = \beta_1 + v_{i1} - v_{i2} = \beta_1 + \epsilon_i \]
where \( \epsilon_i = v_{i1} - v_{i2} \sim N(0, 2\sigma^2_v) \). From the data, the sample odds ratios of the 14 studies are: 5.38, 5.68, 4.32, 1.84, 5.64, 12.77, 1.21, 9.08, 5.09, 3.87, 39.73, 3.18, 4.25, 5.622 whose sample variance supports the non-common-odds-ratio assumption.

The major difficulty of analyzing this dataset is that, along the dimensions of \( \sigma^2_u \) and \( \sigma^2_v \), the marginal likelihood surface is very flat. Figure 5.5 shows this surface when \((\beta_0, \beta)\) is fixed at (-1.9, 1.7). For two variance components, the interval \([0.02, 2.0]\) is evenly gridded into 100 intervals. The marginal likelihood at each grid point was approximated by MC integration with 1,000 MC samples. Table 5.8 compares results arising from different methods. There are no big differences among those methods. Notice that numbers within parentheses should be the corresponding standard error.

Table 5.8: Summary Results, Parameter Estimates (Standard Errors), for Lung Cancer Data

<table>
<thead>
<tr>
<th></th>
<th>RSPL</th>
<th>MCEM*</th>
<th>SSPSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>-1.90(0.18)</td>
<td>-1.93(0.08)</td>
<td>-1.94(0.09)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>1.66(0.22)</td>
<td>1.70(0.08)</td>
<td>1.70(0.05)</td>
</tr>
<tr>
<td>( \sigma^2_u )</td>
<td>0.20(0.16)</td>
<td>0.19(0.06)</td>
<td>0.19(0.01)</td>
</tr>
<tr>
<td>( \sigma^2_v )</td>
<td>0.25(0.13)</td>
<td>0.23(0.05)</td>
<td>0.24(0.06)</td>
</tr>
</tbody>
</table>


### 5.2.3 Salamander Data

The experiment was designed for showing the interbreeding tendency of two isolated salamander populations. Ten males and ten females were included for each population. A pair of salamanders, one male and one female, were coupled by design. If they mate, then a success is recorded. Every female/male was assigned to couple with three different males/females in successive time period. Each pair may or may not belong to the same population. The simplest model shown in Karim and Zeger (1992) is

\[
Y_{ijk} \sim Bin(1, p_{ij})
\]
Figure 5.5: Lung cancer data: Monte Carlo log-likelihood surface of $\sigma^2_u$ and $\sigma^2_v$ when $(\beta_0, \beta) = (-1.9, 1.7)$.

where $i = 1, ..., 20$ indicates female individuals, $j = 1, ..., 20$ indicates male individuals and $k = 1, 2, 3$ indicates independent mating experiments and

$$\eta_{ij} = \log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = X_{ij}\beta + u_i + u_j$$

where $X_{ij}$ indicates the population of the $i$th female and the $j$th male; $u_i \sim N(0, \sigma^2_f)$; and $u_j \sim N(0, \sigma^2_m)$. SSPSA algorithm yielded sequences in Figure 5.6. It shows
that, after 10,000 iterations, all sequences become stable and approach stationary except the one for \( \log(\sigma_f^2) \) which keeps going down. With this observation, we claim that there is no individual effect among female salamanders. Contrasting to SSPSA, RSPL solution supports that there is a tiny / no variation among males but a sizeable variation among females. Other two methods, PQL and Gibbs, suggest a significant individual variation for both male and female. In order to compare these inconsistent results, we plotted the marginal likelihood with fixed regression coefficient \( \beta_0, ..., \beta_3 \), see Figure 5.7. Both two panels of Figure 5.7 show that the maximum of the marginal likelihood locates at the left-lower corner of the \( \sigma_f^2 \) v.s. \( \sigma_m^2 \) plot. This provides the evidence that the SSPSA solution may be more reasonable and accurate.

