ISSN: 2161-718X Vol. 5, No. 1, February 2015

Scientific
Research
Publishing

Open Journal of Statistics

CHART
02
SECTION
SECTION
SECTION
SECTION
LOREM IPSUM DOLOR SIT AMET, CONSECTETUR ADIPISCING ELIT.
01
34%
LOREM IPSUM DOLOR SIT AMET, CONSECTETUR ADIPISCING ELIT.
02
26%
LOREM IPSUM DOLOR SIT AMET, CONSECTETUR ADIPISCING ELIT.
03
48%
LOREM IPSUM DOLOR SIT AMET, CONSECTETUR ADIPISCING ELIT.
04
57%
A>B
78%

ISSN: 2161-718X
9 772161 718001 01

www.scirp.org/journal/ojs

Journal Editorial Board

ISSN 2161-718X (Print)   ISSN 2161-7198 (Online)
http://www.scirp.org/journal/ojs

Open Journal of Statistics, 2015, 5, 1-95
Published Online February 2015 in SciRes. http://www.scirp.org/journal/ojs

Table of Contents

Volume 5 Number 1 February 2015

Open Journal of Statistics (OJS)

Journal Information

SUBSCRIPTIONS

The *Open Journal of Statistics* (Online at Scientific Research Publishing, www.SciRP.org) is published bimonthly by Scientific Research Publishing, Inc., USA.

Subscription rates:

Print: $69 per issue.
To subscribe, please contact Journals Subscriptions Department, E-mail: sub@scirp.org

SERVICES

Advertisements
Advertisement Sales Department, E-mail: service@scirp.org

Reprints (minimum quantity 100 copies)
Reprints Co-ordinator, Scientific Research Publishing, Inc., USA.
E-mail: sub@scirp.org

COPYRIGHT

COPYRIGHT AND REUSE RIGHTS FOR THE FRONT MATTER OF THE JOURNAL:

COPYRIGHT FOR INDIVIDUAL PAPERS OF THE JOURNAL:

REUSE RIGHTS FOR INDIVIDUAL PAPERS:

Note: At SCIRP authors can choose between CC BY and CC BY-NC. Please consult each paper for its reuse rights.

DISCLAIMER OF LIABILITY

Statements and opinions expressed in the articles and communications are those of the individual contributors and not the statements and opinion of Scientific Research Publishing, Inc. We assume no responsibility or liability for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained herein. We expressly disclaim any implied warranties of merchantability or fitness for a particular purpose. If expert assistance is required, the services of a competent professional person should be sought.

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:
E-mail: ojs@scirp.org

Open Journal of Statistics, 2015, 5, 1-9
Published Online February 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.51001

Estimation of Population Ratio in Post-Stratified Sampling Using Variable Transformation

Aloy Chijioke Onyeka, Chinyeaka Hostensia Izunobi, Iheanyi Sylvester Iwueze

Department of Statistics, Federal University of Technology, Owerri, Nigeria
Email: aloyonyeka@futo.edu.ng, chiyeaka2007@yahoo.com, isiwueze@yahoo.com

Received 28 December 2014; accepted 16 January 2015; published 20 January 2015

Abstract

Extending the work carried out by [1], this paper proposes six combined-type estimators of population ratio of two variables in post-stratified sampling scheme, using variable transformation. Properties of the proposed estimators were obtained up to first order approximations, $o\left(n^{-1}\right)$, both for achieved sample configurations (conditional argument) and over repeated samples of fixed size n (unconditional argument). Efficiency conditions were obtained. Under these conditions the proposed combined-type estimators would perform better than the associated customary combined-type estimator. Furthermore, optimum estimators among the proposed combined-type estimators were obtained both under the conditional and unconditional arguments. An empirical work confirmed the theoretical results.

Keywords

Variable Transformation, Combined-Type Estimator, Ratio, Product and Regression-Type Estimators, Mean Squared Error

1. Introduction

The use of information on auxiliary character to improve estimates of population parameters of the study variable is a common practice in sample survey, and sometimes, information on several variables is used to estimate or predict a characteristic of interest. The investigators often collect observations from more than one variable, including the variable of interest y and some auxiliary variables x. The use of these variables (known as auxiliary information in sample survey design) often results in efficient estimate of population parameters (e.g.

How to cite this paper: Onyeka, A.C., Izunobi, C.H. and Iwueze, I.S. (2015) Estimation of Population Ratio in Post-Stratified Sampling Using Variable Transformation. *Open Journal of Statistics*, **5**, 1-9. http://dx.doi.org/10.4236/ojs.2015.51001

mean, ratio, proportion, etc.) under some realistic conditions, especially when there is a strong correlation between the study variables and the auxiliary variables. Many authors have made contributions in this regard, including [2] and [3]. In this context, ratio, product and regression methods of estimation are good examples. Ratio and product-type estimators take advantage of the correlation between the auxiliary variable and the study variable, to improve the estimate of the characteristic of interest. For example, when information is available on the auxiliary variable that is highly positively correlated with the study variable, the ratio method of estimation proposed by [4] is a suitable estimator to estimate the population mean, and when the correlation is negative, the product method of estimation, as envisaged by [5] and [6], is appropriate. However, in some studies, the ratio of the population means (or totals) of the study and auxiliary variables might be of great significance, hence the need to estimate such ratios.

The customary estimator of the population ratio $\left(R=\bar{Y}/\bar{X}\right)$ of the population means of two variables, y and x, under the simple random sampling scheme, is given as $\hat{R}=\bar{y}/\bar{x}$, which is the ratio of the sample means of the two variables ([2] and [7]). The estimator, $\hat{R}=\bar{y}/\bar{x}$, uses information on only two variables, namely the study variable (y) and one auxiliary variable (x). However, several authors, like [7] and [8], have contributed to the problem of estimating the population ratio of two means, often utilizing additional information on one or more auxiliary variables, say $z_i\left(i=1,2,\cdots\right)$. While it is possible to record increased efficiency by introducing such additional auxiliary variables, it is obvious that extra cost is involved in order to obtain information on such additional auxiliary variables. References [1] and [9] have argued that such extra cost could be avoided by using variable transformation of the already observed auxiliary variable, instead of introducing additional (new) auxiliary variables. However, the works carried out by [1] [9] were restricted to estimation of population ratio in simple random sampling scheme. The present study is necessitated by the need to extend to post-stratified sampling scheme, the works on ratio estimation carried out by [1] [9] under the simple random sampling scheme. This is in order to extend to other sampling schemes, the obvious advantage of reduced cost in the use of variable transformation instead of introducing additional (new) auxiliary variables when estimating population ratio of two population parameters.

2. The Proposed Combined-Type Estimators

Let n units be drawn from a population of N units using simple random sampling method and let the sampled units be allocated to their respective strata, where n_h is the number of units that fall into stratum h such that $\sum_{h=1}^{L} n_h = n$. Let y_{hi} and x_{hi} be the i^{th} observation on the study and auxiliary variables, respectively. Consider the following variable transformation of the auxiliary variable, x, under post-stratified sampling scheme.

$$x_{hi}^{*}=\frac{N\bar{X}-nx_{hi}}{N-n},\quad h=1,2,\cdots,L \text{ and } i=1,2,\cdots,N \tag{2.1}$$

An equivalent of the transformation (2.1), in simple random sampling scheme, has been used by authors like [1] [8]-[13]. The associated sample mean estimator of the transformed variable (2.1), in post-stratified sampling scheme, can be written as

$$\bar{x}_{ps}^{*}=\left(1-\pi\right)\bar{X}-\pi\bar{x}_{ps},\quad \text{where}\quad \pi=\frac{n}{N-n} \tag{2.2}$$

and $\bar{x}_{ps}=\sum_{h=1}^{L}\omega_h\bar{x}_h$ and $\bar{y}_{ps}=\sum_{h=1}^{L}\omega_h\bar{y}_h$ are sample mean estimators based on x_{hi} and y_{hi} respectively. Using the sample means $\bar{y}_{ps}$, $\bar{x}_{ps}$ and $\bar{x}_{ps}^{*}$, and assuming that the population mean, $\bar{X}$ of the auxiliary variable x_{hi}, is known, we proposed six combined-type estimators of the population ratio $R=\bar{Y}/\bar{X}$ in post stratified sampling scheme as

$$\hat{R}_{1c}=\frac{\bar{y}_{ps}}{\bar{x}_{ps}-b\left(\bar{x}_{ps}^{*}-\bar{X}\right)} \tag{2.3}$$

$$\hat{R}_{2c}=\frac{\bar{y}_{ps}}{\left(\frac{\bar{x}_{ps}}{\bar{x}_{ps}^{*}}\bar{X}\right)}=\frac{\bar{y}_{ps}\bar{x}_{ps}^{*}}{\bar{x}_{ps}\bar{X}} \tag{2.4}$$

$$\hat{R}_{3c}=\frac{\bar{y}_{ps}}{\left(\frac{\bar{x}_{ps}\bar{x}_{ps}^{*}}{\bar{X}}\right)}=\frac{\bar{y}_{ps}\bar{X}}{\bar{x}_{ps}\bar{x}_{ps}^{*}} \tag{2.5}$$

$$\hat{R}_{4c}=\frac{\bar{y}_{ps}}{\bar{x}_{ps}^{*}} \tag{2.6}$$

$$\hat{R}_{5c}=\frac{\bar{y}_{ps}}{\bar{x}_{ps}^{*}-b\left(\bar{x}_{ps}-\bar{X}\right)} \tag{2.7}$$

$$\hat{R}_{6c}=\frac{\bar{y}_{ps}}{\left(\frac{\bar{x}_{ps}^{*}}{\bar{x}_{ps}}\bar{X}\right)}=\frac{\bar{y}_{ps}\bar{x}_{ps}}{\bar{x}_{ps}^{*}\bar{X}}. \tag{2.8}$$

2.1. Conditional Properties of the Proposed Estimators

Reference [14] defined that under the conditional argument, that is, for the achieved sample configuration, $\underline{n}=\left(n_1,n_2,n_3,\cdots,n_L\right)$ the post stratified estimator, $\bar{y}_{ps}$ is unbiased for the population mean, $\bar{Y}$, with variance

$$V_2\left(\bar{y}_{ps}\right)=\sum_{h=1}^{L}\omega_h^2\left(1-f_h\right)\frac{S_{yh}^2}{n_h}=\sum_{h=1}^{L}\frac{\omega_h^2S_{yh}^2}{n_h}-\frac{1}{N}\sum_{h=1}^{L}\omega_hS_{yh}^2 \tag{2.9}$$

where V_2 refers to conditional variance and S_{yh}^2 is the population variance of y in stratum h. Similarly, Onyeka (2012) obtained the conditional variance of $\bar{x}_{ps}$ and the conditional covariance of $\bar{y}_{ps}$ and $\bar{x}_{ps}$ respectively as:

$$V_2\left(\bar{x}_{ps}\right)=\sum_{h=1}^{L}\omega_h^2\left(1-f_h\right)\frac{S_{xh}^2}{n_h}=\sum_{h=1}^{L}\frac{\omega_h^2S_{xh}^2}{n_h}-\frac{1}{N}\sum_{h=1}^{L}\omega_hS_{xh}^2 \tag{2.10}$$

and

$$C_2\left(\bar{y}_{ps},\bar{x}_{ps}\right)=\sum_{h=1}^{L}\omega_h^2\left(1-f_h\right)\frac{S_{yxh}}{n_h}=\sum_{h=1}^{L}\frac{\omega_h^2S_{yxh}}{n_h}-\frac{1}{N}\sum_{h=1}^{L}\omega_hS_{yxh} \tag{2.11}$$

where S_{xh}^2 is the population variance of x in stratum h, S_{yxh} is the covariance of y and x in stratum h, and C_2 refers to conditional covariance.

Let

$$e_0=\frac{\bar{y}_{ps}-\bar{Y}}{\bar{Y}}\quad\text{and}\quad e_1=\frac{\bar{x}_{ps}-\bar{X}}{\bar{X}}. \tag{2.12}$$

Then, under the conditional argument,

$$E_2\left(e_0\right)=E_2\left(e_1\right)=0 \tag{2.13}$$

$$E_2\left(e_0^2\right)=\frac{V_2\left(\bar{y}_{ps}\right)}{\bar{Y}^2}=\frac{1}{\bar{Y}^2}\sum_{h=1}^{L}\omega_h^2\left(1-f_h\right)\frac{S_{yh}^2}{n_h} \tag{2.14}$$

$$E_2\left(e_1^2\right)=\frac{V_2\left(\bar{x}_{ps}\right)}{\bar{X}^2}=\frac{1}{\bar{X}^2}\sum_{h=1}^{L}\omega_h^2\left(1-f_h\right)\frac{S_{xh}^2}{n_h} \tag{2.15}$$

$$E_2\left(e_0e_1\right)=\frac{C_2\left(\bar{y}_{ps},\bar{x}_{ps}\right)}{\bar{Y}\bar{X}}=\frac{1}{\bar{Y}\bar{X}}\sum_{h=1}^{L}\omega_h^2\left(1-f_h\right)\frac{S_{yxh}}{n_h}. \tag{2.16}$$

Using (2.12), the first proposed estimator, $\hat{R}_{1C}$, given in (2.3), can be re-written up to first order approximation, $o\left(n^{-1}\right)$, in expected value, as

$$\left(\hat{R}_{1c}-R\right)=R\left[e_0-\left(1+b\pi\right)e_1-\left(1+b\pi\right)e_0e_1+\left(1+b\pi\right)^2e_1^2\right] \tag{2.17}$$

and

$$\left[\hat{R}_{1c}-R\right]^2=R^2\left[e_0^2+\left(1+b\pi\right)^2e_1^2-2\left(1+b\pi\right)e_0e_1\right]. \tag{2.18}$$

We take conditional expectation of (2.17) and (2.18), and use (2.13) to (2.16) to make the necessary substitutions. This gives the conditional bias and mean square error of $\hat{R}_{1C}$ respectively as

$$B_2\left(\hat{R}_{1C}\right)=\frac{1}{\bar{X}^2}\left[\left(1+b\pi\right)^2RA_{22}-\left(1+b\pi\right)A_{12}\right] \tag{2.19}$$

and

$$\text{MSE}_2\left(\hat{R}_{1C}\right)=\frac{1}{\bar{X}^2}\left[A_{11}+\left(1+b\pi\right)^2R^2A_{22}-2\left(1+b\pi\right)RA_{12}\right] \tag{2.20}$$

where

$$A_{11}=\sum_{h=1}^{L}\frac{\omega_h^2\left(1-f_h\right)S_{yh}^2}{n_h},\quad A_{22}=\sum_{h=1}^{L}\frac{\omega_h^2\left(1-f_h\right)S_{xh}^2}{n_h},\quad A_{12}=\sum_{h=1}^{L}\frac{\omega_h^2\left(1-f_h\right)S_{yxh}}{n_h}. \tag{2.21}$$

Following similar procedure, we obtain the conditional biases and mean square errors of the six proposed estimators, together with those of the customary combined-type estimator, $\hat{R}_C=\bar{y}_{ps}/\bar{x}_{ps}$, of population ratio $\left(R\right)$, in post-stratified sampling, up to first order approximation, $o\left(n^{-1}\right)$, as:

$$B_2\left(\hat{R}_C\right)=\frac{1}{\bar{X}^2}\left[RA_{22}-A_{12}\right] \tag{2.22}$$

$$B_2\left(\hat{R}_{1C}\right)=\frac{1}{\bar{X}^2}\left(1+b\pi\right)\left[\left(1+b\pi\right)\ RA_{22}-A_{12}\right] \tag{2.23}$$

$$B_2\left(\hat{R}_{2C}\right)=\frac{1}{\bar{X}^2}\left(1+\pi\right)\left[\left(1+\pi\right)\ RA_{22}-A_{12}\right] \tag{2.24}$$

$$B_2\left(\hat{R}_{3C}\right)=\frac{1}{\bar{X}^2}\left[\left(1-\pi+\pi^2\right)\ RA_{22}-\left(1-\pi\right)A_{12}\right] \tag{2.25}$$

$$B_2\left(\hat{R}_{4C}\right)=\frac{1}{\bar{X}^2}\left[\pi^2RA_{22}+\pi A_{12}\right] \tag{2.26}$$

$$B_2\left(\hat{R}_{5C}\right)=\frac{1}{\bar{X}^2}\left(\pi+b\right)\left[\left(\pi+b\right)RA_{22}+A_{12}\right] \tag{2.27}$$

$$B_2\left(\hat{R}_{6C}\right)=\frac{1}{\bar{X}^2}\left(1+\pi\right)\left[\pi RA_{22}+A_{12}\right] \tag{2.28}$$

and

$$\text{MSE}_2\left(\hat{R}_C\right)=\frac{1}{\bar{X}^2}\left[A_{11}+R^2A_{22}-2RA_{12}\right] \tag{2.29}$$

$$\text{MSE}_2\left(\hat{R}_{1C}\right)=\frac{1}{\bar{X}^2}\left[A_{11}+\left(1+b\pi\right)^2R^2A_{22}-2\left(1+b\pi\right)RA_{12}\right] \tag{2.30}$$

$$\mathrm{MSE}_2\left(\hat{R}_{2C}\right)=\frac{1}{\bar{X}^2}\left[A_{11}+(1+\pi)^2R^2A_{22}-2(1+\pi)RA_{12}\right] \tag{2.31}$$

$$\mathrm{MSE}_2\left(\hat{R}_{3c}\right)=\frac{1}{\bar{X}^2}\left[A_{11}+(1-\pi)^2R^2A_{22}-2(1-\pi)RA_{12}\right] \tag{2.32}$$

$$\mathrm{MSE}_2\left(\hat{R}_{4C}\right)=\frac{1}{\bar{X}^2}\left[A_{11}+\pi^2R^2A_{22}+2\pi RA_{12}\right] \tag{2.33}$$

$$\mathrm{MSE}_2\left(\hat{R}_{5C}\right)=\frac{1}{\bar{X}^2}\left[A_{11}+(\pi+b)^2R^2A_{22}+2(\pi+b)RA_{12}\right] \tag{2.34}$$

$$\mathrm{MSE}_2\left(\hat{R}_{6C}\right)=\frac{1}{\bar{X}^2}\left[A_{11}+(1+\pi)^2R^2A_{22}+2(1+\pi)RA_{12}\right]. \tag{2.35}$$

Generally, we have for the proposed six combined-type estimators,

$$\mathrm{MSE}_2\left(\hat{R}_{qc}\right)=\frac{1}{\bar{X}^2}\left[A_{11}+\theta_q^2R^2A_{22}-2\theta_qRA_{12}\right] \tag{2.36}$$

where $q=1,\cdots,6$ and

$$\theta_1=(1+b\pi),\quad \theta_2=(1+\pi),\quad \theta_3=(1-\pi),\quad \theta_4=-\pi,\quad \theta_5=-(\pi+b),\quad \theta_6=-(1+\pi). \tag{2.37}$$

2.2. Unconditional Properties of the Proposed Estimators

Following [14] we obtain the following (unconditional) variances and covariance, for repeated samples of fixed size *n*.

$$V\left(\bar{y}_{ps}\right)=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_hS_{yh}^2 \tag{2.38}$$

$$V\left(\bar{x}_{ps}\right)=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_hS_{xh}^2 \tag{2.39}$$

and

$$\mathrm{Cov}\left(\bar{y}_{ps},\bar{x}_{ps}\right)=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_hS_{yxh} \tag{2.40}$$

where $f=n/N$ is the population sampling fraction. By taking unconditional expectations of (2.17) and (2.18), and using (2.38)-(2.40) to make the necessary substitutions, we obtain the unconditional bias and mean square errors of the first proposed estimator, $\hat{R}_{1c}$, up to first order approximation, $o\left(n^{-1}\right)$, as:

$$B\left(\hat{R}_{1C}\right)=\frac{1}{\bar{X}^2}(1+b\pi)\left(\frac{1-f}{n}\right)\left[(1+b\pi)RA'_{22}-A'_{12}\right] \tag{2.41}$$

and

$$\mathrm{MSE}\left(\hat{R}_{2C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)\left[A'_{11}+(1+\pi)^2R^2A'_{22}-2(1+\pi)RA'_{12}\right] \tag{2.42}$$

where

$$A'_{11}=\sum_{h=1}^{L}\omega_hS_{yh}^2,\quad A'_{22}=\sum_{h=1}^{L}\omega_hS_{xh}^2,\quad A'_{12}=\sum_{h=1}^{L}\omega_hS_{yxh}. \tag{2.43}$$

Following similar procedure, we obtain the unconditional biases and mean square errors of the six proposed estimators, together with those of the customary combined-type estimator, $\hat{R}_C=\bar{y}_{ps}/\bar{x}_{ps}$, of population ratio (R), in post-stratified sampling, up to first order approximation, $o\left(n^{-1}\right)$, as:

$$B\left(\hat{R}_C\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)\left[RA'_{22}-A'_{12}\right] \tag{2.44}$$

$$B\left(\hat{R}_{1C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)(1+b\pi)\left[(1+b\pi)RA'_{22}-A'_{12}\right] \tag{2.45}$$

$$B\left(\hat{R}_{2C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)(1+\pi)\left[(1+\pi)RA'_{22}-A'_{12}\right] \tag{2.46}$$

$$B\left(\hat{R}_{3C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)\left[\left(1-\pi+\pi^2\right) RA'_{22}-(1-\pi)A'_{12}\right] \tag{2.47}$$

$$B\left(\hat{R}_{4C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)\left[R\pi^2 A'_{12}+\pi A'_{12}\right] \tag{2.48}$$

$$B\left(\hat{R}_{5C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)(b+\pi)\left[(b+\pi)RA'_{22}+A'_{12}\right] \tag{2.49}$$

$$B\left(\hat{R}_{6C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)(1+\pi)\left[\pi RA'_{22}+A'_{12}\right] \tag{2.50}$$

and,

$$\mathrm{MSE}\left(\hat{R}_{C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)\left[A'_{11}+R^2A'_{22}-2RA'_{12}\right] \tag{2.51}$$

$$\mathrm{MSE}\left(\hat{R}_{1C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)\left[A'_{11}+R^2(1+\pi b)A'_{22}-2RA'_{12}\right] \tag{2.52}$$

$$\mathrm{MSE}\left(\hat{R}_{2C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)\left[A'_{11}+(1+\pi)^2 R^2A'_{22}-2(1+\pi)RA'_{12}\right] \tag{2.53}$$

$$\mathrm{MSE}\left(\hat{R}_{3C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)\left[A'_{11}+(1-\pi)^2 R^2A'_{22}-2(1-\pi)RA'_{12}\right] \tag{2.54}$$

$$\mathrm{MSE}\left(\hat{R}_{4C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)\left[A'_{11}+\pi^2R^2A'_{22}+2R\pi A'_{12}\right] \tag{2.55}$$

$$\mathrm{MSE}\left(\hat{R}_{5C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)\left[A'_{11}+(b+\pi)^2 R^2A'_{22}+2(b+\pi)RA'_{12}\right] \tag{2.56}$$

$$\mathrm{MSE}\left(\hat{R}_{6C}\right)=\frac{1}{\bar{X}^2}\left(\frac{1-f}{n}\right)\left[A'_{11}+(1+\pi)^2 R^2A'_{22}+2(1+\pi)RA'_{12}\right]. \tag{2.57}$$

Generally, the unconditional mean square errors of the proposed combined-type estimators is obtained as

$$\mathrm{MSE}\left(\hat{R}_{qC}\right)=\frac{1}{\bar{X}^2}\left[\frac{1-f}{n}\right]\left[A'_{11}+\theta_q^2R^2A'_{22}-2\theta_qRA'_{12}\right] \tag{2.58}$$

where θ_q, $q=1,\cdots,6$ is as given in (2.37).

3. Efficiency Comparison

The efficiencies of the six proposed combined-type estimators are first compared with that of the customary combined ratio estimator $\hat{R}_C$ in estimating the population ratio R of two population means under the conditional and unconditional arguments in post-stratified random sampling scheme. Secondly, the performances of the proposed estimators among themselves are investigated. Furthermore, the optimum estimators among the proposed estimators are also obtained. The efficiency comparison is carried out using the mean square errors of the estimators and the results are shown in **Table 1**.