Figure 5.6: SSPSA sequences for Salamander Mating Data with initial value provided by Gibbs estimate, Zeger and Karim, 1991
Table 5.9: Summary Results, Parameter Estimates (Standard Errors), for Salamander Mating Data

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameters</th>
<th>Parameters</th>
<th>Parameters</th>
<th>Parameters</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_0$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\beta_3$</td>
<td>$\phi^2_f$</td>
</tr>
<tr>
<td>GLM</td>
<td>0.69(0.22)</td>
<td>-2.01(0.34)</td>
<td>-0.47(0.31)</td>
<td>2.48(0.46)</td>
<td>-</td>
</tr>
<tr>
<td>SSPSA</td>
<td>0.73(0.08)</td>
<td>-2.11(0.10)</td>
<td>-0.50(0.18)</td>
<td>2.61(0.38)</td>
<td>-</td>
</tr>
<tr>
<td>PQL$^1$</td>
<td>0.79(0.32)</td>
<td>-2.29(0.43)</td>
<td>-0.54(0.39)</td>
<td>2.82(0.50)</td>
<td>0.72</td>
</tr>
<tr>
<td>Gibbs$^2$</td>
<td>1.03(0.43)</td>
<td>-3.01(0.60)</td>
<td>-0.69(0.50)</td>
<td>3.74(0.68)</td>
<td>1.50</td>
</tr>
<tr>
<td>RSPL$^3$</td>
<td>1.16(0.60)</td>
<td>0.78(0.57)</td>
<td>-1.41(0.61)</td>
<td>1.02(0.59)</td>
<td>1.41(0.89)</td>
</tr>
</tbody>
</table>


Figure 5.7: Monte Carlo marginal likelihood with (Left) SSPSA and with (Right) Gibbs fixed-effect coefficient estimates.
Chapter 6

Future Work

In this chapter, we briefly pinpoint some incompleteness of this work and describe some potential applications.

6.1 Properties of SSPSA

As we have claimed that the sample mean, say $\bar{X}$, of the SSPSA sequence is approximately unbiased and its asymptotic variance is available. We are interested in the question that whether or not the random variable $\bar{X}$ follows normal distribution. In time series literature, if the coefficient is fixed then the sample mean of a stationary AR(1) series is asymptotically normal. To my limited knowledge, there is no proof or disproof about the normality when the coefficient is random.

The sufficient condition for RCA model to be ergodic is given by Feigin and Tweedie (1985). Currently, we can show that if the step size $s$ falls in a small interval, see Theorem 5, then conditions given in Feigin and Tweedie (1985) are satisfied. However, this interval is too small to be practical. This drawback is due to a couple of loose inequalities participating in the proof. Improvement is possible.
6.2 Implementation Issues

The major assertion of SSPSA is that the marginal likelihood around the optimal point is quadratic. Based on this, SSPSA yields strictly stationary sequence and hence, the mean and variance of the sequence can be derived analytically when the quadratic function is known. In Section 4.2, the model for the observed response is

\[ y(\theta) = a + b^T \theta + \theta^T B \theta / 2 + \epsilon \]

where \( a \in \mathbb{R}, \, \, b \in \mathbb{R}^d, \, \, \theta \in \mathbb{R}^d, \, \, \epsilon \sim N(0, \sigma^2), \) and \( B \in \mathbb{R}^{d \times d} \) is symmetric and positive definite. Chen and McInroy (2002) proposed an approach for estimating \( B \) utilizing these assumptions on \( B \). However, their solution is not design invariant, i.e. different design points of \( \theta \) end up with very different estimation of \( B \). There is some room to improve.

It is worthwhile to emphasize again that function and parameter scaling are crucial to the SA type algorithm. However, their affection is unknown. For example, when function scaling is done the objective function changes, at least the curvature around the optimal point changes. Although this scaling makes the algorithm possible we would like to know the analytical effect of doing so. The quadratic approximation provides a manageable framework to discover such questions.

6.3 Possible Applications

Conceptually, if an objective function containing several unknown parameters is an integral without closed form then SSPSA can be applied and provide an estimate of the optimum. Of course, it requires that the objective function is locally convex / concave around the optimum with good initial values. Below, we describe a couple of such cases.

A root-finding problem in Finance proposed by Glasserman (2003) is define as

\[ \Pr(L_n > \theta) = p \]
where \( p \in (0, 1) \) is a given number, \( \theta \) is the unknown parameter and \( L_n \), random, denotes the total loss of the portfolio containing \( n \) individuals, usually \( n \) is huge. Given positive and fixed \( c_1, ..., c_n \)

\[
L_n = c_1 Y_1 + \cdots + c_n Y_n
\]

where \( Y_i \)'s indicate the occurrence of default, \( i.e. \ Y_i \sim Bin(1, p_i(Z)) \). In general, the random variable \( Z \) represents certain economical indices, \( e.g. \ GDP, \ interest \ rate \ and/or \ mortgage \ rate, \) and hence \( Y_i \)'s are neither identical nor independent. When all the underlying randomness are assigned the parameter \( \theta \) can be estimated by SSPSA.