4. Numerical Illustration

Here, we use the final year GPA (y) and the level of absenteeism (x) of 2012/2013 graduating students of

Statistics Department, Federal University of Technology Owerri to illustrate the properties of the estimators proposed in the present study. Absenteeism is measured as the average number of days absent from lectures in a month. The class consists of 50 students, with 32 and 18 students respectively falling into low-absenteeism (0 - 3 days per month) and high-absenteeism (4 - 6 days per month) groups or strata. Our interest is to estimate the ratio of final year GPA to absenteeism from lectures, based on a post-stratified sample of 20 out of the 50 graduating students in the class. The data statistics, consisting mainly of population parameters are shown in **Table 2**.

Table 3 shows the percentage relative efficiencies (PRE-1) of the proposed combined-type estimators, $\hat{R}_{qc}$,

Table 1. Efficiency conditions under conditional and unconditional arguments.

Estimator	Conditional argument	Unconditional argument
R_{qc} is better than R_c if:	1) $\lvert\theta_q\rvert<1$ and $\beta<R$ or 2) $\lvert\theta_q\rvert>1$ and $\beta>R$	1) $\lvert\theta_q\rvert<1$ and $\beta'<R$ or 2) $\lvert\theta_q\rvert>1$ and $\beta'>R$
R_{kc} is better than R_{jc} if:	1) $\lvert\theta_j\rvert<\lvert\theta_k\rvert$ and $\lvert\theta_j\rvert<\beta/R$ or 2) $\lvert\theta_j\rvert>\lvert\theta_k\rvert$ and $\lvert\theta_j\rvert>\beta/R$	1) $\lvert\theta_j\rvert<\lvert\theta_k\rvert$ and $\lvert\theta_j\rvert<\beta'/R$ or 2) $\lvert\theta_j\rvert>\lvert\theta_k\rvert$ and $\lvert\theta_j\rvert>\beta'/R$
R_{qc} is optimum if:	$\lvert\theta_q^0\rvert=\beta/R$	$\lvert\theta_q^0\rvert=\beta'/R$

Where $\beta=A_{12}/A_{22}$, $\beta'=A'_{12}/A'_{22}$ and θ_q , $q=1,\cdots,6$ is as given in (2.37).

Table 2. Data statistics for final year GPA (y) and absenteeism from lectures (x).

Population/sample parameters	Stratum 1 (low-absenteeism)	Stratum 2 (high-absenteeism)
$N=50$	$N_1=32$	$N_2=18$
$n=20$	$n_1=12$	$n_2=8$
$(1-f)=0.60$	$(1-f_1)=0.625$	$(1-f_2)=0.556$
$\bar{Y}=2.98$	$\bar{Y}_1=3.16$	$\bar{Y}_2=2.65$
$\bar{X}=3.16$	$\bar{X}_1=2.03$	$\bar{X}_2=5.17$
$R=0.94$	$R_1=1.56$	$R_2=0.51$
$\pi=0.67$	$S_{y1}^2=0.2422$	$S_{y2}^2=0.0389$
$\omega_1=0.64$	$S_{x1}^2=0.9990$	$S_{x2}^2=0.6176$
	$S_{yx1}=-0.2124$	$S_{yx2}=-0.0161$
	$\omega_1=0.64$	$\omega_2=0.36$

Table 3. Percentage relative efficiencies under conditional and unconditional arguments.

Estimator	θ	Conditional argument			Unconditional argument		
		MSE	PRE-1 (%)	PRE-2 (%)	MSE	PRE-1 (%)	PRE-2 (%)
$\hat{R}_{1c}$	0.464	0.00148	259	100	0.00091	262	100
$\hat{R}_{2c}$	1.670	0.00872	44	588	0.00548	44	600
$\hat{R}_{3c}$	0.330	0.00111	346	75	0.00068	352	74
$\hat{R}_{4c}$	−0.670	0.00104	370	70	0.00067	360	73
$\hat{R}_{5c}$	0.130	0.00071	539	48	0.00043	553	47
$\hat{R}_{6c}$	−1.670	0.00576	67	388	0.00370	65	405
$\hat{R}_c$	1.000	0.00384	100	259	0.00240	100	262
$\hat{R}_{qc}^0$		0.00046	836	31	0.00028	854	31

over the customary combined-type estimator, $\hat{R}_c$, under the conditional and under the unconditional arguments. The table also shows the percentage relative efficiency (PRE-2) of the proposed combined-type estimators, $\hat{R}_{1c}$, over the other combined-type estimators, under the conditional and under the unconditional arguments.

Table 3 shows that apart from the estimators, $\hat{R}_{2c}$ and $\hat{R}_{6c}$, the remaining four proposed combined-type estimators, under the conditional and under the unconditional arguments, are more efficient than the customary combined-type estimator, $\hat{R}_c$, for the data under consideration, and their gains in efficiency (PRE-1) are relatively large. Also, using PRE-2, we observe that the proposed combined-type estimator, $\hat{R}_{1c}$, is more efficient than the estimators, $\hat{R}_{2c}$, $\hat{R}_{6c}$, and $\hat{R}_c$, under the conditional and unconditional arguments. The optimum estimator, as expected, has the highest gain in efficiency, both under the conditional and unconditional arguments. However, the customary combined-type estimator, on the other hand, is found to be more efficient than some of the proposed combined-type estimators for the given set of data. This confirms the theoretical results, which showed that the proposed estimators are not always more efficient than the customary combined-type estimator. Notice that $\beta' = -0.16$ and $R = 0.94$ showing that $\beta' < R$ and from the theoretical results in **Table 1**, the proposed estimators would be more efficient than the customary combined-type estimator, under the unconditional argument, if $|\theta_q| < 1$. The empirical results in **Table 3** show that $|\theta_2| > 1$ and $|\theta_6| > 1$, and the proposed estimators $\hat{R}_2$ (PRE-1 = 44%) and $\hat{R}_6$ (PRE-1 = 65%) under the unconditional argument, are less efficient than the customary combined-type estimator, $\hat{R}_c$. Hence the empirical results confirm the theoretical results.

5. Concluding Remarks

The study extends use of variable transformation in estimating population ratio in simple random sampling scheme to post-stratified sampling scheme. Efficiency conditions for preferring the proposed estimators to the customary combined-type estimator are obtained. The study shows that in any given survey, these efficiency conditions should be employed in order to determine the appropriate proposed combined-type estimators to use for the purpose of estimating the population ratio of two variables in post-stratified sampling scheme, using variable transformation.

References

[1] Onyeka, A.C., Nlebedim, V.U. and Izunobi, C.H. (2013) Estimation of Population Ratio in Simple Random Sampling Using Variable Transformation. *Global Journal of Science Frontier Research*, **13**, 57-65.

[2] Sukhatme, P.V. and Sukhatme, B.V. (1970) Sampling Theory of Surveys with Applications. Iowa State University Press, Ames.

[3] Cochran, W.G. (1977) Sampling Techniques. 3rd Edition, John Wiley & Sons, New York.

[4] Cochran, W.G. (1940) The Estimation of the Yields of the Cereal Experiments by Sampling for the Ratio of Grain to Total Produce. *The Journal of Agricultural Science*, **30**, 262-275. http://dx.doi.org/10.1017/S0021859600048012

[5] Robson, D.S. (1957) Application of Multivariate Polykays to the Theory of Unbiased Ratio-Type Estimation. *Journal of the American Statistical Association*, **52**, 511-522. http://dx.doi.org/10.1080/01621459.1957.10501407

[6] Murthy, M.N. (1964) Product Method of Estimation. *Sankhya Series A*, **26**, 294-307.

[7] Singh, M.P. (1965) On the Estimation of Ratio and Product of the Population Parameters. *Sankhya Series B*, **27**, 321-328.

[8] Upadhyaya, L.N., Singh, G.N. and Singh, H.P. (2000) Use of Transformed Auxiliary Variable in the Estimation of Population Ratio in Sample Survey. *Statistics in Transition*, **4**, 1019-1027.

[9] Onyeka, A.C., Nlebedim, V.U. and Izunobi, C.H. (2014) A Class of Estimators for Population Ratio in Simple Random Sampling Using Variable Transformation. *Open Journal of Statistics*, **4**, 284-291. http://dx.doi.org/10.4236/ojs.2014.44029

[10] Srivenkataramana, T. (1980) A Dual of Ratio Estimator in Sample Surveys. *Biometrika*, **67**, 199-204. http://dx.doi.org/10.1093/biomet/67.1.199

[11] Singh, H.P. and Tailor, R. (2005) Estimation of Finite Population Mean Using Known Correlation Coefficient between Auxiliary Characters. *Statistica*, **4**, 407-418.

[12] Tailor, R. and Sharma, B.K. (2009) A Modified Ratio-Cum-Product Estimator of Finite Population Mean Using Known Coefficient of Variation and Coefficient of Kurtosis. *Statistics in Transition—New Series*, **10**, 15-24.

[13] Sharma, B. and Tailor, R. (2010) A New Ratio-Cum-Dual to Ratio Estimator of Finite Population Mean in Simple Random Sampling. *Global Journal of Science Frontier Research*, **10**, 27-31.

[14] Onyeka, A.C. (2012) Estimation of Population Mean in Post-Stratified Sampling Using Known Value of Some Population Parameter(s). *Statistics in Transition—New Series*, **13**, 65-78.

Open Journal of Statistics, 2015, 5, 10-20
Published Online February 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.51002

Comparison of the Sampling Efficiency in Spatial Autoregressive Model

Yoshihiro Ohtsuka[1], Kazuhiko Kakamu[2]

[1]Department of Economics, University of Nagasaki, Nagasaki, Japan
[2]Faculty of Law, Politics and Economics, Chiba University, Chiba, Japan
Email: ohtsuka@sun.ac.jp, kakamu@le.chiba-u.ac.jp

Received 29 December 2014; accepted 18 January 2015; published 22 January 2015

Abstract

A random walk Metropolis-Hastings algorithm has been widely used in sampling the parameter of spatial interaction in spatial autoregressive model from a Bayesian point of view. In addition, as an alternative approach, the griddy Gibbs sampler is proposed by [1] and utilized by [2]. This paper proposes an acceptance-rejection Metropolis-Hastings algorithm as a third approach, and compares these three algorithms through Monte Carlo experiments. The experimental results show that the griddy Gibbs sampler is the most efficient algorithm among the algorithms whether the number of observations is small or not in terms of the computation time and the inefficiency factors. Moreover, it seems to work well when the size of grid is 100.

Keywords

Acceptance-Rejection Metropolis-Hastings Algorithm, Griddy Gibbs Sampler, Markov Chain Monte Carlo (MCMC), Random Walk Metropolis-Hastings Algorithm, Spatial Autoregressive Model

1. Introduction

Spatial models have been widely used in various research fields such as physical, environmental, biological science and so on. Recently, a lot of researches are also emerging in econometrics (e.g., [3] [4] and so on), and [5] gave an excellent survey from the viewpoint of econometrics. When we focus on the estimation methods, properties of several estimation methods are studied. For example, the efficient maximum likelihood (ML) method was proposed by [6], and [7] first formally proved that the quasi maximum likelihood estimator had the usual asymptotic properties, including $\sqrt{n}$ -consistency, asymptotic normality, and asymptotic efficiency. A class of moment estimators was examined by [8] and [9]. The Bayesian approach was first considered by [10] and [11] proposed a Markov chain Monte Carlo (hereafter MCMC) method to estimate the parameters of the

How to cite this paper: Ohtsuka, Y. and Kakamu, K. (2015) Comparison of the Sampling Efficiency in Spatial Autoregressive Model. *Open Journal of Statistics*, **5**, 10-20. http://dx.doi.org/10.4236/ojs.2015.51002

model. We have to mention that in economic analysis typically the sample size is small, for instance, areal data such as state-level data is widely used. The maximum likelihood methods depend on their asymptotic properties while the Bayesian method does not, because the latter evaluates the posterior distributions of the parameters conditioned on the data. Therefore, it is reasonable to examine the properties of Bayesian estimators (see [12]).

Although there are a lot of works using spatial models in a Bayesian framework, previous literature has rarely examined sampling methods for the parameter of spatial correlation. [13] proposed a random walk Metropolis-Hastings (hereafter RMH) algorithm. This method is widely used (e.g., [11] [12] [14] and so on). On the other hand, [2] applied a griddy Gibbs sampler (hereafter GGS) proposed by [1] and showed the GGS got an advantage over the RMH method from a simulated data and estimated the regional electricity demand in Japan. However, [2] has examined only one case. In this paper, we compare the properties of the GGS in the case that the number of observation is small (or large) through the Monte Carlo experiments. Desirable properties for sampling methods in the Bayesian inference are efficiency and well mixing, which yield fast convergence. In addition to these properties, computational requirements and model flexibility are important for applied econometrics. Therefore, the purpose of this paper is to investigate the properties of some sampling algorithms given several parameters of a model.

In this paper, we examine the efficiency of the existing Markov chain Monte Carlo methods for the spatial autoregressive (hereafter SAR) model which is the simplest and most commonly used model in the spatial models. Moreover, we propose an acceptance-rejection Metropolis-Hastings (hereafter ARMH) algorithm as an alternative MH algorithm, which is proposed by [15] because it is well known that the RMH is inefficient. This algorithm is widely used for the acceleration of MCMC convergence, for example, in the time series models (see [16]-[18] and so on). The advantage of this method is that the computational requirement is very small since it is irrelevant to the shape of the full conditional density. Therefore, we apply the algorithm to the SAR model.

We illustrate the properties of these algorithms using simulated data set given the three number of observations and the seven values of spatial correlation. From the results, we find that the GGS is the most efficient method whether the number of observations is small or not in terms of both the computation time and the inefficiency factors. Furthermore, we show that it is efficient when the number of grid in the GGS sampler is one hundred. These results give a benchmark of sampling the spatial correlation parameter of the models.

The rest of this paper is organized as follows. Section 2 summarizes the SAR model. Section 3 discusses the computational strategies of the MCMC methods, and reviews three sampling methods for spatial correlation parameter. Section 4 gives the Monte Carlo experiments using simulated data set and discusses the results. Finally, we summarize the results and provide concluding remarks.

2. Spatial Autoregressive (SAR) Model

Spatial autoregressive model explains the spatial spillover using a weight matrix (see [19]). There are numerous approaches to construct the weight matrix, which plays an important role in the model. For example, those are a first order contiguity matrix, inverse distance one and so on. Among the approaches, [20] recommended the first order contiguity dummies, because they showed that the first order contiguity weight matrix identifies the true model more frequently than the other matrices through the Monte Carlo simulations. Thus, we also utilize the first order contiguity dummies as the weight matrix.

Let $\boldsymbol{C}$ be an $n \times n$ matrix of contiguity dummies, with $c_{ij} = 1$ if areas i and j are adjacent and $c_{ij} = 0$ otherwise (with $c_{ii} = 0$). We standardized the weight matrix as follows

$$w_{ij} = \frac{c_{ij}}{\sum_{j=1}^{n} c_{ij}}$$

and we define $\boldsymbol{W} = \{w_{ij}\}$, where w_{ij} denotes the spatial weight on the j-th unit with respect to the i-th unit. Note that we have $\sum_{j=1}^{n} w_{ij} = 1$ for all i.

Next, let y_i and $\boldsymbol{x}_i$ be a dependent variable and a $1 \times k$ vector of covariates on the i th unit for $i = 1, \cdots, n$, respectively. Then, the SAR model conditioned on the parameters ρ, $\boldsymbol{\beta}$, σ^2 is written as follows:

$$y_i = \rho\sum_{j=1}^{n} w_{ij} y_j + \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}\left(0,\sigma^2\right), \quad |\rho| < 1 \tag{1}$$

where ρ and σ^2 indicates the spatial correlation, and the variance of the disturbance term, respectively. As is shown in [21], we know that $\lambda_{\min}^{-1} = -1$ amd $\lambda_{\max}^{-1} = 1$, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalue of $\boldsymbol{W}$, since we standardize the weight matrix like $\boldsymbol{W}$. Thus, we restrict ρ to $\rho \in (-1,1)$.

Then the likelihood function of the model (1) is given as follows:

$$L\left(\boldsymbol{y} \middle| \boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{X}, \boldsymbol{W}\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \left|\boldsymbol{I}_n - \rho\boldsymbol{W}\right| \exp\left(-\frac{\boldsymbol{e}'\boldsymbol{e}}{2\sigma^2}\right) \tag{2}$$

where $\boldsymbol{y} = \left(y_i,\cdots,y_n\right)'$, $\boldsymbol{X} = \left(\boldsymbol{x}_1',\cdots,\boldsymbol{x}_n'\right)'$, $\boldsymbol{e} = \left(e_i,\cdots,e_n\right)'$, $e_i = y_i - \sum_{j=1}^{n}\rho w_{ij} y_j - \boldsymbol{x}_i\boldsymbol{\beta}$, and $\boldsymbol{I}_n$ is an $n \times n$ unit matrix.

3. Posterior Analysis and Simulation

3.1. Joint Posterior Distribution

Since we adopt the Bayesian approach, we complete the model by specifying the prior distribution over the parameters. We use the following independent prior distribution:

$$\pi\left(\boldsymbol{\beta},\sigma^2,\rho\right) = \pi\left(\boldsymbol{\beta}\right)\pi\left(\sigma^2\right)\pi\left(\rho\right)$$

Given a prior density $\pi\left(\boldsymbol{\beta},\sigma^2,\rho\right)$ and the likelihood function given in (2), the joint posterior distribution can be expressed as

$$\pi\left(\boldsymbol{\beta},\sigma^2,\rho \middle| \boldsymbol{y},\boldsymbol{X},\boldsymbol{W}\right) \propto \pi\left(\boldsymbol{\beta},\sigma^2,\rho\right) L\left(\boldsymbol{y} \middle| \boldsymbol{\beta},\sigma^2,\rho,\boldsymbol{X},\boldsymbol{W}\right) \tag{3}$$

Finally, we assume the following prior distributions:

$$\boldsymbol{\beta} \sim N\left(\mu_0,\boldsymbol{\Sigma}_0\right), \quad \sigma^2 \sim IG\left(\nu_0/2,\lambda_0/2\right), \quad \rho \sim U\left(-1,1\right)$$

where $IG(a,b)$ denotes an inverse gamma distribution with scale and shape parameters a and b.

Since the joint posterior distribution is given by (3), we can now adopt the MCMC method. The Markov chain sampling scheme can be constructed from the full conditional distributions of ρ, β and σ^2.

3.2. Sampling ρ

From (3), the full conditional distribution of ρ is written as

$$p\left(\rho \middle| \boldsymbol{\beta},\sigma^2,\rho,\boldsymbol{y},\boldsymbol{X},\boldsymbol{W}\right) \propto \left|\boldsymbol{I}_n - \rho\boldsymbol{W}\right| \exp\left(-\frac{\boldsymbol{e}'\boldsymbol{e}}{2\sigma^2}\right) \tag{4}$$

As it is difficult to sample from the standard distribution, we examine three approaches for sampling ρ. First, we introduce the GGS, which is applied by [2]. Second, we overview the RMH algorithm, which is extended by [13]. Finally, we propose an ARMH algorithm. These sampling methods are summarized in the following.

3.2.1. Griddy Gibbs Sampler

The GGS was proposed by [1]. This sampling algorithm approximates a cumulative distribution function of the full conditional distribution by each kernel function over a grid of points and uses a numerical integration method, and is sampling method from the full conditional distribution by using the inverse transform method. Let the grid be as follows

$$-1 = a_1 < a_2 < \cdots < a_m < a_{m+1} = 1$$

and $\rho^i\left(i \in \{1,\cdots,m\}\right)$, which is centered in the interval $\left[a_i,a_{i+1}\right]$. Then, the full conditional distribution in the interval $\left[a_i,a_{i+1}\right]$ is approximated as follows

$$\omega_i = \frac{p\left(\rho^i \middle| \beta, \sigma^2, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{W}\right)}{\sum_{h=1}^{m} p\left(\rho^h \middle| \beta, \sigma^2, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{W}\right)}$$

Thus, we select the grid a_i^* with probabilities,

$$h\left(\rho^i\right) = \frac{\omega_i\left(a_{i+1} - a_i\right)}{\sum_{j=1}^{m} \omega_j\left(a_{j+1} - a_j\right)}$$

Finally, we sample ρ from the uniform $\left(a_i^*, a_{i+1}^*\right)$. [22] stated that the choice of the grid of points has to be made carefully and constitute the main difficulty in applying GGS. In this paper, we select the equal interval among $a_{m+1} - a_m$ as in [1]. Then, our numerical experiments examines to choice the size of grid for estimating the spatial correlation.

3.2.2. Random Walk Metropolis-Hastings Algorithm

The RMH method is a simple algorithm because it needs the previous value and a random walk process such as $\phi^{\text{new}} \sim N\left(\phi^{\text{old}}, \tau^2\right)$, where ϕ^{old} is the parameter of the previous sampling, and τ denotes the tuning parameter, respectively. Therefore, the following Metropolis step is used: Sample ρ^{new} from

$$\rho^{\text{new}} \sim N\left(\rho^{\text{old}}, s^2\right)$$

where s is the tuning parameter. In the numerical example below, we select the tuning parameter such that the acceptance rate lies between 0.4 and 0.6 (see [13]). Next, we evaluate the acceptance probability

$$\alpha\left(\rho^{\text{old}}, \rho^{\text{new}}\right) = \min\left(\frac{p\left(\rho^{\text{new}} \middle| \boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{W}\right)}{p\left(\rho^{\text{old}} \middle| \boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{W}\right)}, 1\right)$$

And finally set $\rho = \rho^{\text{new}}$ with probability $\alpha\left(\rho^{\text{old}}, \rho^{\text{new}}\right)$, otherwise $\rho = \rho^{\text{old}}$. The proposal value of ρ is not truncated to the interval $(-1,1)$ because the constraint is part of the target density. Thus, if the proposed value of ρ is not within the interval, the conditional posterior is zero, and the proposal value is rejected with probability one (see [23]). It is well known that the method is not efficient because the convergence is slow for using the previous sampled parameter.

3.2.3. Acceptance-Rejection Metropolis-Hastings Algorithm

An acceptance-rejection Metropolis-Hastings (ARMH) algorithm method was proposed by [15]. This algorithm samples the parameter using the AR and MH steps. Suppose that there is a candidate function $g(\rho)$ such that it is possible to sample directly from $g(\rho)$ by some known method. Then, the AR step proceeds as follows. We sampling the parameter from the candidate function $g(\rho)$, and accepts the candidate draw with probability $p(\rho)/cg(\rho)$. This step is iterated until the candidate draw is accepted.

Next, suppose the candidate ρ^{new} is produced from above AR step. The MH part proceeds as follows. We calculate the acceptance probability, q as following:

$$\begin{aligned}
&\text{If } p\left(\rho^{\text{old}}\right) < cg\left(\rho^{\text{old}}\right), && \text{then } q = 1;\\
&\text{If } p\left(\rho^{\text{old}}\right) \geqslant cg\left(\rho^{\text{old}}\right) \text{ and } p\left(\rho^{\text{new}}\right) < cg\left(\rho^{\text{new}}\right), && \text{then } q = \frac{cg\left(\rho^{\text{old}}\right)}{p\left(\rho^{\text{old}}\right)};\\
&\text{If } p\left(\rho^{\text{old}}\right) \geqslant cg\left(\rho^{\text{old}}\right) \text{ and } p\left(\rho^{\text{new}}\right) \geqslant cg\left(\rho^{\text{new}}\right), && \text{then } q = \min\left[\frac{p\left(\rho^{\text{new}}\right) g\left(\rho^{\text{old}}\right)}{p\left(\rho^{\text{old}}\right) g\left(\rho^{\text{new}}\right)}, 1\right].
\end{aligned}$$

In this step, the candidate draw is accepted with probability q and rejected with probability $1-q$. If the draw is rejected, the previously sampled value is sampled again. If q is small, the probability of sampling the same value consecutively is high, causing high autocorrelation across sample values (see [24]). Hence, we should also make q as close to one as possible.