Nonlinear measurement error model has no closed-form MLE in general. Take logistic regression with a contaminated covariate for example. Denote the true covariate as \( X \) and the observable covariate, a surrogate of \( X \), as \( W = X + U \) where \( U \sim F \) and \( E(U) = 0 \). Moreover, we have

\[
Y_i | X_i \sim Bin(1, p_i) \quad \text{and} \quad \eta_i = \log \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 X_i.
\]

It turns out that the marginal density becomes

\[
\int \Pr(Y_i | W_i = X_i + U) dF_U(u).
\]

Fortunately, SSPSA can handle this kind of problem since we have

\[
\eta_i = \beta_0 + \beta(W_i - U) = \beta_0 + \beta_1 W_i - \beta_1 U
\]

which has exactly the same form of logistic regression with random intercept. The complexity of such a problem can be higher when \( Y_i \)'s are correlated, see Wang, Q (2005) for example. However, SSPSA is conceptually unchanged and easy to implement as long as the likelihood can be written down explicitly.
Bibliography


Appendix A

Some Matrix Operations

Several properties of matrix operations are addressed in this appendix. They may be fundamental but not usually seen from elementary linear algebra books. These materials includes matrix norm, von Neumann series and Kronecker product.

A.1 Matrix Norm

Matrix $p$-norm of a matrix $A \in \mathbb{R}^{d \times d}$ is defined as

$$
\sup_{x \in \mathbb{R}^d} \frac{||Ax||_p}{||x||_p} = ||A||_p
$$

where $||\cdot||_p$ denotes the $p$-norm of a vector. Although all positive integer $p$ are eligible only $p = 1, 2,$ and $\infty$ yield closed form expression of $||A||_p$. Below, several popular norms with closed form expression are listed

1. $p = 1$, $||A||_1^2 = \left( \max_j \sum_{i=1}^d |A_{ij}| \right)^2$

2. $p = 2$, $||A||_2^2 = \lambda_1(A^TA)$

3. $p = \infty$, $||A||_\infty^2 = \left( \max_i \sum_{j=1}^d |A_{ij}| \right)^2$

4. Frobenius, $||A||_F^2 = \text{tr}(A^TA) = \sum_{ij} (A_{ij})^2$
where $A_{ij}$ is the $(i, j)$ element of matrix $A$. Later on, we also use the expression $A = [A_{ij}]_{i,j=1,...,d}$ if it makes the formula easy to understand. Note that the Frobenius norm is not a $p$-norm but has some good properties especially when $A$ is random. We are going to show them in the following sections.

For a given matrix $A \in \mathbb{R}^{d \times d}$, we quote the table from Stewart (1973), pp.183, in order to show the relationship among these matrix norm, see Table A.1. Indeed,

Table A.1: The value of $\tau_{pq}$ when $||A||_p \leq \tau_{pq}||A||_q$, Stewart (1973).

<table>
<thead>
<tr>
<th>$q$</th>
<th>1</th>
<th>2</th>
<th>$F$</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>$\sqrt{d}$</td>
<td>$\sqrt{d}$</td>
<td>$d$</td>
</tr>
<tr>
<td>$p$</td>
<td>$\sqrt{d}$</td>
<td>1</td>
<td>1</td>
<td>$\sqrt{d}$</td>
</tr>
<tr>
<td>$F$</td>
<td>$\sqrt{d}$</td>
<td>$\sqrt{d}$</td>
<td>1</td>
<td>$\sqrt{d}$</td>
</tr>
<tr>
<td>$\infty$</td>
<td>$d$</td>
<td>$\sqrt{d}$</td>
<td>$\sqrt{d}$</td>
<td>1</td>
</tr>
</tbody>
</table>

any two norms in the set $\mathcal{I} = \{1, 2, F, \infty\}$ are proportional up to a positive constant. This provides the potential that we only need to show convergence in one norm and then apply the result to others.

### A.2 von Neuman Series

Consider the sum of a power sequence of a matrix $A \in \mathbb{R}^{d \times d}$: $S_t = \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^t$. Then, assuming $(\mathbf{I} - \mathbf{A})$ is invertible and $||A||_2 < 1$, we have $S_t(\mathbf{I} - \mathbf{A}) = \mathbf{I} - \mathbf{A}^{t+1}$ and thus $S_t = (\mathbf{I} - \mathbf{A})^{-1} - \mathbf{A}^{t+1}(\mathbf{I} - \mathbf{A})^{-1}$. As $t \to \infty$, $||A^t||_2 \leq (||A||_2)^t \to 0$ by assumption. Consequently, we have the von Neumann series: $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots = (\mathbf{I} - \mathbf{A})^{-1}$. 
A.3 Some Properties of Kronecker Product

Denote the Kronecker operator as $\otimes$. For two matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{p \times q}$, the $A \otimes B$ is defined as

$$
\begin{bmatrix}
a_{11}B & \cdots & a_{1m}B \\
\vdots & \ddots & \vdots \\
a_{n1}B & \cdots & a_{nm}B
\end{bmatrix}
$$

Below, we summarize several useful results.