The advantage of this method is that it is free to functional form which differs from the GGS and RMH. In this paper, in order to construct the candidate function, we utilize the result of [7], which showed the consistency and asymptotic normality of quasi-ML estimators of model parameters, to the candidate density. Then, we construct the candidate density $g(\rho)$ as an approximation to the the conditional posterior density by omitting the determinant $|\boldsymbol{I}-\rho\boldsymbol{W}|$ as follows:

$$g(\rho)\sim N\left(\hat{\mu}_\rho,\hat{\sigma}_\rho^2\right) \tag{5}$$

where $\hat{\mu}_\rho=\left(\boldsymbol{y}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{y}\right)^{-1}\left\{\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)'\boldsymbol{W}\boldsymbol{y}\right\}$ and $\hat{\sigma}_\rho^2=\sigma^2\left(\boldsymbol{y}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{y}\right)^{-1}$. Thus we sample ρ^{new} from the distribution, and apply the ARMH algorithm.

3.3. Sampling Other Parameters

The full conditional distributions of β and σ^2 are

$$\boldsymbol{\beta}\sim N\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Sigma}}\right),\quad \text{and}\quad \sigma^2\sim IG\left(\hat{v}/2,\hat{\lambda}/2\right)$$

where $\hat{\beta}=\hat{\boldsymbol{\Sigma}}\left\{\sigma^{-2}\boldsymbol{X}'\left(\boldsymbol{y}-\rho\boldsymbol{W}\boldsymbol{y}\right)+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0\right\}$, $\hat{\boldsymbol{\Sigma}}=\left(\sigma^{-2}\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}_0^{-1}\right)^{-1}$, $\hat{v}=n+v_0$, and $\hat{\lambda}=\boldsymbol{e}'\boldsymbol{e}+\lambda_0$. These parameters are easily sampled from the Gibbs sampler (see [25]).

4. Comparison of MCMC Methods

4.1. Measures of Efficiency for Comparison

In this section, we examine the properties of three MCMC methods by simulated data sets. Desirable properties for sampling methods in MCMC are efficiency and well mixing, which yield fast convergence. [17] compared from the view point of acceptance rate in the AR and MH step. [26] [27] evaluated the efficiency of sampling methods, comparing the inefficiency factor and time of MCMC simulation. Following previous literatures, we also compare inefficiency factor and computational time.

The inefficiency factor is defined as $1+2\sum_{s=1}^{\infty}r_s$ where r_s is the sample autocorrelation at lag s calculated from the sampled values. It is used to measure how well the chain mixes and is the ratio of the numerical variance of the sample posterior mean to the variance of the sample mean from the hypothetical uncorrelated draws (see [28]).

4.2. Data Generating Process and Estimation Procedures

We now explain the simulated data for an experiment. First, we give the weight matrix as an exogenous variable. We construct the spatial weight matrix $\boldsymbol{W}$ as follows: 1) generate c_{ij} for $i>j$ from Bernoulli distribution with a probability of success 0.3, 2) set $c_{ij}=c_{ji}$ for $i\neq j$ and $c_{ij}=0$ for $i=j$, and 3) compute $w_{ij}=c_{ij}\Big/\sum_{j=1}^{n}c_{ij}$ for all i, j. Next, for the independent variables $\boldsymbol{x}_i=\left(1,x_{1i},x_{2i},x_{3i}\right)$, we take the standard normal variates and set the $\boldsymbol{X}$, which are $n\times 4$ covariate matrices.

Given $\boldsymbol{W}$, $\boldsymbol{X}$, $\rho=-0.9,-0.6,-0.3,0,0.3,0.6,0.9$, and $n=50,100,200$, the true data generating process is as follows:

$$y_i=\rho\sum_{j=1}^{n}w_{ij}y_j+\boldsymbol{x}_i\boldsymbol{\beta}+\epsilon_i \tag{6}$$

where the ϵ_i is normally and independently distributed with $E\left(u_i\right)=0$ and $E\left(u_i^2\right)=\sigma^2$. The parameter is set to be $\boldsymbol{\beta}'=\left(\beta_0,\beta_1,\beta_2,\beta_3\right)=\left(1,1,1,1\right)$ and $\sigma^2=0.1$, respectively. The parameters of ρ for simulated data reflect the values obtained in [12]. All the results in this paper were calculated using the Ox version 5.1 (see [29]).

The prior distributions are as follows:

$$\boldsymbol{\beta} \sim N\left(\mathbf{0}, 100 \times \boldsymbol{I}_4\right), \quad \sigma^2 \sim IG\left(1.0/2, 0.01/2\right), \quad \text{and} \quad \rho \sim U\left(-1, 1\right)$$

We perform the MCMC procedure by generating 35,000 draws in a single sample path and discard the first 20,000 draws as the initial burn-in. For the GGS, we consider the number of grid, $m = 50, 100, 300$ for estimating the parameters.

4.3. Results of Comparison

Table 1 reports inefficiency factors by using three methods. Although there are some differences, the performances of the GGS are almost equivalent to those of the ARMH. In addition, these algorithms are more efficient than RMH. For example, from the table in $n = 50$, the inefficiency factors calculated by the ARMH are smaller than the other methods. However, if spatial correlation is positive strong such as $\rho = 0.9$, the value by the GGS $\left(m = 100\right)$ has the smallest inefficiency factor. Next, we focus on the results in $n = 100$. In this case, the GGS $\left(m = 50, 100\right)$ perform the best for $\rho = 0.6, 0.9$, respectively. In the case of $n = 200$, the values of the GGS $\left(m = 100\right)$ and the ARMH are similar in each parameter. We can also find such similarity in sample paths and autocorrelation functions. **Figure 1** shows the results of MCMC simulation in each method in the cases of $\rho = 0.3$, $n = 50$ and $m = 100$. The figure shows that the marginal posterior densities (middle of the figure)

Table 1. Inefficiency factor of models.

Observation: $n = 50$					
Parameter	RMH		GGS		ARMH
ρ		$m = 50$	$m = 100$	$m = 300$	
−0.9	7.2	3.2	3.4	3.4	2.8
−0.6	27.6	4.4	4.4	4.7	3.7
−0.3	15.4	23.7	6.6	6.9	4.3
0	41.6	9.0	10.1	11.5	6.6
0.3	79.8	24.6	19.6	20.9	13.1
0.6	117.0	46.3	45.2	52.3	44.6
0.9	806.1	312.9	223.1	327.2	324.9
Observation: $n = 100$					
Parameter	RMH		GGS		ARMH
ρ		$m = 50$	$m = 100$	$m = 300$	
−0.9	10.7	4.8	5.0	5.3	4.7
−0.6	17.0	6.7	7.3	7.5	4.6
−0.3	34.7	9.0	10.2	10.9	5.0
0	72.4	15.4	16.2	17.8	9.5
0.3	85.1	24.5	25.5	32.6	19.9
0.6	202.3	36.1	56.6	65.8	51.3
0.9	609.1	379.3	338.0	342.1	338.9
Observation: $n = 200$					
Parameter	RMH		GGS		ARMH
ρ		$m = 50$	$m = 100$	$m = 300$	
−0.9	22.2	7.1	8.3	5.7	7.8
−0.6	31.0	11.5	12.4	13.5	9.0
−0.3	64.8	17.6	17.6	19.1	13.8
0	75.7	23.5	26.4	33.6	23.7
0.3	163.6	57.4	67.3	65.6	50.5
0.6	697.3	164.2	117.5	163.3	159.3
0.9	860.4	695.1	628.7	694.0	780.6

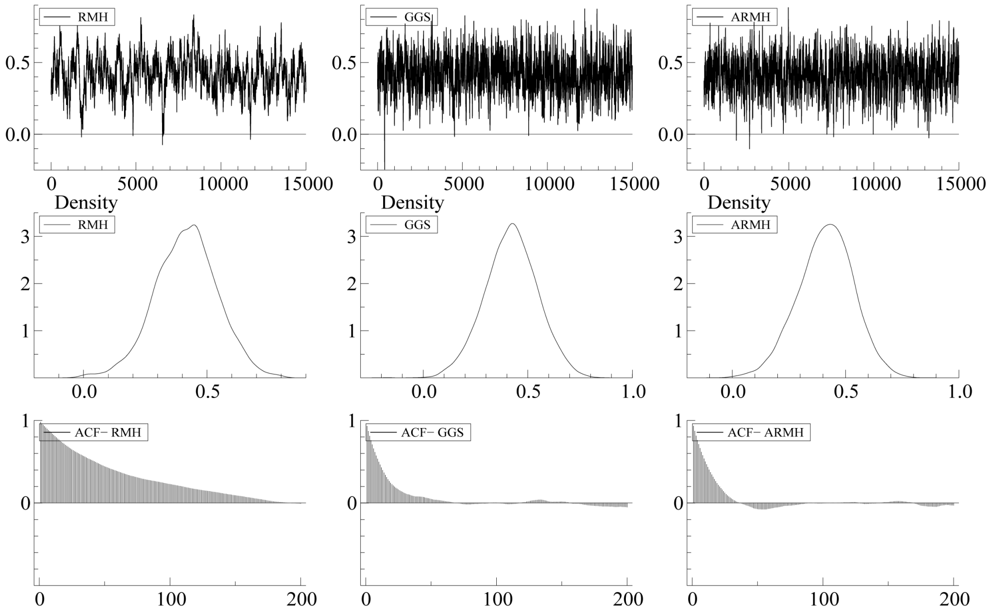

Figure 1. Sample paths, sample autocorrelation and posterior density of $\rho = 0.3$, $n = 50$.

have similar shapes but that the sample paths (top of the figure) and autocorrelation functions (bottom of the figure) are different. From the sample paths, we can find that the ARMH and GGS mix better than the RMH. As same as the sample paths, autocorrelation functions shows the same tendency. The figure of autocorrelation indicates that both GGS and ARMH perform similarly in the autocorrelation disappear. On the contrary, the result for the RMH indicates that serious autocorrelation for parameter at large lag length.

Table 2 shows CPU time on a Pentium Core2 Duo 2.4GHz including discarded and rejected draws. For the GGS, the computation time depends on the number of grid because the increase of grid number causes the cost of computation time. In all cases, the GGS $(m = 50)$ overwhelms the others. If we focus on the case of $n = 50$, the computational time of the GGS $(m = 100)$ are as same as those of the RMH and ARMH methods. Futhermore, if $n = 200$, the GGS needs much shorter time than the RMH and ARMH methods. Summarizing the results of inefficiency factors and computational time, if the number of observation is not only small (like $n = 50$) but also large, then it is suitable to use the GGS. In addition, the choice of grid number affects to the computational time. In this numerical experiments, the results of selecting $m = 100$ seem to work well in terms of inefficiency factors and computational time.

Table 3 shows the results with acceptance probabilities in both AR and MH parts in the ARMH. From the table, the acceptance probabilities in those part are exceed 89%. This result shows that our candidate function seems to work well, and the probabilities of sampling the same value consecutively are low. However, our ARMH algorithm does not improve the values of inefficiency factor. Thus, we think that the SAR model has the problem of identification.

Figure 2 and **Table 4** depict the sample path and the correlation among the parameters in the case of $n = 100$, $\rho = 0.9$, $m = 100$ using the GGS. From β_2 to β_4 and σ^2 in the figure, the MCMC draws seem to be well mixing. In addition, correlations among these parameters are very small. On the other hand, strong correlation between β_0 and ρ can be found from the figure. Moreover, the correlation between β_0 and ρ is -0.995 from the table. Therefore, we assume that the spatial correlation and constant term is weakly identified.

5. Concluding Remarks

This paper reviewed the MCMC estimation procedures for sampling the spatial correlation of SAR model, and proposed the ARMH algorithm as more efficient than the RMH in order to show the property of the GGS proposed by [2]. To illustrate the differences between the estimates of three MCMC methods, we compared these algorithms by simulated data set. From the Monte Carlo experiments, we found that the GGS was the most efficient algorithm with respect to the mixing, efficiency and computational requirement of the MCMC. Moreover,

Table 2. Time of convergence.

Observation: $n = 50$					
Parameter	RMH	GGS			ARMH
ρ		$m = 50$	$m = 100$	$m = 300$	
−0.9	22.17	11.12	22.68	1:05.96	24.24
−0.6	23.22	11.57	23.06	1:05.99	23.95
−0.3	23.31	11.71	23.21	1:07.18	23.99
0	23.27	11.83	23.27	1:07.87	24.01
0.3	23.20	12.26	23.10	1:09.49	23.99
0.6	24.16	12.07	22.70	1:08.36	24
0.9	23.17	12.06	23.64	1:08.66	24.02
Observation: $n = 100$					
Parameter	RMH	GGS			ARMH
ρ		$m = 50$	$m = 100$	$m = 300$	
−0.9	1:31.73	18.67	35.49	1:37.61	1:44.90
−0.6	1:41.36	17.25	35.75	1:37.83	1:42.99
−0.3	1:43.81	19.93	36.64	1:39.25	1:43.22
0	1:40.10	18.30	40.15	1:38.47	1:43.53
0.3	1:40.90	19.90	40.04	1:41.40	1:43.55
0.6	1:41.73	18.91	37.35	1:43.97	1:43.13
0.9	1:43.36	17.93	37.89	1:40.33	1:42.27
Observation: $n = 200$					
Parameter	RMH	GGS			ARMH
ρ		$m = 50$	$m = 100$	$m = 300$	
−0.9	8:40.79	26.88	56.62	2:43.63	9:05.58
−0.6	8:43.81	26.66	56.76	2:45.73	9:07.63
−0.3	8:59.71	26.74	58.84	2:44.22	9:08.03
0	8:57.87	26.92	57.48	2:46.64	8:56.41
0.3	9:03.95	27.02	58.49	2:45.99	8:51.45
0.6	9:12.82	28.24	58.13	2:48.13	9:01.35
0.9	9:22.86	27.10	57.84	2:48.15	8:59.61

Note: Time denotes CPU time on a Pentium Core2 Duo, including discarded and rejected draws.

Table 3. Acceptance probability of the ARMH methods.

Parameter	$n = 50$		$n = 100$		$n = 200$	
ρ	AR step	MH step	AR step	MH step	AR step	MH step
−0.9	0.9866	0.9116	0.9578	0.8975	0.9881	0.9505
−0.6	0.9999	0.9500	0.9999	0.9438	1.0000	0.9724
−0.3	1.0000	0.9848	1.0000	0.9805	1.0000	0.9906
0	1.0000	0.9849	1.0000	0.9787	1.0000	0.9949
0.3	1.0000	0.9670	1.0000	0.9544	1.0000	0.9861
0.6	0.9991	0.9553	0.9958	0.9375	1.0000	0.9802
0.9	0.9997	0.9716	0.9977	0.9649	0.9997	0.9821

Table 4. Correlation of parameters.

	β_1	β_2	β_3	β_4	σ^2
β_2	−0.050				
β_3	0.087	0.106			
β_4	−0.059	0.205	0.183		
σ^2	0.161	−0.007	0.002	−0.025	
ρ	−0.995	0.054	−0.075	0.078	−0.160

Note: True parameter is 0.9. The number of observation set to be 100.

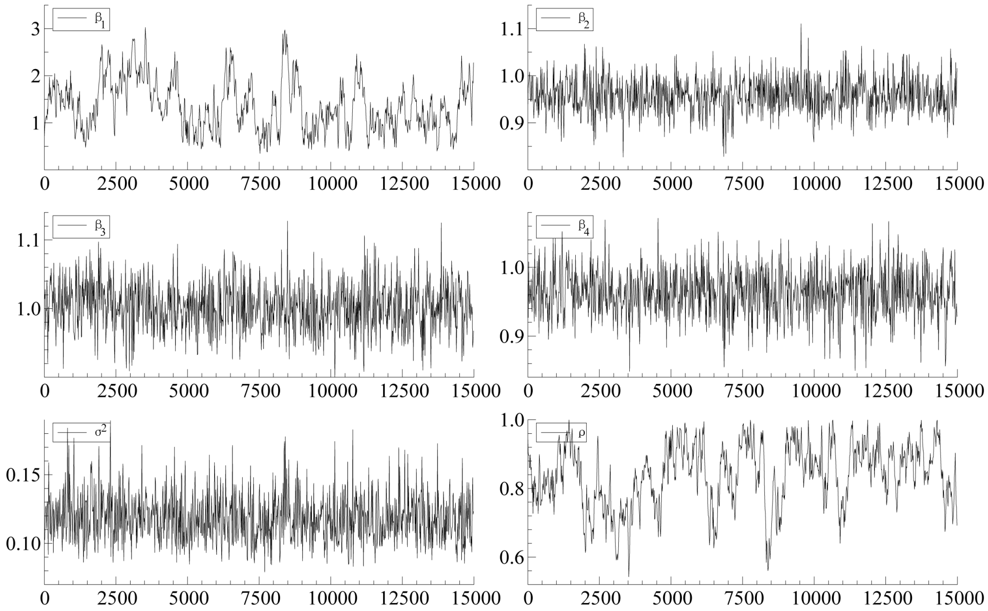

Figure 2. Sample paths of SAR model with GGS ($\rho = 0.9$, $n = 100$, $m = 100$).

the results of selecting $m = 100$ seem to work well in terms of inefficiency factors and computational time. Therefore, the GGS is beneficial algorithm for estimating the spatial parameter as same as the result of [22].

Finally, we will state our remaining issues. In this paper, we found that the GGS was the most efficient algorithm in sampling the intensity of spatial interaction. On the other hand, we showed the problem of the SAR model such that the spatial correlation and constant term was weakly identified. Thus, we have to construct the model which is identified, or appropriate algorithm to sample the intensity of spatial interaction. Furthermore, we found that the number of grids is appropriate when $m = 100$. In this paper, we could not derive the theoretical reason why $m = 100$ was appropriate number of grids, that was, we only showed the results of Monte Carlo experiments. However, it is important to know the properties of the existing sampling methods, though research on the convergence of the GGS algorithm has never been examined. We think that, in this respect, our experiment gives the benchmark in applied econometrics.

Acknowledgements

We gratefully acknowledge helpful discussions and suggestions with Toshiaki Watanabe on several points in the paper. Mototsugu Fukushige gave insightful comments and suggestions when we made a presentation at Japan Association for Applied Economics in June, 2009. This research was partially supported by KAKENHI # 25245035 for K. Kakamu.

References

[1] Ritter, C. and Tanner, M. (1992) Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler.

Journal of the American Statistical Association, **87**, 861-868. http://dx.doi.org/10.1080/01621459.1992.10475289

[2] Ohtsuka, Y. and Kakamu, K. (2009) Estimation of Electric Demand in Japan: A Bayesian Spatial Autoregressive AR(p) Approach. In: Schwartz, L.V., Ed., *Inflation*: *Causes and Effects*, Nova Science Publisher, New York, 156-178.

[3] Anselin, L. (2003) Spatial Externalities, Spatial Multipliers, and Spatial Econometrics. *International Regional Science Review*, **26**, 153-166. http://dx.doi.org/10.1177/0160017602250972

[4] Gelfand, A.E., Banerjee, S., Sirmans, C.F., Tu, Y. and Ong, S.E. (2007) Multilevel Modeling Using Spatial Processes: Application to the Singapore Housing Market. *Computational Statistics and Data Analysis*, **51**, 3567-3579. http://dx.doi.org/10.1080/01621459.1990.10476213

[5] Anselin, L. (2010) Thirty Years of Spatial Econometrics. *Papers in Regional Science*, **89**, 3-25. http://dx.doi.org/10.1111/j.1435-5957.2010.00279.x

[6] Ord, K. (1975) Estimation Methods for Models for Spatial Interaction. *Journal of the American Statistical Association*, **70**, 120-126. http://dx.doi.org/10.1080/01621459.1975.10480272

[7] Lee, L.F. (2004) Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models. *Econometrica*, **72**, 1899-1925. http://dx.doi.org/10.1111/j.1468-0262.2004.00558.x

[8] Conley, T.G. (1999) GMM Estimation with Cross Sectional Dependence. *Journal of Econometrics*, **92**, 1-45. http://dx.doi.org/0.1016/S0304-4076(98)00084-0

[9] Kelejian, H.H. and Prucha, I.R. (1999) A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model. *International Economic Review*, **40**, 509-533. http://dx.doi.org/10.1111/1468-2354.00027

[10] Anselin, L. (1980) A Note on Small Sample Properties of Estimators in A First-order Spatial Autoregressive Model. *Environment and Planning A*, **14**, 1023-1030. http://dx.doi.org/10.1068/a141023

[11] LeSage, J.P. (1997) Regression Analysis of Spatial Data. *The Journal of Regional Analysis and Policy*, **27**, 83-94.

[12] Kakamu, K. and Wago, H. (2008) Small-Sample Properties of Panel Spatial Autoregressive Models: Comparison of the Bayesian and Maximum Likelihood Methods. *Spatial Economic Analysis*, **3**, 305-319. http://dx.doi.org/10.1080/17421770802353725

[13] Holloway, G., Shankar, B. and Rahman, S. (2002) Bayesian Spatial Probit Estimation: A Primer and an Application to HYV Rice Adoption. *Agricultural Economics*, **27**, 383-402. http://dx.doi.org/10.1111/j.1574-0862.2002.tb00127.x

[14] Ohtsuka, Y., Oga, T. and Kakamu, K. (2010) Forecasting Electricity Demand in Japan: A Bayesian Spatial Autoregressive ARMA Approach. *Computational Statistics & Data Analysis*, **54**, 2721-2735. http://dx.doi.org/10.1016/j.csda.2009.06.002

[15] Tierney, L. (1994) Markov Chains for Exploring Posterior Distributions (with Discussion). *Annals of Statistics*, **22**, 1701-1728. http://dx.doi.org/10.1214/aos/1176325750

[16] Chib, S. and Greenberg, E. (1994) Bayes Inference in Regression Models with ARMA(p,q) Errors. *Journal of Econometrics*, **64**, 183-206. http://dx.doi.org/10.1016/0304-4076(94)90063-9

[17] Watanabe, T. (2001) On Sampling the Degree-of-Freedom of Student's-t Disturbances. *Statistics & Probability Letters*, **52**, 177-181. http://dx.doi.org/10.1016/S0167-7152(00)00221-2

[18] Mitsui, H. and Watanabe, T. (2003) Bayesian Analysis of GARCH Option Pricing Models. *Journal of the Japan Statistical Society* (*Japanese Issue*), **33**, 307-324.

[19] LeSage, J.P. and Pace, R.K. (2008) Introduction to Spatial Econometrics (Statistics: A Series of Textbooks and Monographs). Chapman and Hall/CRC, London.

[20] Stakhovych, S. and Bijmolt, T.H.A. (2009) Specification of Spatial Models: A Simulation Study on Weights Matrices. *Papers in Regional Science*, **88**, 389-408. http://dx.doi.org/10.1111/j.1435-5957.2008.00213.x

[21] Sun, D., Tsutakawa, R.K. and Speckman, P.L. (1999) Posterior Distribution of Hierarchical Models Using CAR(1) Distributions. *Biometrika*, **86**, 341-350. http://dx.doi.org/10.1093/biomet/86.2.341

[22] Bauwens, L. and Lubrano, M. (1998) Bayesian Inference on GARCH Models Using the Gibbs Sampler. *The Econometrics Journal*, **1**, 23-46. http://dx.doi.org/10.1111/1368-423X.11003

[23] Chib, S. and Greenberg, E. (1998) Analysis of Multivariate Probit Models. *Biometrika*, **85**, 347-361. http://dx.doi.org/10.1093/biomet/85.2.347

[24] Chib, S. and Greenberg, E. (1995) Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, **49**, 327-335.

[25] Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**, 398-409. http://dx.doi.org/10.1080/01621459.1990.10476213

[26] Asai, M. (2005) Comparison of MCMC Methods for Estimating Stochastic Volatility Models. *Computational Eco-*

nomics, **25**, 281-301. http://dx.doi.org/10.1007/s10614-005-2974-4

[27] Asai, M. (2006) Comparison of MCMC Methods for Estimating GARCH Models. *Journal of the Japan Statistical Society*, **36**, 199-212. http://dx.doi.org/10.14490/jjss.36.199

[28] Chib, S. (2001) Markov Chain Monte Carlo Methods: Computation and Inference. In: Heckman, J.J. and Leamer, E., Eds., *Handbook of Econometrics*, Elsevier, Amsterdam, 3569-3649.

[29] Doornik, J.A. (2006) Ox: An Object Oriented Matrix Programming Language. Timberlake Consultants Press, London.

Open Journal of Statistics, 2015, 5, 21-26
Published Online February 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.51003

Scientific
Research
Publishing

Correct Classification Rates in Multi-Category Discriminant Analysis of Spatial Gaussian Data

Lina Dreižienė[1,2], Kęstutis Dučinskas[1], Laura Paulionienė[1]

[1]Department of Mathematics and Statistics, Klaipėda University, Klaipėda, Lithuania
[2]Institute of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania
Email: l.dreiziene@gmail.com, kestutis.ducinskas@ku.lt, saltyte.laura@gmail.com

Received 28 December 2014; accepted 17 January 2015; published 22 January 2015

Abstract

This paper discusses the problem of classifying a multivariate Gaussian random field observation into one of the several categories specified by different parametric mean models. Investigation is conducted on the classifier based on plug-in Bayes classification rule (PBCR) formed by replacing unknown parameters in Bayes classification rule (BCR) with category parameters estimators. This is the extension of the previous one from the two category cases to the multi-category case. The novel closed-form expressions for the Bayes classification probability and actual correct classification rate associated with PBCR are derived. These correct classification rates are suggested as performance measures for the classifications procedure. An empirical study has been carried out to analyze the dependence of derived classification rates on category parameters.

Keywords

Gaussian Random Field, Bayes Classification Rule, Pairwise Discriminant Function, Actual Correct Classification Rate

1. Introduction

Much work has been done concerning the error rates in two-category discrimination of uncorrelated observations (see e.g. [1]). Several methods for estimations of the error rates in discriminant analysis of spatial data have been recently proposed (see e.g. [2] [3]).

The multi-category problem, however, has very rarely been addressed because most of the methods proposed for two categories do not generalize. Schervish [4] considered the problem of classification into one of three

How to cite this paper: Dreižienė, L., Dučinskas, K. and Paulionienė, L. (2015) Correct Classification Rates in Multi-Category Discriminant Analysis of Spatial Gaussian Data. *Open Journal of Statistics*, **5**, 21-26.
http://dx.doi.org/10.4236/ojs.2015.51003

known normal populations by single linear discriminant function. Techniques for multi-category probability estimation by combining all pairwise comparisons are investigated by several authors (see e.g. [5]). Empirical comparison of different methods of error rate estimation in multi-category linear discriminant analysis for multivariate homoscedastic Gaussian data was performed by Hirst [6]. Bayesian multiclass classification problem for correlated Gaussian observation was empirically studied by Williams [7]. The novel model-free estimation method for multiclass conditional probability based on conditional quintile regression functions is theoretically and numerically studied by Xu [8]. Correct classification rates in multi-category classification of independent multivariate Gaussian observations were provided by Schervish [9]. We generalize results of above to the problem of classification of multivariate spatially correlated Gaussian observations.

We propose the method of multi-category discriminant analysis essentially exploiting the Bayes classification rule that is optimal in the sense of minimum misclassification probability in case of complete statistical certainty (see [10], chapter 6). In practice, however, the complete statistical description of populations is usually not possible. Then having training sample, parametric plug-in Bayes classification rule formed by replacing unknown parameters with their estimators in BCR is being used.

Šaltytė and Dučinskas [11] derived the asymptotic approximation of the expected error rate when classifying the observation of a scalar Gaussian random field into one of two classes with different regression mean models and common variance. This result was generalized to multivariate spatial-temporal regression model in [12]. However, the observations to be classified are assumed to be independent from training samples in all publication listed above. The assumption of independence for the classification of scalar GRF observations was removed by Dučinskas [2]. Multivariate two-category case has been considered in Dučinskas [13] and Dučinskas and Dreižienė [14]. Formulas for the error rates for multiclass classification of scalar GRF observation are derived in [15]. The authors of the above papers have been focused on the maximum likelihood (ML) estimators because of tractability of the covariance matrix of these estimators. In the present paper, we extend the investigation of the performance of the PBCR in multi-category case. The novel closed form expressions for the actual correct classification rate (ACCR) are derived.

By using the derived formulas, the performance of the PBR is numerically analyzed in the case of stationary Gaussian random field on the square lattice with the exponential covariance function. The dependence of the correct classification rate and ACCR values on the range parameter is investigated.

The rest of the paper is organized as follows. Section 2 presents concepts and notions concerning BCR applied to multi-category classification of multivariate Gaussian random field (MGRF) observation. Bayes probability of correct classification is derived. In Section 3, the actual correct classification rate incurred by PBCR is considered and its closed-form expression is derived. Numerical examples, based on simulated data, are presented in Section 4, in order to illustrate theoretical results. The effect of the values of range parameter on the values of ACCR is examined.

2. The Main Concepts and Definitions

The main objective of this paper is to classify a single observation of MGRF $\left\{Z(s): s \in D \subset R^2\right\}$ into one of L categories, say $\Omega_1, \cdots, \Omega_L$.

The model of observation $Z(s)$ in category Ω_l $(l = 1, \cdots, L)$ is

$$Z(s) = \mu_l(s; B_l) + \varepsilon(s).$$

Here μ_l represents a mean component and B_l is a matrix of parameters. The error term is generated by *p*-dimensional zero-mean stationary GRF $\left\{\varepsilon(s): s \in D\right\}$ with covariance function defined by model for all $s, u \in D$

$$\operatorname{cov}\left\{\varepsilon(s), \varepsilon(u)\right\} = r(s-u)\Sigma,$$

where $r(s-u)$ is the spatial correlation function and Σ is the variance-covariance matrix with elements $\left\{\sigma_{ij}\right\}$. So we have deal with so called intrinsic covariance model (see [16]).

Consider the problem of classification of the vector of observation of Z at location s_0 denoted by $Z_0 = Z(s_0)$ into one of L populations specified above with given joint training sample T. Joint training sample T is stratified training sample, specified by $n \times p$ matrix $T = \left(T_1', \cdots, T_L''\right)'$, where T_l is the $n_l \times p$ matrix of n_l observations of $Z(\bullet)$ from Ω_l, $l = 1, \cdots, L$, $n = \sum_{l=1}^{L} n_l$.

Then the model of T is

$$T = M(B) + E,$$

where $B' = \left(B_1', \cdots, B_L'\right)$ is the matrix of category means parameters and E is the $n \times p$ matrix of random errors that has matrix-variate normal distribution *i.e.*

$$E \sim N_{n \times p}\left(0, R \otimes \Sigma\right).$$

Here R denotes the spatial correlation matrix among components (rows) of T. In the rest of the paper the realization (observed value) of training sample T will be denoted by t.

Denote by r_0 the vector of spatial correlations between Z_0 and observations in T and set $\alpha_0 = R^{-1} r_0$, $\rho = 1 - r_0' \alpha_0$, $\mu_l^0 = \mu_l(s_0)$, $l = 1, \cdots, L$.

Notice that in category Ω_l, the conditional distribution of Z_0 given $T = t$ is Gaussian, *i.e.*

$$\left(Z_0 \middle| T = t, \Omega_l\right) \sim N_p\left(\mu_{lt}^0, \Sigma_{ot}\right),$$

where conditional means μ_{lt}^0 are

$$\mu_{lt}^0 = E\left(Z_0 \middle| T = t; \Omega_l\right) = \mu_l^0 + \left(t - M(B)\right)' \alpha_0, \quad l = 1, \cdots, L \tag{1}$$

and conditional covariance matrix Σ_{0t} is

$$\Sigma_{0t} = V\left(Z_0 \middle| T = t; \Omega_l\right) = \rho \Sigma. \tag{2}$$

The marginal and conditional squared Mahalanobis distances between categories Ω_k and Ω_l $(k, l = 1, \cdots, L)$ for observation taken at location $s = s_0$ are specified respectively by

$$\Delta_{kl}^2 = \left(\mu_k^0 - \mu_l^0\right)' \Sigma^{-1} \left(\mu_k^0 - \mu_l^0\right),$$

and

$$d_{klt}^2 = \left(\mu_{kt}^0 - \mu_{lt}^0\right)' \Sigma_{0t}^{-1} \left(\mu_{kt}^0 - \mu_{lt}^0\right) = \Delta_{kl}^2 / \rho.$$

It is easy to notice that d_{kl} does not depend on realizations of T and depends only on their locations.

Under the assumption of completely parametric certainty of populations and for known prior probabilities of populations π_l, $\sum_{l=1}^{L} \pi_l = 1$, Bayes rule minimizing the probability of misclassification is based on the logarithm of the conditional densities ratio.

There is no loss of generality in focusing attention on category L, since the numbering of the categories is arbitrary. Let the set of population parameters is denoted by $\Psi = \{B, \Sigma\}$. Set $r = L - 1$.

Denote the log ratio of conditional densities in categories Ω_L and Ω_l by

$$W_{Ll}\left(Z_0, \Psi\right) = \left(Z_0 - \left(\mu_{Lt} + \mu_{lt}\right)/2\right)' \Sigma_t^{-1} \left(\mu_{Lt} - \mu_{lt}\right) + \gamma_{Ll}, \tag{3}$$

where $\gamma_{kl} = \ln\left(\pi_k / \pi_l\right)$, $k, l = 1, \cdots, L$.

These functions will be called pairwise discriminant functions (PDF).

Then Bayes rule (BR) (see [10], chapter 6) is given by:

$$\text{classify } Z_0 \text{ to population } \Omega_L \text{ if for } l = 1, \cdots, r, \ W_{Ll}\left(Z_0, \Psi\right) \geq 0. \tag{4}$$

3. Probabilities and Rates of Correct Classification

Set $a_{kl} = \Sigma^{-1/2}\left(\mu_k - \mu_l\right)/\rho$ and set M as r-dimensional vector with the l-th components $(l = 1, \cdots, r)$ specified as $m_l = \left|a_{Ll}\right|^2 \big/ 2 + \gamma_{Ll}$, and $V = \left(v_{lm}, l, m = 1, \cdots, r\right)$ with $v_{lm} = a'_{Ll} a_{Lm}$.

Lemma 1. The conditional probability of correct classification for category L due to BCR specified in (4) is

$$PC_L\left(\Psi\right) = \int\limits_{R_r^+} \varphi_r\left(w; M, V\right) \mathrm{d}w .$$

Here $\varphi_r\left(\cdot\right)$ is the probability density function of *r*-variate normal distribution with mean vector M and variance-covariance matrix V.

Proof. Recall, that under the definition (see e.g. [4] [9]) a probability of correct classification due to aforementioned BCR is

$$PC_k = P_{0t}\left(W_{kl}\left(Z_0; \Psi\right) \geq 0, k, l = 1, \cdots L, l \neq k \middle| \Omega_l\right). \tag{5}$$

It is the probability of correct classification of Z_0 when it comes from Ω_l. Probability measure P_{0t} is based on conditional distribution of Z_0 given $T = t$, Ω_k with means and variance-covariance matrix specified in (1), (2). Z_0 may be expressed in form

$$Z_0 = \Sigma_t^{1/2} U + \mu_{kt},$$

where $U \sim N_p\left(0, I_p\right)$, and I_p denotes the p dimensional identity matrix.

After making the substitution of variables I_p in (5) we obtain that

$$E\left(W_{Ll}\left(Z_0\right)\right) = m_l \text{ and } \operatorname{Cov}\left(W_{Ll}\left(Z_0\right), W_{Lm}\left(Z_0\right)\right) = v_{lm}, \ l, \ m = 1, \cdots, r .$$

Set $W\left(Z_0, \Psi\right) = \left(W_{L1}\left(Z_0, \Psi\right), \cdots, W_{LL-1}\left(Z_0, \Psi\right)\right)$, then probability of correct classification can be rewritten in the following way $PC_L\left(\Psi\right) = P\left(W\left(Z_0, \Psi\right) > 0\right)$.

After straightforward calculations we show that $W\left(Z_0, \Psi\right) \sim N_{L-1}\left(M, \Sigma\right)$. That completes the proof of lemma.

In practical applications not all statistical parameters of populations are known. Then the estimators of unknown parameters can be found from training sample. When estimators of unknown parameters are plugged into Bayes discriminant function (BDF), the plug-in BDF is obtained (PBDF). In this paper we assume that true values of parameters B and Σ are unknown.

Let $\hat{B}$ and $\hat{\Sigma}$ be the estimators of B and Σ based on T. Set $\hat{\Psi} = \left\{\hat{B}, \hat{\Sigma}\right\}$.

Then replacing Ψ by $\hat{\Psi}$ in (3) we get the plug-in BDF (PBDF)

$$W_{Ll}\left(Z_0; \hat{\Psi}\right) = \frac{1}{\rho}\left(Z_0 - \left(T - M\left(\hat{B}\right)\right)' \alpha_0 - \left(\hat{\mu}_L + \hat{\mu}_l\right)/2\right)' \hat{\Sigma}^{-1}\left(\hat{\mu}_L - \hat{\mu}_l\right) + \gamma_{Ll} .$$

Then the classification rule based on PBCR is associated with plug-in PDF (PPDF) in the following way: classify Z_0 to population Ω_k if for $l = 1, \cdots, L$ $W_{kl}\left(Z_0, \hat{\Psi}\right) \geq 0$.

Definition 1. The actual correct classification rate incurred by PBCR associated with PPDF is

$$P\hat{C}_k = P_{0t}\left(W_{kl}\left(Z_0; \hat{\Psi}\right) \geq 0, l = 1, \cdots, L, l \neq k \middle| \Omega_k\right).$$

Set $\hat{a}_{Ll} = \Sigma^{1/2} \hat{\Sigma}^{-1}\left(\hat{\mu}_L - \hat{\mu}_l\right) \big/ \sqrt{\rho}$ and $\hat{b}_{Ll} = \left(\mu_L + \alpha'_0\left(M\left(\hat{B}\right) - M\left(B\right)\right) - \left(\hat{\mu}_L + \hat{\mu}_l\right)/2\right)' \hat{\Sigma}^{-1}\left(\hat{\mu}_L - \hat{\mu}_l\right) \Big/ \rho + \gamma_{Ll}$.

Lemma 2. The actual correct classification rate due to PBDR is

$$PC_L\left(\hat{\Psi}\right) = \int\limits_{R_r^+} \varphi_r\left(w; \hat{M}, \hat{V}\right) \mathrm{d}w ,$$

where $\hat{M}$ is *r*-dimensional vector with components $\hat{m}_l = \hat{b}_{Ll}$, $l = 1, \cdots, r$ and $\hat{V} = \left(\hat{v}_{lm} = \hat{a}'_{Ll} \hat{a}_{Lm}, l, m = 1, \cdots, r\right)$.

Proof. It is obvious that in population Ω_l the conditional distribution of BPDF $W_{Ll}\left(Z_0;\hat{\Psi}\right)$ given $T=t$ is Gaussian, *i.e.*,

$$W_{Ll}\left(Z_0;\hat{\Psi}\right)\Big| T=t,\ \ \Omega_L \sim N\left(\hat{m}_l,\hat{v}_{ll}\right).$$

Set $W\left(Z_0,\hat{\Psi}\right)=\left(W_{L1}\left(Z_0,\hat{\Psi}\right),\cdots,W_{LL-1}\left(Z_0,\hat{\Psi}\right)\right)$, then probability of correct classification can be rewritten in the following way:

$$PC_L\left(\hat{\Psi}\right)=P\left(W\left(Z_0,\hat{\Psi}\right)>0\right).$$

After straightforward calculations we show that $W\left(Z_0,\hat{\Psi}\right)\sim N_{L-1}\left(\hat{M},\hat{V}\right)$. That completes the proof of lemma.

4. Example and Discussions

Simulation study in order to compare proposed Bayes probability of correct classification rate and the actual correct classification rate incurred by PBCR was carried out for three class case $(L=3)$. Also the effect of the range parameter on these values is examined.

In this example, observations are assumed to arise from bivariate stationary Gaussian random field $(p=2)$ with constant mean and isotropic exponential correlation function given by $r(h)=\exp\{-h/\theta\}$, where θ is a parameter of spatial correlation (range).

Set $\mu_1=-\mu_2=1_2$, $\mu_3=0_2$ and $\Sigma=I_2$.

Estimators of B and Σ have the following form:

$$\hat{B}=\hat{\mu}=\left(\hat{\mu}_1,\hat{\mu}_2,\hat{\mu}_3\right)'=\hat{\mu}_{ML}=\left(X'R^{-1}X\right)^{-1}X'R^{-1}T,$$

where X denotes design matrix of training sample T and is specified by $X=1_4\oplus 1_4\oplus 1_4$ and

$$\hat{\Sigma}=\left(T-M\left(\hat{B}\right)\right)'R^{-1}\left(T-M\left(\hat{B}\right)\right)\Big/(n-3).$$

Considered set of training locations with indicated class labels is shown in **Figure 1**.

So we have small training sample sizes (*i.e.* $n_1=n_2=n_3=4$) and three different locations to be classified, furthermore we assume equal prior probabilities $\pi_l=1/3$, $l=1,2,3$.

Simulations were performed by geoR: a free and open-source package for geostatistical analysis included in statistical computing software R (http://www.r-project.org/). Each case was simulated 100 times (runs) and $\overline{\text{ACCR}}$ values are calculated by averaging ACCR over the runs. $\overline{\text{ACCR}}$ and CCR values are presented in **Table 1**. As might be expected $\overline{\text{ACCR}}$ values are lower than CCR. All values are increasing while range parameter is increasing. That means the stronger correlation gives better accuracy of proposed classification procedures.

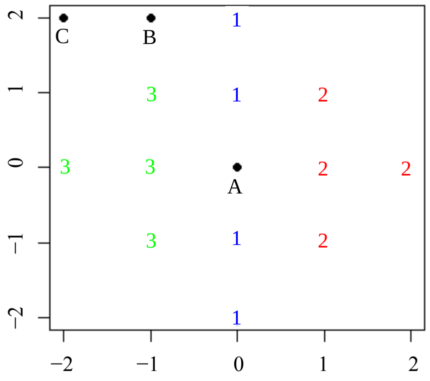

Figure 1. Locations of training sample: “1” are samples from population Ω_1, “2” from Ω_2, “3” from Ω_3, A, B and C denotes the locations of observation to be classified.

Table 1. CCR and $\overline{\text{ACCR}}$ values.

θ	CCR			$\overline{\text{ACCR}}$		
	A (0,0)	B (−1,2)	C (−2,2)	A (0,0)	B (−1,2)	C (−2,2)
1	0.633777	0.634955	0.487689	0.544697	0.427956	0.404184
2	0.882716	0.839326	0.599707	0.636842	0.501045	0.483608
3	0.973217	0.948266	0.716064	0.725356	0.565855	0.480018
4	0.999777	0.955340	0.810298	0.773966	0.639186	0.553055

It's seen in **Figure 1**, that the closest location to be classified is location A and the farthest is location C. CCR and $\overline{\text{ACCR}}$ are largest for location A and smallest for location C. It can be concluded that better accuracy gives closer locations.

References

[1] McLachlan, G.J. (2004) Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York.

[2] Dučinskas, K. (2009) Approximation of the Expected Error Rate in Classification of the Gaussian Random Field Observations. *Statistics and Probability Letters*, **79**, 138-144. http://dx.doi.org/10.1016/j.spl.2008.07.042

[3] Batsidis, A. and Zografos, K. (2011) Errors of Misclassification in Discrimination of Dimensional Coherent Elliptic Random Field Observations. *Statistica Neerlandica*, **65**, 446-461. http://dx.doi.org/10.1111/j.1467-9574.2011.00494.x

[4] Schervish, M.J. (1984) Linear Discrimination for Three Known Normal Populations. *Journal of Statistical Planning and Inference*, **10**, 167-175. http://dx.doi.org/10.1016/0378-3758(84)90068-5

[5] Wu, T.F., Lin, C.J. and Weng, R.C. (2004) Probability Estimates for Multi-Class Classification by Pairwise Coupling. *Journal of Machine Learning Research*, **5**, 975-1005.

[6] Hirst, D. (1996) Error-Rate Estimation in Multiply-Group Linear Discriminant Analysis. *Technometrics*, **38**, 389-399. http://dx.doi.org/10.1080/00401706.1996.10484551

[7] Williams, C.K.I. and Barber, D. (1998) Bayesian Classification with Gaussian Processes. *IEEE Translations on Pattern Analysis and Machine Intelligence*, **20**, 1342-1351. http://dx.doi.org/10.1109/34.735807

[8] Xu, T. and Wang, J. (2013) An Efficient Model-Free Estimation of Multiclass Conditional Probability. *Journal of Statistical Planning and Inference*, **143**, 2079-2088. http://dx.doi.org/10.1016/j.jspi.2013.08.008

[9] Schervish, M.J. (1981) Asymptotic Expansions for the Means and Variances of Error Rates. *Biometrica*, **68**, 295-299. http://dx.doi.org/10.1093/biomet/68.1.295

[10] Anderson, T.W. (2003) An Introduction to Multivariate Statistical Analysis. Wiley, New York.

[11] Šaltytė, J. and Dučinskas, K. (2002) Comparison of ML and OLS Estimators in Discriminant Analysis of Spatially Correlated Observations. *Informatica*, **13**, 297-238.

[12] Šaltytė-Benth, J. and Dučinskas, K. (2005) Linear Discriminant Analysis of Multivariate Spatial-Temporal Regressions. *Scandinavian Journal of Statistics*, **32**, 281-294. http://dx.doi.org/10.1111/j.1467-9469.2005.00421.x

[13] Dučinskas, K. (2011) Error Rates in Classification of Multivariate Gaussian Random Field Observation. *Lithuanian Mathematical Journal*, **51**, 477-485. http://dx.doi.org/10.1007/s10986-011-9142-4

[14] Dučinskas, K. and Dreižienė, L. (2011) Supervised Classification of the Scalar Gaussian Random Field Observations under a Deterministic Spatial Sampling Design. *Austrian Journal of Statistics*, **40**, 25-36.

[15] Dučinskas, K., Dreižienė, L. and Zikarienė, E. (2015) Multiclass Classification of the Scalar Gaussian Random Field Observation with Known Spatial Correlation Function. *Statistics and Probability Letters*, **98**, 107-114. http://dx.doi.org/10.1016/j.spl.2014.12.008

[16] Wackernagel, H. (2003) Multivariate Geostatistics: An Introduction with Applications. Springer-Verlag, Berlin. http://dx.doi.org/10.1007/978-3-662-05294-5

Open Journal of Statistics, 2015, 5, 27-34
Published Online February 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.51004

Separate-Type Estimators for Estimating Population Ratio in Post-Stratified Sampling Using Variable Transformation

Aloy Chijioke Onyeka, Chinyeaka Hostensia Izunobi, Iheanyi Sylvester Iwueze

Department of Statistics, Federal University of Technology, Owerri, Nigeria
Email: aloyonyeka@futo.edu.ng, chiyeaka2007@yahoo.com, isiwueze@yahoo.com

Received 7 January 2015; accepted 25 January 2015; published 30 January 2015

Abstract

The study proposes, along the line of [1], six separate-type estimators for estimating the population ratio of two variables in post-stratified sampling, using variable transformation. Properties of the proposed estimators were obtained up to first order approximations, both for achieved sample configurations (conditional argument) and over repeated samples of fixed size *n* (unconditional argument). Efficiency conditions, under which the proposed separate-type estimators would perform better than the associated customary separate-type estimators in terms of having smaller mean squared errors, were obtained. Furthermore, conditions under which some of the proposed separate-type estimators would perform better than other proposed separate-type estimators were also obtained. The optimum estimators among the proposed separate-type estimators were obtained and an empirical illustration confirmed the theoretical results.

Keywords

Variable Transformation, Separate-Type Estimator, Optimum Estimators, Ratio, Product and Regression-Type Estimators, Mean Squared Error

1. Introduction

Information on auxiliary character has been used by many authors [2]-[9] in sample survey to improve estimates of population parameters of the study variable, and sometimes, information on several variables is used to estimate or predict a characteristic of interest, such as mean, total, ratio, and proportion. Reference [1] proposed the following six (6) estimators of the population ratio $\left(R = \bar{Y}/\bar{X}\right)$ of the population means of two variables, y and x, under the simple random sampling scheme.