1. $A_1 \otimes (A_2 + A_3) = A_1 \otimes A_2 + A_1 \otimes A_3$

2. $(aA_1) \otimes A_2 = A_1 \otimes (aA_2) = a(A_1 \otimes A_2)$ where $a \in \mathbb{R}$

3. $(A_1 \otimes A_2)(A_3 \otimes A_4) = (A_1A_3) \otimes (A_2A_4)$

4. $(A_1 \otimes A_2)^T = A_1^T \otimes A_2^T$

5. $\text{vec}(A_1AA_2) = (A_2^T \otimes A_1)\text{vec}(A)$ where $\text{vec}(V)$ is a function that piles up the matrix $V$ one column by another into a vector.

6. Suppose that $P \in \mathbb{R}^{d \times d}$ is an orthogonal matrix. Then, $e_i \otimes P_j$, $i, j = 1, \ldots, d$ are eigenvectors of $P \otimes P$.

7. Suppose a symmetric matrix $A \in \mathbb{R}^{d \times d}$ can be factored into $P\Lambda P^T$, the spectral decomposition, where $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_d\}$. Then, the eigenpairs of $A \otimes A$ are $(\lambda_i \times \lambda_j, e_i \otimes P_j)$, for $i, j = 1, \ldots, d$. 
Appendix B

Moments of Random Directions

In this appendix, we discuss the moments of random directions mentioned in Chapter 4.

B.1 Bernoulli Random Direction

Let $\Delta = (\Delta_1, \ldots, \Delta_d)^T$ be a random vector where $\Delta_i$’s are i.i.d. random variables taking value 1 or -1 with equal probability which implies $E(\Delta_i) = 0$ and $\text{Var}(\Delta_i) = 1$ for $i = 1, \ldots, d$. Immediately, we have $||\Delta||^2 = d$, $E(\Delta) = 0$, $E(\Delta_i \Delta_j) = I(i = j)$ because $\Delta_i$ and $\Delta_j$ are independent if $i \neq j$, and so $E(\Delta \Delta^T) = I$. Therefore, we have $(\xi^2, \nu^2) = (1, d)$. Notice that Bernoulli random direction points the vertices of $d$-dimensional hypercube.

B.2 Random Vector on Unit Sphere

Let $\Delta = (\Delta_1, \ldots, \Delta_d)^T$ be a vector randomly chosen from the surface area of a unit $d$-dimensional ball. Before deriving $E(\Delta)$ and $E(\Delta \Delta^T)$, be aware of that the surface
of a $d$-dimensional ball with radius $r$ is

$$2 \frac{\pi^{d/2} r^{d-1}}{\Gamma(d/2)} = a_d \times r^{d-1}$$

where $a_d = 2\pi^{d/2}/\Gamma(d/2)$. Now, we have sufficient tools to derive these expectations. To be convenient, let $x = \Delta$ and $\tilde{x}_j = (x_{j+1}, \ldots, x_d)^T$. First, we evaluate $E(\Delta_1) = \int_{[x^T x = 1]} \frac{x_1}{a_d} \, dx = \frac{1}{a_d} \int_{x_1 \in [-1,1]} x_1 \int_{[x_1^2 x_1 = 1 - x_1^2]} \frac{x_1 x_2}{a_d} \, dx = \frac{a_d-1}{a_d} \int_{-1}^{1} x_1 (1 - x_1^{2})^{(d-2)/2} \, dx_1$.

Note that the integrand above is an odd function so $E(\Delta_1) = 0$ and thus, $E(\Delta) = 0$. Similarly, we have

$$\int_{[x^T x = 1]} \frac{x_1 x_2}{a_d} \, dx = \frac{1}{a_d} \int_{x_1^2 + x_2^2 \leq 1} \int_{[x_1^2 x_2 = 1 - x_1^2 - x_2^2]} \frac{x_1 x_2}{a_d} \, dx = \frac{a_d-2}{a_d} \int_{[x_1^2 + x_2^2 \leq 1]} x_1 x_2 (1 - x_1^2 - x_2^2)^{(d-3)/2} \, dx_1 dx_2.$$