How to cite this paper: Onyeka, A.C., Izunobi, C.H. and Iwueze, I.S. (2015) Separate-Type Estimators for Estimating Population Ratio in Post-Stratified Sampling Using Variable Transformation. *Open Journal of Statistics*, **5**, 27-34.
http://dx.doi.org/10.4236/ojs.2015.51004

$$\hat{R}_1 = \frac{\bar{y}}{\bar{x} - b\left(\bar{x}^* - \bar{X}\right)} \quad \left(\text{regression-type estimator of sample mean, } \bar{x}\right) \tag{1.1}$$

$$\hat{R}_2 = \frac{\bar{y}}{\left(\frac{\bar{x}}{\bar{x}^*}\bar{X}\right)} = \frac{\bar{y}\bar{x}^*}{\bar{x}\bar{X}} \quad \left(\text{ratio-type estimator of sample mean, } \bar{x}\right) \tag{1.2}$$

$$\hat{R}_3 = \frac{\bar{y}}{\left(\frac{\bar{x}\bar{x}^*}{\bar{X}}\right)} = \frac{\bar{y}\bar{X}}{\bar{x}\bar{x}^*} \quad \left(\text{product-type estimator of sample mean, } \bar{x}\right) \tag{1.3}$$

$$\hat{R}_4 = \frac{\bar{y}}{\bar{x}^*} \quad \left(\text{transformed mean estimator, } \bar{x}^*\right) \tag{1.4}$$

$$\hat{R}_5 = \frac{\bar{y}}{\bar{x}^* - b\left(\bar{x} - \bar{X}\right)} \quad \left(\text{regression-type estimator of transformed mean, } \bar{x}^*\right) \tag{1.5}$$

$$\hat{R}_6 = \frac{\bar{y}}{\left(\frac{\bar{x}^*}{\bar{x}}\bar{X}\right)} = \frac{\bar{y}\bar{x}}{\bar{x}^*\bar{X}} \quad \left(\text{ratio-type estimator of transformed mean, } \bar{x}^*\right) \tag{1.6}$$

where, $\bar{y}$, $\bar{x}$ and $\bar{x}^*$ are sample means of the variables y_i, x_i and x_i^* respectively,

$$x_i^* = \frac{N\bar{X} - nx_i}{N-n}, \quad i = 1, 2, \cdots, N \tag{1.7}$$

$$\bar{x}^* = \left(1+\pi\right)\bar{X} - \pi\bar{x}, \quad \pi = \frac{n}{N-n} \tag{1.8}$$

and b is a suitable constant, often chosen to be very close to the population regression coefficient of y on x.

Reference [1] noted that authors like [8] [10]-[14] had used the variable transformation (1.7) or its equivalence in their respective studies. The obvious advantage of variable transformation is the introduction of an additional auxiliary (transformed) variable without additional cost, since the new auxiliary variable is a transformation of an already observed auxiliary variable. The work carried out by [1] was restricted to simple random sampling scheme. The present study extends the work carried out by [1] to post-stratified random sampling, by considering six (6) separate-type estimators of the population ratio of two variables in post-stratified random sampling, proposed along the line of the estimators proposed by [1] under the simple random sampling scheme.

2. The Proposed Separate-Type Estimators

Let n units be drawn from a population of N units using simple random sampling method and let the sampled units be allocated to their respective strata, where n_h is the number of units that fall into stratum h such that $\sum_{h=1}^{L} n_h = n$. Let y_{hi} and x_{hi} be the i^{th} observation on the study and auxiliary variables. Consider the following variable transformation of the auxiliary variable, x, under post-stratified sampling scheme.

$$x_{hi}^* = \frac{N\bar{X} - nx_{hi}}{N-n}, \quad h = 1, 2, \cdots, L \ \text{ and } \ i = 1, 2, \cdots, N \tag{2.1}$$

with the associated sample mean

$$\bar{x}_{ps}^* = \left(1-\pi\right)\bar{X} - \pi\bar{x}_{ps} \quad \text{where} \quad \pi = \frac{n}{N-n} \tag{2.2}$$

where $\bar{x}_{ps} = \sum_{h=1}^{L}\omega_h\bar{x}_h$ and $\left(\bar{y}_{ps}\sum_{h=1}^{L}\omega_h\bar{y}_h\right)$ are sample mean estimators based on x_{hi} and y_{hi} respectively. Using the sample means $\bar{y}_{ps}$, $\bar{x}_{ps}$ and $\bar{x}_{ps}^*$, and assuming that the population mean, $\bar{X}$ of the auxiliary variable x, is known, we proposed six separate-type estimators of the population ratio $R = \bar{Y}/\bar{X}$ in post stratified sampling scheme, following [1], as

$$\hat{R}_{1S} = \sum_{h=1}^{L} \omega_h \frac{\bar{y}_h}{\bar{x}_h - b\left(\bar{x}_h^* - \bar{X}_h\right)} \tag{2.3}$$

$$\hat{R}_{2S} = \sum_{h=1}^{L} \omega_h \frac{\bar{y}_h}{\left(\frac{\bar{x}_h}{\bar{x}_h^*} \bar{X}_h\right)} = \sum_{h=1}^{L} \omega_h \left(\frac{\bar{y}_h \bar{x}_h^*}{\bar{x}_h \bar{X}_h}\right) \tag{2.4}$$

$$\hat{R}_{3S} = \sum_{h=1}^{L} \omega_h \frac{\bar{y}_h}{\left(\frac{\bar{x}_h \bar{x}_h^*}{\bar{X}_h}\right)} = \sum_{h=1}^{L} \omega_h \left(\frac{\bar{y}_h \bar{X}_h}{\bar{x}_h \bar{x}_h^*}\right) \tag{2.5}$$

$$\hat{R}_{4S} = \sum_{h=1}^{L} \omega_h \frac{\bar{y}_h}{\bar{x}_h^*} \tag{2.6}$$

$$\hat{R}_{5S} = \sum_{h=1}^{L} \omega_h \frac{\bar{y}_h}{\bar{x}_h^* - b\left(\bar{x}_h - \bar{X}_h\right)} \tag{2.7}$$

$$\hat{R}_{6S} = \sum_{h=1}^{L} \omega_h \frac{\bar{y}_h}{\left(\frac{\bar{x}_h^*}{\bar{x}_h} \bar{X}_h\right)} = \sum_{h=1}^{L} \omega_h \left(\frac{\bar{y}_h \bar{x}_h}{\bar{x}_h^* \bar{X}_h}\right). \tag{2.8}$$

2.1. The Conditional Properties of the Proposed Separate-Type Estimators

Let

$$e_{0h} = \frac{\bar{y}_h - \bar{Y}_h}{\bar{Y}_h} \quad \text{and} \quad e_{1h} = \frac{\bar{x}_h - \bar{X}_h}{\bar{X}_h}. \tag{2.9}$$

Then under the conditional argument,

$$E_2\left(e_{0h}\right) = E\left(e_{1h}\right) = 0 \tag{2.10}$$

$$E_2\left(e_{1h}^2\right) = \frac{V_2\left(\bar{x}_h\right)}{\bar{X}_h^2} = \frac{1}{\bar{X}_h^2}\left(1 - f_h\right)\frac{S_{xh}^2}{n_h} \tag{2.11}$$

$$E_2\left(e_{1h}^2\right) = \frac{V_2\left(\bar{x}_h\right)}{\bar{X}_h^2} = \frac{1}{\bar{X}_h^2}\left(1 - f_h\right)\frac{S_{xh}^2}{n_h} \tag{2.12}$$

$$E_2\left(e_{0h}e_{1h}\right) = \frac{C_2\left(\bar{y}_h, \bar{x}_h\right)}{\bar{Y}_h \bar{X}_h} = \frac{1}{\bar{Y}_h \bar{X}_h}\left(1 - f_h\right)\frac{S_{yxh}}{n_h} \tag{2.13}$$

where E_2 refers to conditional expectation. Notice that the first proposed estimator (2.3) can be rewritten as

$$\hat{R}_{1S} = \sum_{h=1}^{L} \omega_h \hat{R}_{1h} \tag{2.14}$$

where

$$\hat{R}_{1h} = \frac{\bar{y}_h}{\bar{x}_h - b\left(\bar{x}_h^* - \bar{X}_h\right)} \tag{2.15}$$

such that expanding up to first order approximation, $o\left(n^{-1}\right)$, in expected value, we obtain

$$\left(\hat{R}_{1h} - R_h\right) = R_h\left(e_{0h} - \left(1 + b\pi\right)e_{1h} - \left(1 + b\pi\right)e_{0h}e_{1h} + \left(1 + b\pi\right)^2 e_{1h}^2\right) \tag{2.16}$$

and

$$\left(\hat{R}_{1h}-R_h\right)^2=R_h^2\left(e_{0h}^2+\left(1+b\pi\right)^2e_{1h}^2-2\left(1+b\pi\right)e_{0h}e_{1h}\right). \tag{2.17}$$

We take conditional expectation of (2.16) and (2.17) and use (2.10) to (2.13) to make the necessary substitutions to obtain the conditional bias and mean square error of $\hat{R}_{1h}$ respectively as

$$B_2\left(\hat{R}_{1h}\right)=\frac{\left(1-f_h\right)}{n_h}\frac{1}{\bar{X}_h^2}\left[R_h\left(1+b\pi\right)^2S_{xh}^2-\left(1+b\pi\right)S_{yxh}\right] \tag{2.18}$$

and

$$\text{MSE}_2\left(\hat{R}_{1h}\right)=\frac{\left(1-f_h\right)}{n_h}\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+\left(1+b\pi\right)^2R_h^2S_{xh}^2-2\left(1+b\pi\right)R_hS_{yxh}\right] \tag{2.19}$$

so that, using (2.14)

$$B_2\left(\hat{R}_{1S}\right)=\sum_{h=1}^{L}\omega_hB_2\left(\hat{R}_{1h}\right)=\sum_{h=1}^{L}\omega_h\frac{\left(1-f_h\right)}{n_h}\frac{1}{\bar{X}_h^2}\left[R_h\left(1+b\pi\right)^2S_{xh}^2-\left(1+b\pi\right)S_{yxh}\right] \tag{2.20}$$

and

$$\text{MSE}_2\left(\hat{R}_{1S}\right)=\sum_{h=1}^{L}\omega_h^2B_2\left(\hat{R}_{1h}\right)=\sum_{h=1}^{L}\omega_h^2\frac{\left(1-f_h\right)}{n_h}\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+\left(1+b\pi\right)^2R_h^2S_{xh}^2-2\left(1+b\pi\right)R_hS_{yxh}\right]. \tag{2.21}$$

Following similar procedure, we obtain the conditional biases and mean square errors of the six proposed separate-type estimators, together with those of the customary separate-type estimator, $\hat{R}_S=\sum_{h=1}^{L}\omega_h\frac{\bar{y}_h}{\bar{x}_h}$, of population ratio $\left(R\right)$ in post-stratified sampling, up to first order approximation as:

$$B_2\left(\hat{R}_S\right)=\sum_{h=1}^{L}\omega_h\frac{\left(1-f_h\right)}{n_h}\frac{1}{\bar{X}_h^2}\left[R_hS_{xh}^2-S_{yxh}\right] \tag{2.22}$$

$$B_2\left(\hat{R}_{1S}\right)=\sum_{h=1}^{L}\omega_h\frac{\left(1-f_h\right)}{n_h}\frac{1}{\bar{X}_h^2}\left[\left(1+b\pi\right)^2R_hS_{xh}^2-\left(1+b\pi\right)S_{yxh}\right] \tag{2.23}$$

$$B_2\left(\hat{R}_{2S}\right)=\sum_{h=1}^{L}\omega_h\frac{\left(1-f_h\right)}{n_h}\frac{1}{\bar{X}_h^2}\left[\left(1+\pi\right)R_hS_{xh}^2-\left(1+\pi\right)S_{yxh}\right] \tag{2.24}$$

$$B_2\left(\hat{R}_{3S}\right)=\sum_{h=1}^{L}\omega_h\frac{\left(1-f_h\right)}{n_h}\frac{1}{\bar{X}_h^2}\left[\left(1-\pi+\pi^2\right)R_hS_{xh}^2-\left(1-\pi\right)S_{yxh}\right] \tag{2.25}$$

$$B_2\left(\hat{R}_{4S}\right)=\sum_{h=1}^{L}\omega_h\frac{\left(1-f_h\right)}{n_h}\frac{1}{\bar{X}_h^2}\left[R_h\pi^2S_{xh}^2+\pi S_{yxh}\right] \tag{2.26}$$

$$B_2\left(\hat{R}_{5S}\right)=\sum_{h=1}^{L}\omega_h\frac{\left(1-f_h\right)}{n_h}\frac{1}{\bar{X}_h^2}\left[\left(\pi+b\right)^2R_hS_{xh}^2+\left(\pi+b\right)S_{yxh}\right] \tag{2.27}$$

$$B_2\left(\hat{R}_{6S}\right)=\sum_{h=1}^{L}\omega_h\frac{\left(1-f_h\right)}{n_h}\frac{1}{\bar{X}_h^2}\left[\pi\left(1+\pi\right)R_hS_{xh}^2+\left(1+\pi\right)S_{yxh}\right] \tag{2.28}$$

and,

$$\text{MSE}_2\left(\hat{R}_S\right)=\sum_{h=1}^{L}\omega_h^2\frac{\left(1-f_h\right)}{n_h}\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+R_h^2S_{xh}^2-2R_hS_{yxh}\right] \tag{2.29}$$

$$\text{MSE}_2\left(\hat{R}_{1S}\right)=\sum_{h=1}^{L}\omega_h^2\frac{\left(1-f_h\right)}{n_h}\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+\left(1+b\pi\right)^2R_h^2S_{xh}^2-2\left(1+b\pi\right)R_hS_{yxh}\right] \tag{2.30}$$

$$\mathrm{MSE}_2\left(\hat{R}_{2S}\right)=\sum_{h=1}^{L}\omega_h^2\frac{(1-f_h)}{n_h}\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+(1+\pi)^2R_h^2S_{xh}^2-2(1+\pi)R_hS_{yxh}\right] \tag{2.31}$$

$$\mathrm{MSE}_2\left(\hat{R}_{3S}\right)=\sum_{h=1}^{L}\omega_h^2\frac{(1-f_h)}{n_h}\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+(1-\pi)^2R_h^2S_{xh}^2-2(1-\pi)R_hS_{yxh}^2\right] \tag{2.32}$$

$$\mathrm{MSE}_2\left(\hat{R}_{4S}\right)=\sum_{h=1}^{L}\omega_h^2\frac{(1-f_h)}{n_h}\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+\pi^2R_h^2S_{xh}^2+2\pi R_hS_{yxh}\right] \tag{2.33}$$

$$\mathrm{MSE}_2\left(\hat{R}_{5S}\right)=\sum_{h=1}^{L}\omega_h^2\frac{(1-f_h)}{n_h}\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+(b+\pi)^2R_h^2S_{xh}^2+2(b+\pi)R_hS_{yxh}\right] \tag{2.34}$$

$$\mathrm{MSE}_2\left(\hat{R}_{6S}\right)=\sum_{h=1}^{L}\omega_h^2\frac{(1-f_h)}{n_h}\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+(1+\pi)^2R_h^2S_{xh}^2+2(1+\pi)R_hS_{yxh}\right]. \tag{2.35}$$

Generally, the conditional mean square errors of the proposed separate-type estimators are obtained as:

$$\mathrm{MSE}_2\left(\hat{R}_{qS}\right)=\sum_{h=1}^{L}\omega_h^2\frac{(1-f_h)}{n_h}\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+\theta_q^2R_h{}^2S_{xh}^2-2\theta_qR_hS_{yxh}\right] \tag{2.36}$$

where $q=1,\cdots,6$ and

$$\theta_1=(1+b\pi),\ \ \theta_2=(1+\pi),\ \ \theta_3=(1-\pi),\ \ \theta_4=-\pi,\ \ \theta_5=-(\pi+b),\ \ \theta_6=-(1+\pi). \tag{2.37}$$

2.2. The Unconditional Properties of the Proposed Separate-Type Estimators

We take unconditional expectation of the conditional biases and mean square errors of (2.22) to (2.37) to obtain the unconditional properties of the separate type estimators as:

$$B\left(\hat{R}_S\right)=\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left(\frac{1-f_h}{n_h}\right)\left[R_hS_{xh}^2-S_{yxh}\right] \tag{2.38}$$

$$B\left(\hat{R}_{1S}\right)=\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left(\frac{1-f_h}{n_h}\right)\left[(1+b\pi)^2R_hS_{xh}^2-(1+b\pi)S_{yxh}\right] \tag{2.39}$$

$$B\left(\hat{R}_{2S}\right)=\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left(\frac{1-f_h}{n_h}\right)\left[(1+\pi)R_hS_{xh}^2-(1+\pi)S_{yxh}\right] \tag{2.40}$$

$$B\left(\hat{R}_{3S}\right)=\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left(\frac{1-f_h}{n_h}\right)\left[\left(1-\pi+\pi^2\right)R_hS_{xh}^2-(1-\pi)S_{yxh}\right] \tag{2.41}$$

$$B\left(\hat{R}_{4S}\right)=\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left(\frac{1-f_h}{n_h}\right)\left[R_h\pi^2S_{xh}^2+\pi S_{yxh}\right] \tag{2.42}$$

$$B\left(\hat{R}_{5S}\right)=\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left(\frac{1-f_h}{n_h}\right)\left[(b+\pi)^2R_hS_{xh}^2+(b+\pi)S_{yxh}\right] \tag{2.43}$$

$$B\left(\hat{R}_{6S}\right)=\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left(\frac{1-f_h}{n_h}\right)\left[\pi(1+\pi)R_hS_{xh}^2+(1+\pi)S_{yxh}\right] \tag{2.44}$$

and,

$$\mathrm{MSE}\left(\hat{R}_S\right)=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+R_h^2S_{xh}^2-2R_hS_{yxh}\right] \tag{2.45}$$

$$\mathrm{MSE}\left(\hat{R}_{1S}\right)=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+(1+b\pi)^2R_h^2S_{xh}^2-2(1+b\pi)R_hS_{yxh}\right] \tag{2.46}$$

$$\mathrm{MSE}\left(\hat{R}_{2S}\right)=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+(1+\pi)^2R_h^2S_{xh}^2-2(1+\pi)R_hS_{yxh}\right] \tag{2.47}$$

$$\mathrm{MSE}\left(\hat{R}_{3S}\right)=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+(1-\pi)^2R_h^2S_{xh}^2-2(1-\pi)R_hS_{yxh}^2\right] \tag{2.48}$$

$$\mathrm{MSE}\left(\hat{R}_{4S}\right)=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+\pi^2R_h^2S_{xh}^2+2\pi R_hS_{yxh}\right] \tag{2.49}$$

$$\mathrm{MSE}\left(\hat{R}_{5S}\right)=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+(b+\pi)^2R_h^2S_{xh}^2+2(b+\pi)R_hS_{yxh}\right] \tag{2.50}$$

$$\mathrm{MSE}\left(\hat{R}_{6S}\right)=\left(\frac{1-f}{n}\right)\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+(1+\pi)^2R_h^2S_{xh}^2+2(1+\pi)R_hS_{yxh}\right]. \tag{2.51}$$

Generally, the unconditional mean square errors of the proposed separate-type estimators of the population ratio are obtained as:

$$\mathrm{MSE}\left(\hat{R}_{qs}\right)=\left[\frac{1-f}{n}\right]\sum_{h=1}^{L}\omega_h\frac{1}{\bar{X}_h^2}\left[S_{yh}^2+\theta_q^2R_h^2S_{xh}^2-2\theta_qR_hS_{yxh}\right]. \tag{2.52}$$

3. Efficiency Comparison

The efficiencies of the six proposed separate-type estimators, $\hat{R}_{qs}$, were first compared with that of the customary separate-type estimator $\hat{R}_s$ in estimating the population ratio, R, of two population means under the conditional and unconditional arguments in post stratified random sampling scheme. Secondly, the performances of the proposed estimators among themselves were also compared, and finally, the optimum estimators among the proposed estimators were obtained. The efficiency conditions were based on estimators with smaller mean squared errors, and the results are shown in **Table 1**.

4. Numerical Illustration

Here, we use the final year GPA (y) and the level of absenteeism (x) of 2012/2013 graduating students of Statistics department, Federal University of Technology Owerri to illustrate the properties of the estimators proposed in the present study. Absenteeism is the average number of days absent from lectures in a month. The

Table 1. Efficiency conditions under the conditional and unconditional arguments.

Estimator	Conditional argument	Unconditional argument
R_{qs} is better than R_s if:	1) $\lvert\theta_q\rvert<1$ and $\beta''<1$ or 2) $\lvert\theta_q\rvert>1$ and $\beta''>1$	1) $\lvert\theta_q\rvert<1$ and $\beta^*<1$ or 2) $\lvert\theta_q\rvert>1$ and $\beta^*>1$
R_{js} is better than R_{ks} if:	1) $\lvert\theta_j\rvert<\lvert\theta_k\rvert$ and $\lvert\theta_j\rvert<\beta''$ or 2) $\lvert\theta_j\rvert>\lvert\theta_k\rvert$ and $\lvert\theta_j\rvert>\beta''$	1) $\lvert\theta_j\rvert<\lvert\theta_k\rvert$ and $\lvert\theta_j\rvert<\beta^*$ or 2) $\lvert\theta_j\rvert>\lvert\theta_k\rvert$ and $\lvert\theta_j\rvert>\beta^*$
R_{qs} is optimum if:	$\lvert\theta_q^0\rvert=\beta''$	$\lvert\theta_q^0\rvert=\beta^*$

Where $\beta''=\dfrac{\sum_{h=1}^{L}\dfrac{\omega_h^2(1-f_h)S_{yxh}R_h}{n_h\bar{X}_h^2}}{\sum_{h=1}^{L}\dfrac{\omega_h^2(1-f_h)S_{xh}^2R_h^2}{n_h\bar{X}_h^2}}$, $\beta^*=\dfrac{\sum_{h=1}^{L}\dfrac{\omega_hR_hS_{yxh}}{\bar{X}_h^2}}{\sum_{h=1}^{L}\dfrac{\omega_hR_h^2S_{xh}^2}{\bar{X}_h^2}}$, and θ_q, $q=1,\cdots,6$.

class consists of 50 students, with 32 and 18 students respectively falling into low-absenteeism (0 - 3 days per month) and high-absenteeism (4 - 6 days per month) groups or strata. Our interest is to estimate the ratio of final year GPA to absenteeism from lectures, based on a post-stratified sample of 20 out of the 50 students in the class. The data statistics, consisting mainly of population parameters, are shown in **Table 2**.

Table 3 shows the percentage relative efficiencies (PRE-1) of the proposed separate-type estimators, $\hat{R}_{qs}$, over the customary separate-type estimator, $\hat{R}_s$, under the conditional argument and unconditional arguments. The table also shows the percentage relative efficiency (PRE-2) of one of the proposed separate-type estimators, $\hat{R}_{1s}$, over the other separate-type estimators, under the conditional and unconditional arguments.

Table 3 shows that apart from the estimators, $\hat{R}_{2s}$ and $\hat{R}_{6s}$, the remaining four proposed separate-type estimators, under the conditional and unconditional arguments, are more efficient than the customary separate-type estimator, $\hat{R}_s$, for the data under consideration, and their gains in efficiency (PRE-1) are relatively large. Also, using, PRE-2 we observe that the proposed separate-type estimator, $\hat{R}_{1s}$, is more efficient than the estimators, $\hat{R}_{2s}$, $\hat{R}_{6s}$, and $\hat{R}_s$, under the conditional argument and unconditional arguments. The optimum estimator, as expected, has the highest gain in efficiency. However, the customary separate-type estimator is found to be more efficient than some of the proposed separate-type estimators for the given data set. This confirms the theoretical results which shows that the proposed estimators are not always more efficient than the customary separate estimators. Hence, the empirical results confirm the theoretical results.

5. Concluding Remarks

The present study extended the use of variable transformation in estimating population ratio in simple random

Table 2. Data statistics for final year GPA (*y*) and absenteeism from lectures (*x*).

Population/sample parameters	Stratum 1 (low-absenteeism)	Stratum 2 (high-absenteeism)
$N = 50$	$N_1 = 32$	$N_2 = 18$
$n = 20$	$n_1 = 12$	$n_2 = 8$
$(1-f) = 0.60$	$(1-f_1) = 0.625$	$(1-f_2) = 0.556$
$\bar{Y} = 2.98$	$\bar{Y}_1 = 3.16$	$\bar{Y}_2 = 2.65$
$\bar{X} = 3.16$	$\bar{X}_1 = 2.03$	$\bar{X}_2 = 5.17$
$R = 0.94$	$R_1 = 1.56$	$R_2 = 0.51$
$\pi = 0.67$	$S_{y1}^2 = 0.2422$	$S_{y2}^2 = 0.0389$
$b = -0.80$	$S_{x1}^2 = 0.9990$	$S_{x2}^2 = 0.6176$
	$S_{yx1} = -0.2124$	$S_{yx2} = -0.0161$
	$\omega_1 = 0.64$	$\omega_2 = 0.36$

Table 3. Efficiency comparison of proposed separate-type estimators.

Estimators	θ	Conditional argument			Unconditional argument		
		MSE	PRE-1 (%)	PRE-2 (%)	MSE	PRE-1 (%)	PRE-2 (%)
$\hat{R}_{1c}$	0.464	0.00563	311	100	0.00336	311	100
$\hat{R}_{2c}$	1.670	0.04259	41	757	0.02539	41	757
$\hat{R}_{3c}$	0.330	0.00381	459	68	0.00227	459	68
$\hat{R}_{4c}$	−0.670	0.00468	374	83	0.00279	373	83
$\hat{R}_{5c}$	0.130	0.00194	900	35	0.00116	899	35
$\hat{R}_{6c}$	−1.670	0.03103	56	551	0.01850	56	551
$\hat{R}_c$	1.000	0.01748	100	311	0.01042	100	311
$\hat{R}_{qc}^0$		0.00104	1678	19	0.00062	1673	19

sampling scheme to post-stratified sampling scheme where we proposed six separate-type estimators. Efficiency conditions under which the proposed estimators performed better than the customary separate-type estimators were obtained. Both the theoretical and empirical comparisons show that the proposed estimators are not always better or more efficient than the customary separate-type estimator of the population ratio in post-stratified sampling. Consequently, in any given survey, these efficiency conditions should be employed to determine the appropriate separate-type estimators to use for estimating the population ratio of two variables in post-stratified sampling scheme using variable transformation. The major advantage of the proposed estimators is the use of additional (transformed) auxiliary variable without additional cost, since the additional auxiliary variable is a transformation of an already observed auxiliary variable.

References

[1] Onyeka, A.C., Nlebedim, V.U. and Izunobi, C.H. (2013) Estimation of Population Ratio in Simple Random Sampling Using Variable Transformation. *Global Journal of Science Frontier Research*, **13**, 57-65.

[2] Cochran, W.G. (1940) The Estimation of the Yields of the Cereal Experiments by Sampling for the Ratio of Grain to Total Produce. *The Journal of Agricultural Science*, **30**, 262-275. http://dx.doi.org/10.1017/S0021859600048012

[3] Robson, D.S. (1957) Application of Multivariate Polykays to the Theory of Unbiased Ratio-Type Estimation. *Journal of the American Statistical Association*, **52**, 511-522. http://dx.doi.org/10.1080/01621459.1957.10501407

[4] Murthy, M.N. (1964) Product Method of Estimation. *Sankhya*, *Series A*, **26**, 294-307.

[5] Singh, M.P. (1965) On the Estimation of Ratio and Product of the Population Parameters. *Sankhya*, *Series B*, **27**, 321-328.

[6] Sukhatme, P.V. and Sukhatme, B.V. (1970) Sampling Theory of Surveys with Applications. Iowa State University Press, Ames.

[7] Cochran, W.G. (1977) Sampling Techniques. 3rd Edition, John Wiley & Sons, New York.

[8] Onyeka, A.C. (2013) Dual to Ratio Estimators of Population Mean in Post-Stratified Sampling Using Known Value of Some Population Parameters. *Global Journal of Science Frontier Research*, **13**, 13-23.

[9] Onyeka, A.C., Nlebedim, V.U. and Izunobi, C.H. (2014) A Class of Estimators for Population Ratio in Simple Random Sampling Using Variable Transformation. *Open Journal of Statistics*, **4**, 284-291. http://dx.doi.org/10.4236/ojs.2014.44029

[10] Srivenkataramana, T. (1980) A Dual of Ratio Estimator in Sample Surveys. *Biometrika*, **67**, 199-204. http://dx.doi.org/10.1093/biomet/67.1.199

[11] Upadhyaya, L.N., Singh, G.N. and Singh, H.P. (2000) Use of Transformed Auxiliary Variable in the Estimation of Population Ratio in Sample Survey. *Statistics in Transition*, **4**, 1019-1027.

[12] Singh, H.P. and Tailor, R. (2005) Estimation of Finite Population Mean Using Known Correlation Coefficient between Auxiliary Characters. *Statistica*, *Anno LXV*, **4**, 407-418.

[13] Tailor, R. and Sharma, B.K. (2009) A Modified Ratio-Cum-Product Estimator of Finite Population Mean Using Known Coefficient of Variation and Coefficient of Kurtosis. *Statistics in Transition—New Series*, **10**, 15-24.

[14] Sharma, B. and Tailor, R. (2010) A New Ratio-Cum-Dual to Ratio Estimator of Finite Population Mean in Simple Random Sampling. *Global Journal of Science Frontier Research*, **10**, 27-31.

Open Journal of Statistics, 2015, 5, 35-45
Published Online February 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.51005

Implicit Hypotheses Are Hidden Power Droppers in Family-Based Association Studies of Secondary Outcomes

Jean Gaschignard[1,2*], Quentin B. Vincent[1,2], Jean-Philippe Jaïs[1,2,3], Aurélie Cobat[1,2], Alexandre Alcaïs[1,2,4]

[1]Laboratoire de Génétique des Maladies Infectieuses, Institut National de la Santé et de la Recherche Médicale, Paris, France
[2]Université Paris Descartes, Sorbonne Paris Cité, Institut Imagine, Paris, France
[3]Biostatistique et Informatique Médicale, Hôpital Necker, Paris, Farnce
[4]URC, CIC, Necker and Cochin Hospitals, Paris, France
Email: *jean.gaschignard@inserm.fr, alexandre.alcais@inserm.fr

Received 8 January 2015; accepted 26 January 2015; published 30 January 2015

Abstract

Family-based tests of association between a genetic marker and a disease constitute a common design to dissect the genetic architecture of complex traits. The FBAT software is one of the most popular tools to perform such studies. However, researchers are also often interested in the genetic contribution to a more specific manifestation of the phenotype (e.g. severe vs. non-severe form) known as a secondary outcome. Here, what we demonstrate is the limited power of the classical formulation of the FBAT statistic to detect the effect of genetic variants that influence a secondary outcome, in particular when these variants also impact on the onset of the disease, the primary outcome. We prove that this loss of power is driven by an implicit hypothesis, and we propose a derivation of the original FBAT statistic, free from this implicit hypothesis. Finally, we demonstrate analytically that our new statistic is robust and more powerful than FBAT for the detection of association between a genetic variant and a secondary outcome.

Keywords

Family-Based Association Test, FBAT, Genetic Association Studies, Null Hypothesis, Secondary Outcome, Homogeneity Test

[*]Corresponding author.

How to cite this paper: Gaschignard, J., Vincent, Q.B., Jaïs, J.-P., Cobat, A. and Alcaïs, A. (2015) Implicit Hypotheses Are Hidden Power Droppers in Family-Based Association Studies of Secondary Outcomes. *Open Journal of Statistics*, **5**, 35-45.
http://dx.doi.org/10.4236/ojs.2015.51005

1. Introduction

The aim of genetic epidemiological studies is to identify the genetic factors influencing the development of common diseases. Genetic epidemiology combines classical epidemiological data (assessment of risk factors known to affect the expression of the phenotype studied) and genetic information (familial relationships, typing of genetic marker) and proposes a large range of tools to address the initial question, the use of one depending on the nature of your sample and the size of your wallet. Over the past ten years, however, our understanding of the pattern of genetic variation at the genome scale, coupled to an unprecedented decrease in the cost of measuring this variation, has put (genome-wide) association studies at the front. Although the vast majority of genetic association study designs are derived from usual case-control retrospective epidemiological studies (*i.e.* that compare the distribution of allelic/genotypic frequencies between a group of cases and a group of controls), one is quite specific to the field of genetic epidemiology and relies on the collection and analysis of families. Such family-based tests of association between a genetic item (allele, genotype...) and the disease under study offer interesting features as compared to case-control designs (Laird and Lange [1]; Chen and Abecasis [2]). They are robust against population stratification, allow the inference of both haplotype phase and missing genotypes (Chen and Abecasis [2]; Burdick *et al.* [3]), and can identify peculiar allelic segregation, for example, due to imprinting effect (Vincent *et al*. [4]).

The Transmission Desequilibrium Test (TDT) has emerged as the first popular family-based test of association (Spielman *et al.* [5]). It tests whether the transmission of a given allele from a heterozygote parent to an affected child is different from what is expected in the absence of any association between the genetic marker and the disease under study. The null hypothesis is written as $p = 0.5$ where p is the proportion of a given allele that has been transmitted to affected children by heterozygote parents. Whereas the TDT could only analyze binary traits in samples of pure trios (*i.e.* two parents and a single affected child), Laird *et al.* [6] proposed a more comprehensive approach designed to handle binary, quantitative or censored traits, multiple genetic models (e.g. additive, dominant or recessive) and more complex family structures (e.g. families with multiple children). This approach uses a natural measure of association between two variables, *i.e.* the covariance between phenotypes and genotypes, and relies on a score-test. It has been implemented in the popular *Family Based Association Test* software (FBAT, Laird *et al.* [6]; Rabinowitz and Laird [7]; Lange and Laird [8]). In this context of familial samples, FBAT has proved very efficient in identifying alleles associated with many phenotypes, whether binary or quantitative (e.g. Mira *et al.* [9]; Cobat *et al.* [10]).

Although developed to handle a large variety of tests according to the nature of both the traits and their genetic determinants, it is intrinsically designed to test primary outcomes (e.g. affected vs. unaffected) as the null hypothesis is based on the same underlying principles as the TDT (*i.e.* $p = 0.5$). However, in many cases researchers are interested in the genetic contribution to a more specific phenotype (e.g. severe vs. non-severe form), here denoted as a secondary outcome. Here, what we demonstrate is the limited power of the classical formulation of the FBAT statistic to detect the effect of genetic variants that influence a secondary outcome, in particular when these variants also impact on the onset of the disease, the primary outcome. We prove that this loss of power is driven by an implicit hypothesis and we propose a derivation of the original FBAT statistic, free from this implicit hypothesis. Finally, we demonstrate analytically that our new statistic is robust and more powerful than FBAT for the detection of association between a genetic variant and a secondary outcome.

2. Original FBAT Statistic

For sake of simplicity and without major loss of generality, we consider the analysis of a diallelic marker in a sample of trios with no missing parental data under an additive genetic model. Using the same notations as in the original FBAT paper (Laird *et al.* [6]),

$$\text{let } S = \sum_i T_i X_i$$

in which X_i represents the genotype at the locus being tested and T_i the phenotype of the child of family i. The expectation of X_i is calculated conditioned on the parental genotypes under the null hypothesis of no association.

$$\text{Let } E = E(S) = \sum_i E_i = \sum_i T_i E(X_i)$$

$$\text{Let } V = \text{Var}(S) = \sum_i T_i^2 \text{Var}(X_i)$$

$$\text{FBAT} = \frac{(S-E)^2}{V}$$

$$\text{FBAT} \underset{H_0}{\sim} \chi^2_{1df}$$

Under an additive model, X_i is the number of copy of the allele under study (0, 1 or 2). As the most common way to code the phenotype is $T = 1$ for affected individuals and $T = 0$ for unaffected ones. In a sample with no missing parental data, unaffected individuals do not contribute to the statistic; however, in the presence of missing parental data, such unaffected individuals will indirectly impact on the statistic as they can be used to infer missing parental genotypes under some conditions (Knapp [11]). S is generally written as:

$$S = \sum_{i \in \text{affected}} 1 \times X_i + \sum_{i \in \text{unaffected}} 0 \times X_i = \sum_{i \in \text{affected}} X_i.$$

The null hypothesis of no association between the phenotype and a given allele is the random transmission of this allele from heterozygote parents to (affected) children. By noting p the transmission probability of this allele, the null H_0 and alternate H_1 hypotheses can be written as:

$$H_0 : p = \frac{1}{2}$$

$$H_1 : p \neq \frac{1}{2}.$$

The tested allele will be considered "at risk" or "protective" for the disease, if $p > \frac{1}{2}$ or $p < \frac{1}{2}$, respectively[1].

3. FBAT Statistic to Test Secondary Outcomes

It is common practice to study a "primary" phenotype (e.g. disease yes/no) but as stated in the introduction, researchers are often interested in the genetic contribution to a "secondary" phenotype (e.g. severe vs. non-severe form of the disease). At first glance, FBAT could be used to test this hypothesis by computing the original statistic independently in the two modalities of the secondary outcome (e.g. severe and non-severe). Denoting D_1 and D_2 the two modalities of the secondary outcome, p_1 and p_2 the transmission probabilities of the tested allele to D_1 and D_2 children, respectively, we have:

$$S_1 = \sum_{i \in 1} T_i X_i, \quad S_1 = \sum_{i \in 1} X_i, \quad \text{FBAT}_1 = \frac{(S_1 - E_1)^2}{V_1}$$

$$H_0 : p_1 = \frac{1}{2}$$

$$H_1 : p_1 \neq \frac{1}{2}$$

$$S_2 = \sum_{i \in 2} T_i X_i, \quad S_2 = \sum_{i \in 2} X_i, \quad \text{FBAT}_2 = \frac{(S_2 - E_2)^2}{V_2}$$

$$H_0 : p_2 = \frac{1}{2}$$

$$H_1 : p_2 \neq \frac{1}{2}.$$

[1] More precisely, in the general case, the null hypothesis of FBAT is "no association OR no linkage" and therefore the alternate hypothesis is "association AND linkage". H_0 can be written as a composite hypothesis: "no association AND no linkage" ∪ "no association AND linkage ∪ "association AND no linkage". In the particular case of a sample limited to trios, there is no linkage information, and the hypotheses are: H_0 = association, H_1 = no association.

However, because of the bivariate nature of the phenotype under study (*i.e.* disease AND severe form or disease AND non-severe form), rejection of the null hypothesis cannot distinguish between alleles associated with the disease *per se* (*i.e.* independently of its severity) or alleles specifically associated with the severity of the disease. FBAT offers no immediate solution to study such secondary outcomes, *i.e.* to distinguish between alleles impacting the primary (e.g. disease *per se*) or the secondary (e.g. severe vs. non-severe) outcome. Below we propose two new tests denoted as FBAT$_{\text{het}}$ and FBAT$_{\text{het free}}$ that can be used to directly assess the association between a marker allele and a secondary outcome.

3.1. The FBAT$_{\text{het}}$ Test

A first straightforward idea is to perform a homogeneity test of the allelic transmission rate between the two subgroups D_1 and D_2.

$$\text{Let FBAT}_{\text{het}}\left(D_1,D_2\right)=\text{homogeneity}\left(S_1,S_2\right)$$

$$=\frac{\left(S_1-E_1\right)^2}{V_1}+\frac{\left(S_2-E_2\right)^2}{V_2}-\frac{\left(S_1-E_1+S_2-E_2\right)^2}{V_1+V_2}$$

$$\text{FBAT}_{\text{het}}=\frac{\left(\frac{S_1-E_1}{V_1}-\frac{S_2-E_2}{V_2}\right)^2}{\frac{1}{V_1}+\frac{1}{V_2}}=\frac{\left(\frac{S_1-E_1}{V_1}-\frac{S_2-E_2}{V_2}\right)^2}{\frac{1}{V_1^2}V_1+\frac{1}{V_2^2}V_2}$$

FBAT$_{\text{het}}$ = FBAT with the phenotypes coded as $T=\frac{1}{V_1}$ for individuals D_1 and $T=-\frac{1}{V_2}$ for individuals D_2.

Indeed,

$$S\left(T_1=\frac{1}{V_1},T_2=-\frac{1}{V_2}\right)=\sum_{i\in 1}\frac{1}{V_1}X_i-\sum_{i\in 2}\frac{1}{V_2}X_i=\frac{1}{V_1}S_1-\frac{1}{V_2}S_2$$

$$E=\frac{1}{V_1}E_1-\frac{1}{V_2}E_2 \quad\text{and}\quad V=\frac{1}{V_1^2}V_1+\frac{1}{V_2^2}V_2$$

$$\text{and}\quad \text{FBAT}\left(T_1=\frac{1}{V_1},T_2=-\frac{1}{V_2}\right)=\frac{\left(\frac{S_1-E_1}{V_1}-\frac{S_2-E_2}{V_2}\right)^2}{\frac{1}{V_1}+\frac{1}{V_2}}=\text{FBAT}_{\text{het}}.$$

The two hypotheses can then be written as:

$$H_0: p_1=p_2=\frac{1}{2}$$

$$H_1: p_1\neq\frac{1}{2}\cup p_2\neq\frac{1}{2}.$$

Note that under an additive genetic model and in a sample of trios with no missing parental data, coding $T_1=\frac{1}{V_1}$ and $T_2=-\frac{1}{V_2}$ is equivalent to coding $T_1=\frac{1}{n_1}$ and $T_2=-\frac{1}{n_2}$, where n_1 and n_2 are the number of heterozygote parents of children with phenotype D_1 and D_2 (see Appendix A)[2].

[2] FBAT$_{\text{het}}$ can be implemented in FBAT by using the offset option "-o" while coding $T_1=1$ and $T_2=0$: the software then calculates, for each allele, an offset μ used to transform the phenotypic values in $T_1=1-\mu$ and $T_2=-\mu$ that minimizes the variance of the statistics. We show in Appendix B that using the offset option is equivalent to coding $T_1=\frac{1}{V_1}$ and $T_2=-\frac{1}{V_2}$, thus testing for secondary outcome. Here, one should not code unaffected individuals as 0 but as missing to avoid that the controls interfere in the calculation of the statistics. FBAT software can be downloaded from: http://www.biostat.harvard.edu/fbat/fbat.htm.

3.2. The FBAT$_{\text{het free}}$ Test

A somewhat hidden/under evaluated constraint of FBAT$_{\text{het}}$ is that the null hypothesis forces the transmission probabilities in both groups to be 0.5. Although valid and likely efficient in quite a number of practical situations, this can dramatically impact the power of the test in the study of a secondary outcome. A simple example being that carrying one copy of the allele is sufficient to develop the disease *per se* but that carrying two alleles will be associated with developing a severe form of the disease.

We propose a new statistic denoted as FBAT$_{\text{het free}}$ that relaxes this 0.5 constraint. Consider a diallelic locus (A and a) and denote n_{A1} (n_{A2}) the number of transmissions of allele A from Aa heterozygote parents to their children with phenotype D_1 (D_2). Then $\frac{n_{A1}+n_{A2}}{n_1+n_2}=\frac{n_A}{N}$ is the mean number of transmission of allele A from Aa heterozygote parents to affected children (whether D_1 or D_2).

Whereas in the above-mentioned FBAT and FBAT$_{\text{het}}$ tests the expected transmission of the allele of interest under the null hypothesis of no association is 0.5, in FBAT$_{\text{het free}}$ it is $\frac{n_A}{N}$. We can calculate S, E and V for FBAT, FBAT$_{\text{het}}$ and FBAT$_{\text{het free}}$. The contribution to $S-E$ of each transmission of an allele A from any Aa parent is 1/2 in FBAT and FBAT$_{\text{het}}$, and $\frac{n_A}{N}$ in FBAT$_{\text{het free}}$. Similarly, its contribution to V is 1/4 in FBAT and FBAT$_{\text{het}}$, and $\left(1-\frac{n_A}{N}\right)\frac{n_A}{N}$ in FBAT$_{\text{het free}}$ (**Figure 1**). Note that for all three statistics, the expectancy and variance of a trio including two heterozygote parents are twice those of a trio with only one heterozygote parent. Symmetrically, Aa heterozygote parents transmitting allele a each contributes for 1/2 and $\left(1-\frac{n_A}{N}\right)$ to $S-E$, and for 1/4 and $\left(1-\frac{n_A}{N}\right)\frac{n_A}{N}$ to V in FBAT or FBAT$_{\text{het}}$ and FBAT$_{\text{het free}}$, respectively. Then with $T_1=\frac{1}{n_1}$ and $T_2=-\frac{1}{n_2}$, we have:

$$\text{FBAT}_{\text{het free}}=\frac{\left(\frac{n_{A1}}{n_1}\left(1-\frac{n_A}{N}\right)+\frac{n_{a1}}{n_1}\left(0-\frac{n_A}{N}\right)-\frac{n_{A2}}{n_2}\left(1-\frac{n_A}{N}\right)-\frac{n_{a2}}{n_2}\left(0-\frac{n_A}{N}\right)\right)^2}{\frac{n_1}{n_1^2}\frac{n_An_a}{N^2}+\frac{n_2}{n_2^2}\frac{n_An_a}{N^2}}$$

$$=\frac{N}{n_1n_2n_An_a}\left(n_{A1}n_{a2}-n_{a1}n_{A2}\right)^2.$$

It is shown in Appendix C that FBAT$_{\text{het free}}$ is a Pearson's chi-squared test. In summary, the hypotheses of the FBAT$_{\text{het free}}$ test can be written as:

$$H_0: p_1=p_2$$

$$H_1: p_1\neq p_2.$$

As opposed to FBAT and FBAT$_{\text{het}}$, the implicit/hidden 0.5 constraint has disappeared.

3.3. Comparison of FBAT$_{\text{het}}$ and FBAT$_{\text{het free}}$

To illustrate the magnitude of the differential power of FBAT$_{\text{het}}$ and FBAT$_{\text{het free}}$, we could have gone for large simulation studies. However, we show analytically in Appendix D that:

$$\text{FBAT}_{\text{het}}=\rho\text{FBAT}_{\text{hetfree}}\quad\text{with}\quad\rho=\frac{4n_A\left(N-n_A\right)}{N^2},\ \rho\in[0,1].$$

Trio with 1 heterozygote parent
Aa
aa
Trio with 2 heterozygote parents
Aa
Aa

Child genotype		*aa*	*Aa*	*AA*
	X_i	0	1	2
FBAT &FBAT$_{\text{het}}$	$p(X_i)$	$\frac{1}{2}$	$\frac{1}{2}$	0
	$X_i - E(X)$	$-\frac{1}{2}$	$\frac{1}{2}$	
FBAT$_{\text{het free}}$	$p(X_i)$	$1-\frac{n_A}{N}$	$\frac{n_A}{N}$	0
	$X_i - E(X)$	$-\frac{n_A}{N}$	$1-\frac{n_A}{N}$	

Child genotype		*aa*	*Aa*	*AA*
	X_i	0	1	2
FBAT &FBAT$_{\text{het}}$	$p(X_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
	$X_i - E(X)$	-1	0	1
FBAT$_{\text{het free}}$	$p(X_i)$	$\left(1-\frac{n_A}{N}\right)^2$	$2\left(1-\frac{n_A}{N}\right)\frac{n_A}{N}$	$\left(\frac{n_A}{N}\right)^2$
	$X_i - E(X)$	$-\frac{2n_A}{N}$	$1-\frac{2n_A}{N}$	$2\left(1-\frac{n_A}{N}\right)$

For FBAT and FBAT$_{\text{het}}$ $E(X)=\frac{1}{2}$ $E(X)=1$

$\operatorname{Var}(X)=\frac{1}{4}$ $\operatorname{Var}(X)=\frac{1}{2}$

For FBAT$_{\text{het free}}$, $E(X) = \frac{n_A}{N}$ $E(X)=\frac{2n_A}{N}$

$\operatorname{Var}(X)=\left(1-\frac{n_A}{N}\right)\frac{n_A}{N}$ $\operatorname{Var}(X)=2\left(1-\frac{n_A}{N}\right)\frac{n_A}{N}$

Figure 1. Contribution of a trio to FBAT, FBAT$_{\text{het}}$ and FBAT$_{\text{het free}}$ according to the number of heterozygote parents. In a trio with one (left panel) and two (right panel) heterozygote parents, the expected genotypes *aa*, *Aa* and *AA* of the child will vary according to the statistics used. In FBAT and FBAT$_{\text{het}}$, the transmission probability of an allele A from an heterozygote parent is $\frac{1}{2}$, whereas it is $\frac{n_A}{N}$ for FBAT$_{\text{het free}}$ (with *N* denoting the total number of alleles transmitted from heterozygote parents in the whole sample, n_A the number of alleles A transmitted, and $\frac{n_A}{N}$ the mean transmission of allele *A*).