Let $x_1 = r \cos \theta$ and $x_2 = r \sin \theta$ where $r \in [0,1]$ and $\theta \in [0,2\pi]$. Then,

$$\int_{[x^T x = 1]} \frac{x_1 x_2}{a_d} \, dx = \frac{a_d-2}{a_d} \int_{0}^{2\pi} \cos \theta \sin \theta d\theta \int_{0}^{1} r^3 (1 - r^2)^{(d-3)/2} \, dr = 0.$$

This provides that $E(\Delta_i \Delta_j) = 0$ for all $i \neq j$. Next, with the same trick,

$$\int_{[x^T x = 1]} \frac{x_1^2}{a_d} \, dx = \frac{a_d-1}{a_d} \int_{-1}^{1} x_1^2 (1 - x_1^2)^{d/2-1} \, dx_1 = \frac{a_d-1}{a_d} \int_{0}^{1} u^{3/2-1} (1 - u)^{d/2-1} \, du$$

$$= \frac{a_d-1}{a_d} \frac{\Gamma(3/2)\Gamma(d/2)}{\Gamma(d+3/2)} = \frac{2}{d^2 - 1} \left( \frac{\Gamma(d/2)}{\Gamma((d-1)/2)} \right)^2 = m_2$$

when $d \geq 2$. It turns out that using random direction on a unit sphere yields $({\xi}_2^2, \nu) = (m_2, 1)$. Note that $m_2$ is a function of dimension $d$. In order to match the notation in Chapter 4, let $\xi_2^2(d) = m_2$. As we claimed that $\xi_2^2(d)/\nu^2$ is always less than 1, we demonstrate this fact in Figure B.1 by showing that $\xi_2^2(d) < 1$ for $d > 1$.

Next, we derive some quantities required in Chapter 4, $m_{22} = E(\Delta_1^2 \Delta_2^2)$ and
Figure B.1: The $\xi^2_s$ function

\[
m_4 = E\Delta^4_1.
\]

\[
m_{22} = \int_{[x^T x = 1]} \frac{x_1^2 x_2^2}{a_d} dx = \frac{a_{d-2}}{a_d} \int_{[x_1^2 + x_2^2 \leq 1]} x_1^2 x_2^2 (1 - x_1^2 - x_2^2)^{(d-3)/2} dx_1 dx_2
\]

\[
= \frac{a_{d-2}}{a_d} \int_0^1 r^5 (1 - r^2)^{(d-3)/2} \int_0^{2\pi} \cos^2 \theta \sin^2 \theta d\theta
\]

\[
= \frac{a_{d-2}}{a_d} \times \frac{\Gamma(3) \Gamma \left( \frac{d-1}{2} \right)}{2 \Gamma \left( \frac{d-1}{2} + 3 \right)} \times \frac{\pi}{4} = \frac{d - 2}{8} \times \frac{1}{(\frac{d-1}{2} + 2)(\frac{d-1}{2} + 1)(\frac{d-1}{2})}
\]

and

\[
m_4 = \int_{[x^T x = 1]} \frac{x_1^4}{a_d} dx = \frac{a_{d-1}}{a_d} \int_{x_1^2 \leq 1} x_1^4 (1 - x_1^2)^{d/2-1} dx_1
\]

\[
= \frac{a_{d-1}}{a_d} \int_0^1 u^{3/2} (1 - u)^{d/2-1} du = \frac{3}{4} \frac{\left[ \Gamma \left( \frac{d}{2} \right) \right]^2}{\Gamma \left( \frac{d-1}{2} \right) \Gamma \left( \frac{d+5}{2} \right)}
\]

where $u = x_1^2$. 
Additionally, notice that $E X_1 X_2 = E(X_1 E(X_2|X_1))$ and

$$E(X_2|X_1 = x_1, X^T X = 1) = E(X_2|\bar{X}_1^T \bar{X}_1 = 1 - x_1^2)$$

$$= \int_{|\bar{x}_1^T \bar{x}_1 - x_1^2|} \frac{x_2}{a_{d-1}} \, d\bar{x}_1 = \frac{a_{d-2}}{a_{d-1}} \int_{\sqrt{1-x_1^2}}^{\sqrt{1-x_1^2}} x_2 (1 - x_1^2 - x_2^2)^{(d-3)/2} \, dx_2 = 0$$

because the integrand is an odd function in terms of $x_2$. Similarly, we can infer that $E(\prod_{j=1}^{d} X_j^{k_j})$ is zero if at least one $k_j$’s is odd.
Appendix C

Properties of Working Matrices

The working matrix is defined as $A = I - sp^{T}B$ where $B$ is positive definite and $\lambda_1^B = 1$. In this appendix, we derive the largest eigenvalue of $A$ which is always larger than 1. Nevertheless, we also show that the Frobenius norm of products of random matrix $A_t$ converges to 0 under some condition on the step size.