The distribution of ρ according to $\frac{n_A}{N}$ is shown in **Figure 2**. As an example, consider a sample of 300 trios with an affected child (150 D_1 and 150 D_2), all with one herterozygote parent. Consider the mean transmission of allele *A* is 0.7 in D_1 and 0.8 in D_2. Then $\frac{n_A}{N}=0.75$, $\rho=0.75$, $\text{FBAT}_{\text{het}}=3$ and $\text{FBAT}_{\text{het free}}=4$, $p\left(\text{FBAT}_{\text{het}}\right)=0.083$ and $p\left(\text{FBAT}_{\text{het free}}\right)=0.046$.

When there is an equivalent number of transmissions of alleles A and a from Aa heterozygote parents to their children, $n_A=n_a=\frac{N}{2}$ and $\rho=1$. In practice, this is observed when the mean transmission of allele

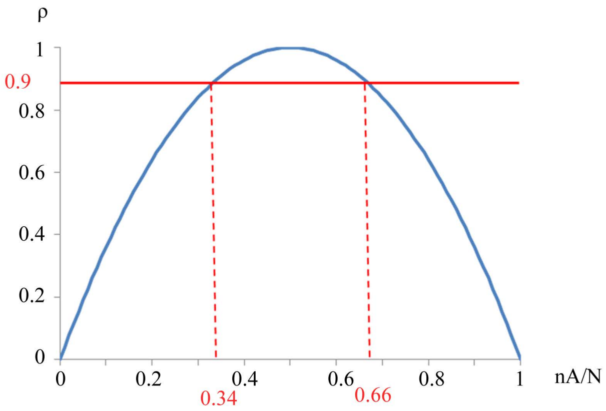

Figure 2. Distribution of ρ according to $\frac{n_A}{N}$. $\rho = \frac{4n_A(N-n_A)}{N^2}$ is the link function between FBAT$_{\text{het}}$ and FBAT$_{\text{het free}}$. When the mean transmission of allele A among affected cases is close to 0.5, ρ is also close from 1. When $\frac{n_A}{N} \in [0.34; 0.66]$, $\rho > 0.9$.

A among all affected individuals $(D_1 + D_2)$ is 0.5. In that particular case, $\text{FBAT}_{\text{het}} = \text{FBAT}_{\text{het free}}$. In all other cases, $\rho < 1$ and $\text{FBAT}_{\text{het}} < \text{FBAT}_{\text{het free}}$ as shown in **Figure 3**.

4. Discussion

Family-based association studies have gained popularity to dissect the genetic architecture of complex traits and FBAT is likely the most popular tool to perform such studies. We have shown that at first glance it can be conveniently used to test for secondary outcomes, e.g. genetic heterogeneity between severe and non-severe forms of a disease. As an example, in a sample of trios, one can weight each "sub-phenotype" (severe and non-severe) by the inverse of the variance of each statistic. We called this test FBAT$_{\text{het}}$, for which the null and alternative hypotheses are $H_0 : p_1 = p_2 = \frac{1}{2}$ and $H_1 : p_1 \neq \frac{1}{2}$ or $p_2 \neq \frac{1}{2}$, respectively.

However, in the previous test, the transmission probabilities under the null hypothesis are fixed to 0.5 in both groups. This may not be optimal in the context of secondary outcomes when the transmission of the tested allele has already been found to significantly differ from 0.5 with respect to the primary outcome. We show that it is possible to relax this constraint by modifying the expectation in the FBAT$_{\text{het}}$ statistic so that the test is defined as $H_0 : p_1 = p_2$ and $H_1 : p_1 \neq p_2$, which are the classical hypotheses in the vast majority of homogeneity tests. This new test, FBAT$_{\text{het free}}$, is proven to be equivalent to a classical test for homogeneity. FBAT$_{\text{het free}}$ is the most powerful test when the mean transmission to affected children ($D_1 + D_2$, primary outcome) is not 0.5. Stated differently, each time an allele is found associated with the disease *per se*, FBAT$_{\text{het free}}$ will be the most powerful to detect heterogeneity between the transmission rates of this allele across the modalities of the secondary outcome.

For sake of simplicity, we have derived our main statistic FBAT$_{\text{het free}}$ in the context of the analysis of a diallelic marker under an additive genetic model in a sample of trios with no missing parental data. However, generalization to other genetic models and more complex family structures should be possible by using, for a given marker, the estimated mean transmission of the allele under study among affected individuals, in preference to the actual 0.5 that prevents testing $p_1 = p_2$. By doing so, one will be able to take advantage of all the features of FBAT ranging from the analysis of all kinds of phenotypes to the simultaneous testing of several alleles either in a classic multivariate way or taking into account the phase through haplotypic analysis.

Acknowledgements

We thank Laurent Abel, Jean-Laurent Casanova and all members of the Epidemiological Group for their support

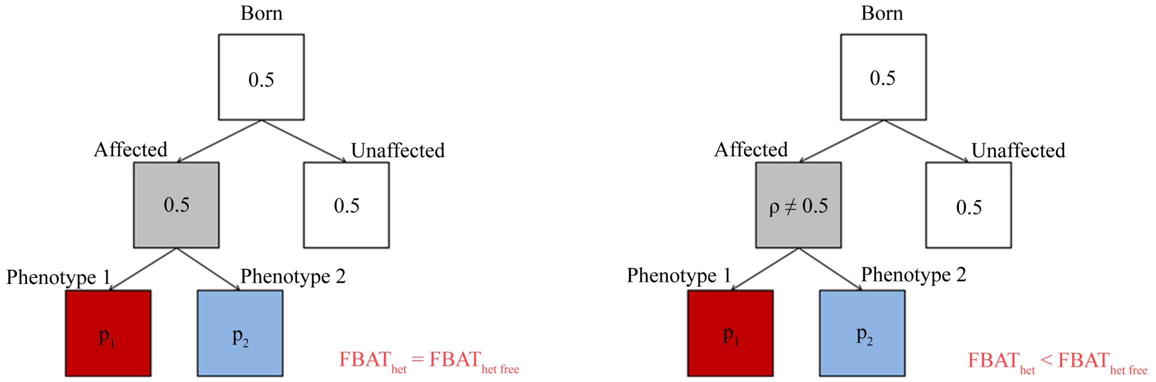

Figure 3. Power of $FBAT_{het}$ vs. $FBAT_{het\ free}$ according to the mean transmission rate of the tested allele among the affected children.

and constructive criticism. JG is funded by the Fondation pour la Recherche Médicale, and QV by the Institut Imagine. This work was supported by the Programme Blanc de l'Agence National de la Recherche.

References

[1] Laird, N.M. and Lange, C. (2006) Family-Based Designs in the Age of Large-Scale Gene-Association Studies. *Nature Reviews Genetics*, **7**, 385-394. http://dx.doi.org/10.1038/nrg1839

[2] Chen, W.M. and Abecasis, G.R. (2007) Family-Based Association Tests for Genomewide Association Scans. *American Journal of Human Genetics*, **81**, 913-926. http://dx.doi.org/10.1086/521580

[3] Burdick, J.T., Chen, W.M., Abecasis, G.R. and Cheung, V.G. (2006) In Silico Methods for Inferring Genotypes in Pedigrees. *Nature Genetics*, **38**, 1002-1004. http://dx.doi.org/10.1038/ng1863

[4] Vincent, Q., Alcais, A., Alter, A., Schurr, E. and Abel, L. (2006) Quantifying Genomic Imprinting in the Presence of Linkage. *Biometrics*, **62**, 1071-1080. http://dx.doi.org/10.1111/j.1541-0420.2006.00610.x

[5] Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993) Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-Dependent Diabetes Mellitus (IDDM). *American Journal of Human Genetics*, **52**, 506-516.

[6] Laird, N.M., Horvath, S. and Xu, X. (2000) Implementing a Unified Approach to Family-Based Tests of Association. *Genetic Epidemiology*, **19**, S36-S42. http://dx.doi.org/10.1002/1098-2272(2000)19:1+<::AID-GEPI6>3.0.CO;2-M

[7] Rabinowitz, D. and Laird, N. (2000) A Unified Approach to Adjusting Association Tests for Population Admixture with Arbitrary Pedigree Structure and Arbitrary Missing Marker Information. *Human Heredity*, **50**, 211-223. http://dx.doi.org/10.1159/000022918

[8] Lange, C. and Laird, N.M. (2002) Power Calculations for a General Class of Family-Based Association Tests: Dichotomous Traits. *American Journal of Human Genetics*, **71**, 575-584. http://dx.doi.org/10.1086/342406

[9] Mira, M.T., Alcais, A., Van Thuc, N., Moraes, M.O., Di Flumeri, C., Hong Thai, V., Chi Phuong, M., Thu Huong, N., Ngoc Ba, N., Xuan Khoa, P., *et al*. (2004) Susceptibility to Leprosy Is Associated with PARK2 and PACRG. *Nature*, **427**, 636-640. http://dx.doi.org/10.1038/nature02326

[10] Cobat, A., Gallant, C.J., Simkin, L., Black, G.F., Stanley, K., Hughes, J., Doherty, T.M., Hanekom, W.A., Eley, B., Jais, J.P., *et al*. (2009) Two Loci Control Tuberculin Skin Test Reactivity in an Area Hyperendemic for Tuberculosis. *Journal of Experimental Medicine*, **206**, 2583-2591. http://dx.doi.org/10.1084/jem.20090892

[11] Knapp, M. (1999) The Transmission/Disequilibrium Test and Parental-Genotype Reconstruction: The Reconstruction-Combined Transmission/Disequilibrium Test. *American Journal of Human Genetics*, **64**, 861-870. http://dx.doi.org/10.1086/302285

Appendix A. Proof That Coding $T_1 = \frac{1}{V_1}$ and $T_2 = -\frac{1}{V_2}$ Is Equivalent to $T_1 = \frac{1}{n_1}$ and $T_2 = -\frac{1}{n_2}$ under an Additive Genetic Model

Let N_1 and N_2 be the number of trios with phenotype D_1 and D_2, and N_{id} (N_{is}) the number of trios with double (d) or single (s) heterozygote parent (s). Let n_i be the number of heterozygote parents. Then

$$n_i = 2N_{id} + N_{is}.$$

Let V_s and V_d be the unitary variance for trios with 1 or 2 heterozygote parents.

For FBAT and FBAT$_{\text{het}}$, $V_s = \frac{1}{4}$ and $V_d = \frac{1}{2} = 2V_s$. Then

$$V_1 = \sum_{\substack{\&\text{pheno 1}\\ \text{1 parent}}} \operatorname{Var}(S_j) + \sum_{\substack{\&\text{pheno 1}\\ \text{2 parents}}} \operatorname{Var}(S_j) = N_{1s}V_s + N_{1d}V_d = \frac{N_{1s}}{4} + \frac{N_{1d}}{2} = \frac{n_1}{4}$$

$$\text{and } V_2 = \sum_{\substack{\&\text{pheno 2}\\ \text{1 parent}}} \operatorname{Var}(S_j) + \sum_{\substack{\&\text{pheno 2}\\ \text{2 parents}}} \operatorname{Var}(S_j) = \frac{n_2}{4}.$$

Given that $\text{FBAT}(T_1 = x, T_2 = y) = \text{FBAT}(T_1 = kx, T_2 = ky)$, coding $T_1 = \frac{1}{V_1}$ and $T_2 = -\frac{1}{V_2}$ is equivalent to $T_1 = \frac{1}{n_1}$ and $T_2 = -\frac{1}{n_2}$ for FBAT and FBAT$_{\text{het}}$.

For FBAT$_{\text{het free}}$, $V_s = \left(1 - \frac{n_A}{N}\right)\frac{n_A}{N}$ and $V_d = 2\left(1 - \frac{n_A}{N}\right)\frac{n_A}{N} = 2V_s$. Then

$$V_1 = \sum_{\substack{\&\text{pheno 1}\\ \text{1 parent}}} \operatorname{Var}(S_j) + \sum_{\substack{\&\text{pheno 1}\\ \text{2 parents}}} \operatorname{Var}(S_j) = (N_{1s} + 2N_{1d})\left(1 - \frac{n_A}{N}\right)\frac{n_A}{N} = n_1\left(1 - \frac{n_A}{N}\right)\frac{n_A}{N}$$

$$\text{and } V_2 = \sum_{\substack{\&\text{pheno 2}\\ \text{1 parent}}} \operatorname{Var}(S_j) + \sum_{\substack{\&\text{pheno 2}\\ \text{2 parents}}} \operatorname{Var}(S_j) = n_2\left(1 - \frac{n_A}{N}\right)\frac{n_A}{N}.$$

Then coding $T_1 = \frac{1}{V_1}$ and $T_2 = -\frac{1}{V_2}$ is also equivalent to $T_1 = \frac{1}{n_1}$ and $T_2 = -\frac{1}{n_2}$ for FBAT$_{\text{het free}}$.

Appendix B. Proof That $\mu = \frac{n_1}{n_1 + n_2}$ Is the Offset That Minimizes the Variance under an Additive Genetic Model

Let μ be the offset.

$$T_1 = 1 - \mu \quad \text{and} \quad T_2 = -\mu$$

With the same notations as in Appendix A,

$$\begin{aligned}
\operatorname{Var} &= \sum_{\substack{\&\text{pheno 1}\\ \text{1 parent}}} \operatorname{Var}(S_j) + \sum_{\substack{\&\text{pheno 1}\\ \text{2 parents}}} \operatorname{Var}(S_j) + \sum_{\substack{\&\text{pheno 2}\\ \text{1 parent}}} \operatorname{Var}(S_j) + \sum_{\substack{\&\text{pheno 2}\\ \text{2 parents}}} \operatorname{Var}(S_j) \\
&= N_{1s}T_1^2V_s + N_{1d}T_1^2V_d + N_{2s}T_2^2V_s + N_{2d}T_2^2V_d \\
&= (1-\mu)^2 (N_{1s}V_s + N_{1d}V_d) + (-\mu)^2 (N_{2s}V_s + N_{2d}V_d).
\end{aligned}$$

For FBAT, $V_s = \frac{1}{4}$, $V_d = \frac{1}{2} = 2V_s$ and

$$\mathrm{Var} = \frac{1}{4}\left((1-\mu)^2 n_1 + \mu^2 n_2\right)$$

and $\min_{\mu}(\mathrm{Var}) = \min_{\mu}\left((1-\mu)^2 n_1 + \mu^2 n_2\right)$ is obtained for $\mu = \frac{n_1}{n_1+n_2}$.

For $\mathrm{FBAT}_{\text{het free}}$, $V_s = \left(1-\frac{n_A}{N}\right)\frac{n_A}{N}$, $V_d = 2\left(1-\frac{n_A}{N}\right)\frac{n_A}{N} = 2V_s$ and

$$\mathrm{Var} = \left(1-\frac{n_A}{N}\right)\frac{n_A}{N}\left((1-\mu)^2 n_1 + \mu^2 n_2\right)$$

and $\min_{\mu}(\mathrm{Var}) = \min_{\mu}\left((1-\mu)^2 n_1 + \mu^2 n_2\right)$ is also obtained for $\mu = \frac{n_1}{n_1+n_2}$.

Appendix C. Proof That FBAT_{het free} Is a Pearson's χ^2

With the notations of the manuscript, let us write the table of contingency of the transmission of alleles *A* and *a* in two phenotypic groups.

Transmission	*A*	*a*	Total
D_1	n_{A1}	n_{a1}	n_1
D_2	n_{A2}	n_{a2}	n_2
	n_A	n_a	N

Pearson's $\chi^2(n_{A1}, n_{a1}, n_{a2}, n_{A2})$

$$= \frac{\left(n_{A1} - \frac{n_A n_1}{N}\right)^2}{\frac{n_A n_1}{N}} + \frac{\left(n_{a1} - \frac{n_a n_1}{N}\right)^2}{\frac{n_a n_1}{N}} + \frac{\left(n_{A2} - \frac{n_A n_2}{N}\right)^2}{\frac{n_A n_2}{N}} + \frac{\left(n_{a2} - \frac{n_a n_2}{N}\right)^2}{\frac{n_a n_2}{N}}$$

$$= \frac{(n_{A1}N - n_A n_1)^2}{N n_A n_1} + \frac{(n_{a1}N - n_a n_1)^2}{N n_a n_1} + \frac{(n_{A2}N - n_A n_2)^2}{N n_A n_2} + \frac{(n_{a2}N - n_a n_2)^2}{N n_a n_2}$$

$$= \left(\frac{1}{N n_A n_1} + \frac{1}{N n_a n_1} + \frac{1}{N n_A n_2} + \frac{1}{N n_a n_2}\right)(n_{A1}n_{a2} - n_{a1}n_{A2})^2$$

$$= \frac{N}{n_1 n_2 n_A n_a}(n_{A1}n_{a2} - n_{a1}n_{A2})^2$$

$$= \mathrm{FBAT}_{\text{het free}}.$$

Appendix D. Proof That FBAT_{free} = ρFBAT_{het free}

With the notations used in the main text, for $\mathrm{FBAT}_{\text{het}}$,

$$S - E = \frac{1}{n_1}\left(n_{A1}\left(1-\frac{1}{2}\right) + n_{a1}\left(0-\frac{1}{2}\right)\right) - \frac{1}{n_2}\left(n_{A2}\times\left(1-\frac{1}{2}\right) + n_{a1}\left(0-\frac{1}{2}\right)\right)$$

and $V = \frac{1}{4}\left(\frac{1}{n_1}\right)^2 (n_{A1} + n_{a1}) + \frac{1}{4}\left(\frac{1}{n_2}\right)^2 (n_{A2} + n_{a2}) = \left(\frac{1}{n_1}\right)^2 \frac{n_1}{4} + \left(\frac{1}{n_2}\right)^2 \frac{n_2}{4}$.

Then $\mathrm{FBAT}_{\mathrm{het}} = \frac{(S-E)^2}{V} = \frac{\left(\frac{n_{A1}}{n_1}\left(1-\frac{1}{2}\right)+\frac{n_{a1}}{n_1}\left(0-\frac{1}{2}\right)-\frac{n_{A2}}{n_2}\left(1-\frac{1}{2}\right)-\frac{n_{a2}}{n_2}\left(0-\frac{1}{2}\right)\right)^2}{\frac{1}{4}\frac{n_1}{n_1^2}+\frac{1}{4}\frac{n_2}{n_2^2}}$

$$= \frac{4\left(n_{A1}n_{a2}-n_{a1}n_{A2}\right)^2}{Nn_1n_2} = \frac{4n_An_a}{N^2}\frac{N\left(n_{A1}n_{a2}-n_{a1}n_{A2}\right)^2}{n_An_an_1n_2} = \rho\mathrm{FBAT}_{\text{het free}},$$

with $\rho = \frac{4n_A\left(N-n_A\right)}{N^2}$.

Open Journal of Statistics, 2015, 5, 46-50
Published Online February 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.51006

Statistical Significance of Geographic Heterogeneity Measures in Spatial Epidemiologic Studies

Min Lian[1,2]

[1]Department of Medicine, Washington University School of Medicine, St. Louis, Missouri, USA
[2]Alvin J. Siteman Cancer Center, Barnes-Jewish Hospital and Washington University School of Medicine, St. Louis, Missouri, USA
Email: mlian@wustl.edu

Received 16 January 2015; accepted 3 February 2015; published 6 February 2015

Abstract

Assessing geographic variations in health events is one of the major tasks in spatial epidemiologic studies. Geographic variation in a health event can be estimated using the neighborhood-level variance that is derived from a generalized mixed linear model or a Bayesian spatial hierarchical model. Two novel heterogeneity measures, including median odds ratio and interquartile odds ratio, have been developed to quantify the magnitude of geographic variations and facilitate the data interpretation. However, the statistical significance of geographic heterogeneity measures was inaccurately estimated in previous epidemiologic studies that reported two-sided 95% confidence intervals based on standard error of the variance or 95% credible intervals with a range from 2.5th to 97.5th percentiles of the Bayesian posterior distribution. Given the mathematical algorithms of heterogeneity measures, the statistical significance of geographic variation should be evaluated using a one-tailed *P* value. Therefore, previous studies using two-tailed 95% confidence intervals based on a standard error of the variance may have underestimated the geographic variation in events of their interest and those using 95% Bayesian credible intervals may need to re-evaluate the geographic variation of their study outcomes.

Keywords

Spatial Epidemiology, Heterogeneity, Statistical Significance, 95% Confidence Interval, 95% Credible Interval

1. Introduction

Spatial epidemiology is an important methodology to deal with spatial-correlated issues in epidemiologic studies.

How to cite this paper: Lian, M. (2015) Statistical Significance of Geographic Heterogeneity Measures in Spatial Epidemiologic Studies. *Open Journal of Statistics*, **5**, 46-50. http://dx.doi.org/10.4236/ojs.2015.51006

One of its core tasks is to determine geographic variations and quantify the magnitude of geographic variations in diseases, health behaviors, or environmental exposures [1]. Some published epidemiologic studies inappropriately estimated the statistical significance of geographic heterogeneity measures of examined events.

The generalized linear mixed model and the Bayesian spatial hierarchical model are the most commonly applied to fit the data with a multilevel spatial structure. A geographic variation can be directly quantified as neighborhood-level variance $\left(\sigma^2\right)$ from parameter estimations of the multilevel model fitting. However, this variance has no meaningful unit and is difficult to interpret. Spatial statisticians and epidemiologists have developed two state-of-the-art heterogeneity measures, the median odds ratio (MOR, Equation (1)) [2]-[4] and the interquartile odds ratio (IqOR, Equation (2)) [5], to facilitate the interpretation of geographic heterogeneity of an event.

$$\begin{aligned}\text{MOR} &= \exp\left(Z_{0.75}\times\sqrt{2\times\text{VAR}}\right)\\ &= \exp\left(0.9539\times\sqrt{\text{VAR}}\right),\end{aligned} \tag{1}$$

where VAR is the neighborhood-level variance, while $Z_{0.75}$ is the Z value of the Gaussian distribution at the 75th percentile (0.6745).

$$\begin{aligned}\text{IqOR} &= \exp\left(\left(Z_{0.875}-Z_{0.125}\right)\times\sqrt{\text{VAR}}\right)\\ &= \exp\left(2.3007\times\sqrt{\text{VAR}}\right),\end{aligned} \tag{2}$$

where $Z_{0.875}$ and $Z_{0.125}$ are the Z values of the Gaussian distribution at the 87.5th and 12.5th percentiles (1.1504, −1.1504), respectively.

Both MOR and IqOR are derived from the variance and are always greater than or equal to one. Larger values of MOR and IqOR denote greater geographic variations in the event of interest. The MOR reflects the average difference of risk when comparing two subjects who have the same individual characteristics and are selected randomly from two different neighborhoods. The IqOR represents the average difference of risk when comparing the first quartile of study subjects residing in neighborhoods with the highest risk to the fourth quartile of study subjects residing in neighborhoods with the lowest risk [3] [5]. Similarly, the median rate ratio (MRR) and the interquartile rate ratio (IqRR) can be estimated in a prospective study, and the median hazards ratio (MHR) and the interquartile hazard ratio (IqHR) [6] are for time-to-event studies. To facilitate the explanation, the MOR and IqOR are applied in the following discussions.

2. Issues in Determining the Statistical Significance of Geographic Heterogeneity Measures

Geographic variations can be qualitatively assessed by using neighborhood-level variance estimation derived from a generalized linear mixed model. The modeling conducted by a commonly used statistical analysis package, such as the SAS, also gives a *Z* value and a corresponding *P* value based on an approximately normal distribution of the estimated parameter. With the standard error of the variance from the multilevel model fitting, a 95% CI is able to be computed mathematically. However, one cannot perform a generalized linear mixed analysis to estimate the statistical significance and 95% CIs of the MOR and IqOR because both MOR and IqOR are derived from the variance and do not have their own standard errors.

Alternatively, a Bayesian spatial hierarchical model with a Markov Chain Monte Carlo (MCMC) simulation has been used to estimate geographic heterogeneities. In this setting, the 95% Bayesian credible interval (CrI), defined by the 2.5th and 97.5th percentiles of Bayesian posterior distribution of the geographic heterogeneity measure, has been commonly reported.