**Proposition 7.** Let $B_d \times d$ be a symmetric, full rank and positive definite matrix. Thus, for any $x \in \mathbb{R}^d$ satisfying $x^T x = 1$, we have $x^T B^2 x \geq (x^T B x)^2$.

**Proof.** Since $B$ has full rank, its spectral decomposition gives $B = \Gamma \Lambda \Gamma^T$ with $\Lambda = \text{diag}\{\lambda_1, ..., \lambda_d\}$, $\lambda_i > 0$ for all $i = 1, ..., d$. Besides, since span $= \{\Gamma_1, ..., \Gamma_d\} = \mathbb{R}^d$, the whole $d$--dimensional real space, for any valid $x$, we can find a vector $a \in \mathbb{R}^d$ such that $a = \Gamma x$. So we have $a^T a = x^T x = 1$. Moreover, we have $x^T B x = a^T \Gamma^T \Gamma \Lambda \Gamma^T \Gamma a = \sum_{i=1}^{d} a_i^2 \lambda_i$ and $x^T B^2 x = a^T \Gamma^T \Gamma \Lambda^2 \Gamma^T \Gamma a = \sum_{i=1}^{d} a_i^2 \lambda_i^2$.

Now, since $\sum_i a_i^2 = 1$, a valid probability mass, we can treat above equations like $E_x(\lambda)$ and $E_x(\lambda^2)$. But, we know $0 \leq \text{Var}_x(\lambda) = E_x(\lambda^2) - \{E_x(\lambda)\}^2$.

**Proposition 8.** The eigenvalues of $A = I - sp^{T}B$ are $(d - 1)$ multiples of 1 and $1 - sT_1$ where $T_1 = \rho^T B \rho$. Further, suppose the spectral decomposition of $A$ is $\Gamma \Lambda \Gamma^{-1}$. Then,
1. The corresponding eigenvector of eigenvalues \(1 - sT\) is \(\rho\), i.e. \(\Gamma_d = \rho\).

2. The last row of \(\Gamma^{-1}\) is \(h = B\rho/T_1\).

Proof. Let \(D_1 = \text{span}\{B\rho\}\). Suppose \(v \in D_1^\perp\) then
\[
Av = v - s\rho^T Bv = v - 0.
\]
Besides,
\[
A\rho = \rho - s\rho^T B\rho = (1 - sT_1)\rho.
\]
Note that \(h\) should be orthogonal to all \(v \in D_1^\perp\). Thus, \(h\) should belong to \(D_1\), i.e. \(h = a \times B\rho\) for some constant \(a\). In addition, \(\rho^T h = 1\) so \(a^{-1} = \rho^T B\rho = T_1\). \(\square\)

**Proposition 9.** Let \(A\) be the matrix introduced above. The eigenvalues of \(A^T A\) are \((d - 2)\) multiples of \(1\), \(r^+ = \eta(s) + \sqrt{C_\psi(s)}/2\) and \(r^- = \eta(s) - \sqrt{C_\psi(s)}/2\) where \(C_\psi(s)\) is defined as \((C.1)\) and \(\eta(s)\) as \((C.2)\). Moreover, \(0 \leq r^- \leq 1 \leq r^+\) for all \(s > 0\) and \(r^+ = 1\) whenever one of the following conditions holds: 1) \(s = 0\); 2) \(s \in [0, 2T_1/T_2]\) and \(B = \lambda I\), \(\lambda > 0\).

Proof. Note that
\[
A^T A = I - \frac{s}{d} \Delta^T B - \frac{s}{d} B \Delta^T + \frac{s^2}{d} B \Delta \Delta^T B = I - s\rho^T B - sB\rho^T + s^2 B\rho^T B
\]
where \(\rho = \Delta/\sqrt{d}\). Now we have \(\rho^T \rho = 1\). Define the space \(D_2 = \text{span}\{B\rho, \rho\}\). Then, for any vector \(c \in \mathbb{R}^d\) which is perpendicular to \(D_2\), we have \(c^T B\rho = 0\), \(c^T \rho = 0\) and hence
\[
A^T Ac = c - s\rho (\rho^T Bc) - sB\rho (\rho^T c) + s^2 B\rho (\rho^T Bc) = c.
\]
Moreover, there exists \(d - 2\) such \(c\)’s which are orthogonal one another and their corresponding eigenvalues are all equal to \(1\). In this sequel, the remaining \(2\) eigenvectors lie in \(D_2\). Let \(e\) denote an eigenvector lying on \(D_2\) with corresponding eigenvalue \(r\), i.e. \(A^T A e = re\). In addition, such an \(e\) can be expressed as
\[
e = \begin{bmatrix} B\rho & \rho \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}
\]
for some \( a, b \in \mathbb{R} \). Consequently,

\[
A^T A e = e - s \rho (\rho^T B e) - s B \rho (\rho^T e) + s B \rho (\rho^T Be).
\]

In addition, we have

\[
\rho^T Be = a \rho^T B^2 \rho + b \rho^T B \rho = a T_2 + b T_1
\]

\[
\rho^T e = a \rho^T B \rho + b \rho^T \rho = a T_1 + b
\]

which yield

\[
A^T A e = [B \rho \ \rho] \begin{bmatrix}
a + s^2 (a T_2 + b T_1) - s (a T_1 + b) \\
b - s (a T_2 + b T_1)
\end{bmatrix} = [B \rho \ \rho] \begin{bmatrix}
ra \\
r b
\end{bmatrix}.
\]

Accordingly, we have following equations

\[
\begin{cases}
(1 - r) a + s^2 (a T_2 + b T_1) - s (a T_1 + b) = 0 \\
(1 - r) b - s (a T_2 + b T_1) = 0
\end{cases}
\]

which implies that \( 1 - r \) should be the root of

\[
\psi(1 - r) = (1 - r)^2 + (1 - r)(s^2 T_2 - 2 s T_1) + s^2 (T_1^2 - T_2).
\]

The condition function of \( \psi \) is

\[
C_\psi(s) = s^2 \left\{ (s T_2 - 2 T_1)^2 + 4(T_2 - T_1^2) \right\} = s^2(s^2 T_2^2 - 4 s T_1 T_2 + 4 T_2). \tag{C.1}
\]

Because of \( T_2 \geq T_1^2 \), by Proposition 7, \( C_\psi(s) \geq 0 \) for all \( s \in \mathbb{R}^+ \). Therefore, the resulting eigenvalues are \( r^+ = \eta(s) + \sqrt{C_\psi(x)/2} \) and \( r^- = \eta(s) - \sqrt{C_\psi(x)/2} \) where

\[
\eta(s) = 1 - s T_1 + s^2 T_2/2. \tag{C.2}
\]

However, we hope that \( 0 \leq r^- \leq r^+ \leq 1 \). It is easy to show that \( C_\psi(s) \leq 4 \eta^2(s) \) for all \( s > 0 \). Unfortunately, \( C_\psi(s) \leq 4(1 - \eta(s))^2 \) holds (with equality) only if \( s = 0 \) or \([T_1^2 = T_2] \cap [0 < s < 2T_1/T_2] \). But \( C_\psi(s) \geq 4(1 - \eta(s))^2 \) for all \( s > 0 \). Thus, we have \( 0 \leq r^- \leq 1 \leq r^+ \). Note that the condition \( T_1^2 = T_2 \) implies that \( \text{Var}(\lambda) = 0, \) \( i.e. \) all \( \lambda \)'s are identical which implies condition 2. So we proved. \( \square \)
Corollary 1. If $B = \lambda \times I$ then eigenvalues of $A^T A$ are $(d - 1)$ multiples of 1 and $(1 - \lambda s)^2$ providing $0 \leq \lambda s \leq 2$.

Proposition 10.

$$(d - 1) < \|A^T A\|_F^2 \leq d$$

for any $s \in [0, 1]$.

Proof. With the fact that

$$(r^+ + r^-)(s) = 2\eta(s) = 2 - 2sT_1 + s^2T_2 = \left(s - \frac{T_1}{T_2}\right)^2 - \left(\frac{T_1}{T_2} - 2\right)$$

we know that the minium of function $(r^+ + r^-)(s)$ is between 1 and 2. So we proved.

Proposition 11. Define $D_t = A_{t-1} \times \cdots \times A_0$. The Frobenius norm of sequence $\{\prod_{j=0}^{t} A_j\}$ is a monotone decreasing sequence providing

$$\rho^T D_t D_t^T B \rho \geq \frac{s}{2} \rho^T B D_t D_t^T B \rho.$$  \hspace{1cm} (C.3)

for some small $s > 0$ and for all $t \in \mathbb{N}$.

Proof. Note that $\prod_{j=0}^{t} A_j = A_t A_{t-1} \cdots A_0 = A_t D$ and also the Frobenius norm of $A_t D$ is sum of the Euclidean norm of vector $A_t D_j$’s. Suppose the spectral decomposition of $A_t$ is defined as $A_t = \Gamma \Lambda \Gamma^{-1}$. By Proposition 8, $\Lambda = diag\{1, \ldots, 1, r\}$ where $0 < r = 1 - sT_1 \leq 1$ for some small $s > 0$. Define $X = \Gamma^{-1} D$, $\zeta_i = (0, \ldots, 0, X_{id})^T$, 

$$\cdots$$
and $S = \Gamma^T \Gamma$. Then,

$$\|A_iD\|_F^2 = \text{tr}(D^T \Gamma^{-T} \Lambda \Gamma^T \Gamma \Gamma^{-1} D) = \sum_{j=1}^d X_j^T \Lambda S \Lambda X_j$$

$$= \sum_{j=1}^d (X_j - \zeta_j \times (1 - r))^T S (X_j - \zeta_j \times (1 - r))$$

$$= \|D\|_F^2 - 2(1 - r) \sum_{j=1}^d X_j^T \zeta_j + (1 - r)^2 \sum_{j=1}^d X_j^2 S_{dd}$$

$$= \|D\|_F^2 - 2(1 - r) \sum_{j=1}^d \rho^T D_i D_i^T \Gamma^{-T} e_d + (1 - r)^2 S_{dd} e_d^T \Gamma^{-1} D D^T \Gamma^{-T} e_d$$

$$= \|D\|_F^2 - 2(1 - r) \rho^T DD^T h + (1 - r)^2 S_{dd} h^T DD^T h = \|D\|_F^2 - \epsilon.$$

Thus, the condition that $\|A_iD\|_F \leq \|D\|_F$ is $\epsilon \geq 0$. So the conclusion holds as long as $2\rho^T DD^T h - (1 - r) h^T DD^T h \geq 0$. Note that, in our case, $S_{dd} = 1$ and $h = B \rho / T_1$.

After some algebra, we have the condition (C.3).

**Proposition 12.** Suppose $S$ and $W$ are two matrices in $\mathbb{R}^{d \times d}$ where $S$ is symmetric but $W$ may be not. Then,

$$\text{tr}(SW) \leq \text{tr}(S) \text{tr}(W)$$

providing $\lambda_1^S \leq \text{tr}(S)$ and $\lambda_1^W \geq 0$.

**Proof.** Denote the spectral decomposition of $S$ and $W$ as $\Gamma_S \Lambda_S \Gamma_S^T$ and $\Gamma_W \Lambda_W \Gamma_W^{-1}$.

$$\text{tr}(SW) = \text{tr}(\Gamma_S \Lambda_S \Gamma_S^T W) = \text{tr}(\Lambda_S \Gamma_S^T W \Gamma_S) = \sum_{j=1}^d \lambda_j^S \Gamma_j^T \Gamma_j W \Gamma_S$$

$$\leq \lambda_1^S \text{tr}(\Gamma_S^T W \Gamma_S) \leq \text{tr}(S) \times \text{tr}(W) \text{tr}(\Gamma_S^T \Gamma_S) = \text{tr}(S) \times \text{tr}(W).$$

**Corollary 2.** Suppose $\lambda_1^B \leq \text{tr}(B) = k$, $k > 0$. Then, $\rho^T BD_i D_i^T B \rho \leq k \times \rho^T D_i D_i^T B \rho$, i.e. (C.3) holds with probability 1, for all $t \in \mathcal{N}$ and $s \leq 2/k$.

**Proof.** According to Proposition 12, we have

$$0 \leq \rho^T BD_i D_i^T B \rho = \text{tr}(BD_i D_i^T B \rho \rho^T) \leq \text{tr}(B) \text{tr}(\rho^T D_i D_i^T B \rho) = k \times \rho^T D_i D_i^T B \rho$$
if the smallest eigenvalue of matrix $D_tD_T^TB\rho\rho^T$ is positive. This is true because

$$
\rho^T H B \rho = \rho^T \Gamma_H \Lambda_H \Gamma_H^T G \rho = \text{tr}(\Lambda_H \Gamma_H^T B \rho \rho^T \Gamma_H) \geq \lambda_d^H \sum_{j=1}^d \Gamma_{Hj}^T B \rho \rho^T \Gamma_{Hj} \\
= \lambda_d^H \sum_{j=1}^d \rho^T \Gamma_{Hj} \Gamma_{Hj}^T B \rho = \lambda_d^H \rho^T \Gamma_H \Gamma_H^T B \rho = \lambda_d^H T_1 \geq 0
$$

where $H = D_tD_t^T$ and thus, $\lambda_d^H \geq 0$. So we showed.