In the estimation of a fixed effect of an exposure, its statistical significance can be identified if the 95% confidence/credible interval of its regression coefficient does not cross zero. However, this empirical method conflicts with the nature of geographic heterogeneity measures. Two unreasonable results are usually reported in the studies in which the 95% CI or CrI of geographic heterogeneity measures were used to determine their statistical significance. The 95% CI of the variance could cross zero based on an approximately normal distribution $\left(\bar{X}\pm1.96\times\text{SE}\right)$. This is unreasonable because the variance should always be greater than or equal to 0. In addition, the 2.5th percentile of the Bayesian posterior distribution of the variance is always greater than 0 and consequently the MOR and IqOR are always greater than one. This leads to the overestimation of geographic disparities.

3. Example and Solution

3.1. Example

A simulation analysis was performed to illustrate the issues relevant to the statistical significance of spatial heterogeneity measures. It is assumed that a population of colorectal cancer (CRC) survivors come randomly from 100 neighborhoods, each with 5 - 20 patients, and that the probability of smoking for each patient is 0.2 - 0.5. A random simulation generated a dataset that included 1245 patients and 420 smokers. Multilevel logistic regression is applied to quantify small-area geographic variation in smoking behavior among these CRC patients (Equation (3)).

$$\begin{aligned} y_{ij} &\sim \text{Binomial}\left(p_{ij}\right) \\ \text{logit}\left(p_{ij}\right) &= \beta_0 + \beta_1 N_i + \beta_2 X_{ij} + u_i \end{aligned} \tag{3}$$

where p_{ij} is the probability of smoking for patient j who resides in neighborhood i; β_0 is the intercept; β_1 and β_2 are the fixed coefficients of neighborhood- and individual-level covariates, respectively; N_i is characteristics of neighborhood *i*; and X_{ij} is a vector of individual-level covariates; u_i is the random effect between neighborhoods with a normal assumption: $u_i \sim N\left(0,\sigma^2\right)$.

To simplify the explanation, an empty model without neighborhood- and individual-level covariates was fit to estimate the overall geographic heterogeneity of smoking among these CRC patients using the Bayesian hierarchical approach with a MCMC simulation in WinBUGS (Version 1.4.3, MRC, UK). After 50,000 iterations for the convergence, additional 50,000 iterations were run to obtain the posterior estimates of three spatial heterogeneity measures. Because the dataset was simulated randomly, the geographic variation in smoking was expected to be small.

Table 1 shows the Bayesian parameter estimates of three heterogeneity measures. Based on an approximately normal assumption, the 95% CIs of three geographic measures were computed as $\mu \pm 1.96 \times \sigma$. Alternatively, the 95% CrIs of three geographic measures were expressed as the range from their 2.5th to their 97.5th percentiles. However, the inconsistent results were observed when comparing the 95% CIs of the variance, MOR and IqOR to their 95% CrIs. The 95% CI of the variance crossed zero and the 95% CIs of both MOR and IqOR crossed 1, suggesting no significant geographic variation in smoking behavior among CRC survivors. In contrast, the 95% CrI of the variance was more than zero and the 95% CrIs of the MOR and IqOR were greater than one, suggesting a significant geographic variation in smoking behavior.

3.2. Solution

Table 2 shows that, the variance is a non-negative measure, and MOR and IqOR are never less than one. The null hypothesis of the statistical test should be that the variance equals to zero and both MOR and IqOR equal to one, that is, there is no significant geographic variation in the event of interest. Meanwhile, the alternative hypothesis of the statistical test should be that the variance is greater than zero, and both MOR and IqOR are greater than one. Therefore, the statistical test is theoretically one-tailed, rather than two-tailed. The critical value for the significance level at 0.05 is 1.645 instead of 1.960. The statistical significance should be denoted directly using one-tailed (right-tailed) *P* value. One may not report the 95% CI or the interval between the 2.5th

Table 1. Three Bayesian estimates of three spatial heterogeneity measures in the simulated example.

Measure	Mean (μ)	SD (σ)	2.50%	Median	97.50%	95% confidence interval	95% credible interval
VAR*	0.007	0.009	0.001	0.004	0.033	0.007 (−0.011, 0.025)	0.004 (0.001, 0.033)
MOR†	1.074	0.045	1.022	1.061	1.189	1.074 (0.986, 1.162)	1.061 (1.022, 1.189)
IqOR‡	1.190	0.125	1.054	1.155	1.517	1.190 (0.945, 1.435)	1.155 (1.054, 1.517)

*VAR, variance; †MOR, median odds ratio; ‡IqOR, interquartile odds ratio.

Table 2. Three spatial heterogeneity measures and their statistical hypotheses.

Measure	Range	Null hypothesis (H_0)	Alternative hypothesis (H_1)
VAR*	≥0	VAR = 0	VAR > 0
MOR†	≥1	MOR = 1	MOR > 1
IqOR‡	≥1	IqOR = 1	IqOR > 1

*VAR, variance; †MOR, median odds ratio; ‡IqOR, interquartile odds ratio.

and the 97.5th percentiles of Bayesian posterior distribution (95% CrI) of geographic heterogeneity measures to avoid the misinterpretation of geographic variations. In fact, a one-tailed *P* value for the variation/heterogeneity estimation has been given from a generalized linear mixed model fitting using common statistical analysis packages, such as the SAS. For the heterogeneity estimation from a Bayesian hierarchical model, one should compute the corresponding statistics, based on the prior distribution of the variance, to obtain their one-tailed *P* value to determine its statistical significance. In the simulated example, since the *Z* value for the variance is: $(0.007-0)/0.009=0.778$, the geographic variation in smoking among CRC survivors is not statistically significant using 1.645 as the cutoff for the significance level at 0.05.

4. Discussions

The purpose of this study was to point out an inappropriate method that was used to determine the statistical significance of geographic heterogeneity measures. The simulated data suggested that empirically reporting of the 95% CI/CrI of geographic heterogeneity measures may lead to misunderstanding of the statistical significance of geographic variations of an event.

According to the nature of geographic heterogeneity measures, the statistical inference should be one-tailed (right-tailed). It is inappropriate to report a two-tailed 95% CI/CrI of a heterogeneity measure in spatial epidemiologic studies. It could mislead one in understanding the statistical significances of heterogeneity measures. In the studies using standard errors to obtain two-tailed *P* values or 95% CIs, geographic variations in the events may be underestimated because a two-tailed test is more conservative than a one-tailed test. In contrast, the studies using the interval between the 2.5th and the 97.5th percentiles of a Bayesian posterior distribution to obtain a 95% CrI may overestimate the statistical significance of geographic variation of the event because a Bayesian 95% CrI never crosses zero for the variance and one for both MOR and IqOR. The issue of statistical significance of geographic heterogeneity measures, which was discussed in this paper, is also extendible to a general multilevel study aiming to investigate the variation(s) in one or multiple event(s) of interest across a non-spatial higher level, such as healthcare providers or medical service facilities.

Acknowledgements

This work is supported partly by a Career Development Award (K07 CA178331) and a Research Award (R21 CA169807) funded by the National Cancer Institute, and a Research Award (R01 AA021492) funded by the National Institute of Alcohol Abuse and Alcoholism, both at the National Institutes of Health. I also thank the service provided by the Health Behavior, Communication and Outreach Core, which is supported by the National Cancer Institute Cancer Center Support grant (P30 CA091842) awarded to the Alvin J. Siteman Cancer Center at Barnes-Jewish Hospital and Washington University School of Medicine. I declare no conflict of interest.

References

[1] Elliott, P. and Wartenberg, D. (2004) Spatial Epidemiology: Current Approaches and Future Challenges. *Environmental Health Perspectives*, **112**, 998-1006. http://dx.doi.org/10.1289/ehp.6735

[2] Larsen, K., Petersen, J.H., Budtz-Jorgensen, E. and Endahl, L. (2000) Interpreting Parameters in the Logistic Regression Model with Random Effects. *Biometrics*, **56**, 909-914. http://dx.doi.org/10.1111/j.0006-341X.2000.00909.x

[3] Larsen, K. and Merlo, J. (2005) Appropriate Assessment of Neighborhood Effects on Individual Health: Integrating Random and Fixed Effects in Multilevel Logistic Regression. *American Journal of Epidemiology*, **161**, 81-88. http://dx.doi.org/10.1093/aje/kwi017

[4] Merlo, J., Chaix, B., Ohlsson, H., Beckman, A., Johnell, K., Hjerpe, P., Rastam, L. and Larsen, K. (2006) A Brief Conceptual Tutorial of Multilevel Analysis in Social Epidemiology: Using Measures of Clustering in Multilevel Logistic Regression to Investigate Contextual Phenomena. *Journal of Epidemiology and Community Health*, **60**, 290-297. http://dx.doi.org/10.1136/jech.2004.029454

[5] Chaix, B., Merlo, J., Subramanian, S.V., Lynch, J. and Chauvin, P. (2005) Comparison of a Spatial Perspective with the Multilevel Analytical Approach in Neighborhood Studies: The Case of Mental and Behavioral Disorders Due to Psychoactive Substance Use in Malmo, Sweden, 2001. *American Journal of Epidemiology*, **162**, 171-182. http://dx.doi.org/10.1093/aje/kwi175

[6] Chaix, B., Rosvall, M. and Merlo, J. (2007) Assessment of the Magnitude of Geographical Variations and Socioeconomic Contextual Effects on Ischaemic Heart Disease Mortality: A Multilevel Survival Analysis of a Large Swedish Cohort. *Journal of Epidemiology and Community Health*, **61**, 349-355. http://dx.doi.org/10.1136/jech.2006.047597

Abbreviations and Acronyms

VAR: variance;
MOR: median odds ratio;
MRR: median rate ratio;
MHR: median hazard ratio;
IqOR: interquartile odds ratio;
IqRR: interquartile rate ratio;
IqHR: interquartile hazard ratio.

Open Journal of Statistics, 2015, 5, 51-59
Published Online February 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.51007

Combining Likelihood Information from Independent Investigations

L. Jiang, A. Wong

Department of Mathematics and Statistics, York University, Toronto, Canada
Email: august@mathstat.yorku.ca, august@yorku.ca

Received 23 January 2015; accepted 11 February 2015; published 15 February 2015

Abstract

Fisher [1] proposed a simple method to combine *p*-values from independent investigations without using detailed information of the original data. In recent years, likelihood-based asymptotic methods have been developed to produce highly accurate *p*-values. These likelihood-based methods generally required the likelihood function and the standardized maximum likelihood estimates departure calculated in the canonical parameter scale. In this paper, a method is proposed to obtain a *p*-value by combining the likelihood functions and the standardized maximum likelihood estimates departure of independent investigations for testing a scalar parameter of interest. Examples are presented to illustrate the application of the proposed method and simulation studies are performed to compare the accuracy of the proposed method with Fisher's method.

Keywords

Canonical Parameter, Fisher's Expected Information, Modified Signed Log-Likelihood Ratio Statistic, Standardized Maximum Likelihood Estimate Departure

1. Introduction

Supposed that k independent investigations are conducted to test the same null hypothesis and the *p*-values are $p_1,\cdots,p_k$ respectively. Fisher [1] proposed a simple method to combine these *p*-values to obtain a single *p*-value (p) without using the detailed information concerning the original data nor knowing how these *p*-values were obtained. His methodology is based on the following two results from distribution theories:

1) If U is distributed as Uniform(0, 1), then $-2\log U$ is distributed as Chi-square with 2 degrees of freedom $\left(\chi_2^2\right)$;

2) If $X_1,\cdots,X_k$ are independently distributed as $\chi_{v_1}^2,\cdots,\chi_{v_k}^2$, then $X_1+\cdots+X_k$ is distributed as $\chi_{v_1+\cdots+v_k}^2$.

Since $p_1,\cdots,p_k$ are independently distributed as Uniform(0, 1), then the combined *p*-value p is

How to cite this paper: Jiang, L. and Wong, A. (2015) Combining Likelihood Information from Independent Investigations. *Open Journal of Statistics*, **5**, 51-59. http://dx.doi.org/10.4236/ojs.2015.51007

$$p = P\left(\chi^2_{2k} \geq -2\sum_{i=1}^{k} \log p_i\right). \tag{1}$$

For illustration, Fisher [1] reported the *p*-values of three independent investigations: 0.145, 0.263 and 0.087. Thus the combined *p*-value is

$$p = P\left(\chi^2_{2(3)} \geq -2\left[\log(0.145) + \log(0.263) + \log(0.087)\right]\right) = P\left(\chi^2_6 \geq 11.417\right) = 0.0763$$

which gives moderate evidence against the null hypothesis. Fisher [1] described the procedure as a "simple test of the significance of the aggregate".

As an illustrative example is the study of rate of arrival. It is common to use a Poisson model to model the number of arrivals over a specific time interval. Let $X_1, \cdots, X_n$ be the number of arrivals in *n* consecutive unit time intervals and denote $x = \sum_{i=1}^{n} X_i$ be the total number of arrivals over the *n* consecutive unit time intervals. Moreover, let θ be the rate of arrival in an unit time interval. We observed a total of 14 arrivals over 20 consecutive unit time intervals. In other words, $n = 20$, $x^0 = 14$ and we are interested in assessing $\theta = 1$. Then the null distribution of X is Poisson (20) and, based on the observed $x^0 = 14$, the mid-*p*-value is

$$p_1(1) = \sum_{i=1}^{13} \frac{e^{-20} 20^i}{i!} + \frac{1}{2}\frac{e^{-20} 20^{14}}{14!} = 0.0855.$$

An alternate way of investigating the rate of arrival over a period of time is by modeling the time to first arrival, *T* with the exponential model with rate θ. We observed $t^0 = 2$, and, again, we are interested in assessing $\theta = 1$. Then the null distribution of T is the exponential with rate 1, and, based on the observed $t^0 = 2$, the *p*-value is

$$p_2(1) = P(T > 1) = \exp(-2(1)) = 0.1353.$$

By Fisher's way of combining the *p*-values, we have

$$P\left(\chi^2_2 > -2\left[\log\left(p_1(1)\right) + \log\left(p_2(1)\right)\right]\right) = 0.0116$$

which gives strong evidence that θ is greater than 1.

In recent years, many likelihood-based asymptotic methods have been developed to produce highly accurate *p*-values. In particular, both the Lugannani and Rice's [2] method and the Barndorff-Nielsen's [3] [4] method produced *p*-values which have third-order accuracy, *i.e.* the rate of convergence is $O\left(n^{-3/2}\right)$. Fraser and Reid [5] showed that both methods required the signed log-likelihood ratio statistic and the standardized maximum likelihood estimate departure calculated in the canonical parameter scale. In this paper, we proposed a method to combine likelihood functions and the standardized maximum likelihood estimates departure calculated in the canonical parameter scale obtained from independent investigations to obtain a combined *p*-value.

In Section 2, a brief review of the third-order likelihood-based method for a scalar parameter of interest is presented. In Section 3, the relationship between the score variable and the locally defined canonical parameter is determined. Using the results in Section 3, a new way of combining likelihood information is proposed in Section 4. Examples and simulation results are presented in Section 5 and some concluding remarks are recorded in Section 6.

2. Third-Order Likelihood-Based Method for a Scalar Parameter of Interest

Fraser [6] showed that for a sample $x^0 = \left(x_1^0, \cdots, x_n^0\right)$ from a canonical exponential family model with log-likelihood function

$$\ell(\varphi) = \ell\left(\varphi; x^0\right) = \ell\left(\varphi; x_1^0, \cdots, x_n^0\right) = \log \prod_{i=1}^{n} f\left(x_i^0; \varphi\right)$$

where

$$f(x; \varphi) = \exp\left\{\varphi t(x) - c(\varphi)\right\} h(x)$$

and φ is the scalar canonical parameter of interest. The *p*-value function $p(\varphi) = P\left(\hat{\varphi} \leq \hat{\varphi}^0; \varphi\right)$ can be approximated with third-order accuracy using either the Lugannani and Rice [2] formula

$$p(\varphi)=\Phi(r)+\phi(r)\left\{\frac{1}{r}-\frac{1}{q}\right\} \tag{2}$$

or the Barndorff-Nielsen [3] [4] formula

$$p(\varphi)=\Phi\left(r+\frac{1}{r}\log\frac{q}{r}\right) \tag{3}$$

where r is the signed log-likelihood ratio statistic

$$r=r(\varphi)=\operatorname{sign}\left(\hat{\varphi}^0-\varphi\right)\left\{2\left[\ell\left(\hat{\varphi}^0\right)-\ell(\varphi)\right]\right\}^{1/2} \tag{4}$$

q is the standardized maximum likelihood departure calculated in the canonical parameter scale:

$$q=q(\varphi)=\left(\hat{\varphi}^0-\varphi\right)j_{\varphi\varphi}^{1/2}\left(\hat{\varphi}^0\right) \tag{5}$$

$\hat{\varphi}^0$ is the maximum likelihood estimate of φ satisfying $\left.\frac{\mathrm{d}\ell(\varphi)}{\mathrm{d}\varphi}\right|_{\hat{\varphi}^0}=0$, and

$$j_{\varphi\varphi}\left(\hat{\varphi}^0\right)=-\left.\frac{\mathrm{d}^2\ell(\varphi)}{\mathrm{d}\varphi^2}\right|_{\hat{\varphi}^0}$$

is the observed information evaluated at $\hat{\varphi}^0$. Jensen [7] showed that (2) and (3) are asymptotically equivalent up to third-order accuracy. In literature, there exists many applications of these methods, for example, see Brazzale *et al*. [8].

Fraser and Reid [5] [9] generalized the methodology to any model with log likelihood function $\ell(\theta)=\ell(\theta;x)$. They defined the locally defined canonical parameter be

$$\varphi=\varphi(\theta)=\left.\frac{\mathrm{d}\ell(\theta)}{\mathrm{d}V}\right|_{x^0}=\left.\frac{\mathrm{d}\ell(\theta)}{\mathrm{d}x}\cdot V\right|_{x^0} \tag{6}$$

where

$$V=\left.\frac{\mathrm{d}x}{\mathrm{d}\theta}\right|_{\left(x^0,\hat{\theta}^0\right)}=-\left.\left\{\frac{\partial z(\theta,x)}{\partial x}\right\}^{-1}\left\{\frac{\partial z(\theta,x)}{\partial\theta}\right\}\right|_{\left(x^0,\hat{\theta}^0\right)} \tag{7}$$

is the rate of change of x with respect to the change of θ at $\left(x^0,\hat{\theta}^0\right)$, and $z(\theta,x)$ is a pivotal quantity. Define s be the score variable satisfying

$$s=\left.\frac{\mathrm{d}\ell(\theta)}{\mathrm{d}\varphi}\right|_{\hat{\theta}^0} \tag{8}$$

with $\hat{\theta}^0$ being the maximum likelihood estimate of θ obtained from $\ell(\theta)$ at the observed data point x^0. The signed log-likelihood ratio statistic r is

$$r=r(\theta)=\operatorname{sign}\left(\hat{\theta}^0-\theta\right)\left\{2\left[\ell\left(\hat{\theta}^0\right)-\ell(\theta)\right]\right\}^{1/2} \tag{9}$$

and the standardized maximum likelihood departure $q(\theta)$ re-calibrated in the φ scale is

$$q=q(\theta)=\left(\hat{\varphi}^0-\varphi\right)j_{\varphi\varphi}^{1/2}\left(\hat{\varphi}^0\right).$$

Since $\hat{\varphi}^0=\varphi\left(\hat{\theta}^0\right)$, by applying the chain rule in differentiation, we have

$$j_{\varphi\varphi}\left(\hat{\varphi}^0\right)=j_{\theta\theta}\left(\hat{\theta}^0\right)\left[\varphi_\theta\left(\hat{\theta}^0\right)\right]^{-2}$$

where $\varphi_\theta\left(\hat{\theta}^0\right)=\left.\frac{\mathrm{d}\varphi(\theta)}{\mathrm{d}\theta}\right|_{\hat{\theta}^0}$. Therefore, $q(\theta)$ can be written as

$$q = q(\theta) = \left(\hat{\varphi}^0 - \varphi\right)\left\{ j_{\theta\theta}\left(\hat{\theta}^0\right)\left[\varphi_\theta\left(\hat{\theta}^0\right)\right]^{-2}\right\}^{1/2}. \tag{10}$$

Applications of the general method discussed above can be found is Reid and Fraser [10] and Davison *et al*. [11].

Note that V in (7) can be viewed as the sensitivity direction and is examined in Fraser *et al*. [12] for the study of the sensitivity analysis of the third-order method. And $\left.\frac{\mathrm{d}s}{\mathrm{d}\theta}\right|_{\left(s^0,\hat{\theta}^0\right)}$ gives the rate of change of the score variable with respect to the change of θ at the observed data point in the tangent exponential model.

3. Relationship between the Score Variable and the Locally Defined Canonical Parameter

In Bayesian analysis, Jeffreys [13] proposed to use the prior density which is proportional to the square root of the Fisher's expected information. This prior is invariant under reparameterization. In other words, the scalar parameter

$$\beta = \beta(\theta) = \int \left[E\left(-\frac{\mathrm{d}^2\ell(\gamma)}{\mathrm{d}\gamma^2}\right)\right]^{1/2} \mathrm{d}\gamma$$

yields an information function $E\left(-\frac{\mathrm{d}^2\ell(\beta)}{\mathrm{d}\beta^2}\right)$ that is constant in value. Since Fisher's expected information might be difficult to obtain, we can approximate it by the observed information evaluated at the maximum likelihood estimate $\hat{\theta}$ which is

$$j_{\theta\theta}\left(\hat{\theta}\right) = -\left.\frac{\mathrm{d}^2\ell(\theta)}{\mathrm{d}\theta^2}\right|_{\hat{\theta}}.$$

Hence, $\beta\left(\hat{\theta}\right)$ is approximately invariant under reparameterization.

Fraser *et al*. [12] showed that

$$z = \int^{\hat{\varphi}} j_{\varphi\varphi}^{1/2}(\gamma)\mathrm{d}\gamma - \int^{\varphi} j_{\varphi\varphi}^{1/2}(\gamma)\mathrm{d}\gamma \tag{11}$$

is a pivotal quantity to the second-order. A change of variable from the maximum likelihood estimate of locally defined canonical parameter $\hat{\varphi}$ to the score variable s for the first integral of (11) yields

$$z = \int^{s} j_{\varphi\varphi}^{-1/2}\left(\hat{\varphi}(\gamma)\right)\mathrm{d}\gamma - \int^{\varphi} j_{\varphi\varphi}^{1/2}(\gamma)\mathrm{d}\gamma \tag{12}$$

which relates the score varaible to the locally defined canonical parameter. Taking the total derivative of (12), and evaluate at the observed data point, we have

$$\left.\frac{\mathrm{d}s}{\mathrm{d}\theta}\right|_{\left(s^0,\hat{\theta}^0,\hat{\varphi}^0\right)} = j_{\varphi\varphi}^{1/2}\left(\hat{\varphi}^0\right) j_{\varphi\varphi}^{1/2}\left(\varphi\left(\hat{\theta}^0\right)\right).$$

Moreover, at $\hat{\theta}^0$,

$$\mathrm{d}\varphi = \varphi_\theta\left(\hat{\theta}^0\right)\mathrm{d}\theta.$$

Therefore, the rate of change of the score variable with respect to the change of the locally defined canonical parameter at the observed data point is

$$w = \left.\frac{\mathrm{d}s}{\mathrm{d}\varphi}\right|_{\left(s^0,\hat{\theta}^0,\hat{\varphi}^0\right)} = j_{\varphi\varphi}^{1/2}\left(\hat{\varphi}^0\right) j_{\varphi\varphi}^{1/2}\left(\varphi\left(\hat{\theta}^0\right)\right)\varphi_\theta\left(\hat{\theta}^0\right). \tag{13}$$

This describes how the locally defined canonical parameter φ moves the score variable s.

4. Combining Likelihood Information

Assume we have k independent investigations, each of them is used to obtain inference concerning a scalar parameter θ. Denote the log-likelihood function for the i^{th} investigation be $\ell_i(\theta)$ and the corresponding canonical parameter is $\varphi_i(\theta)$. Note that if $\varphi_i(\theta)$ is not explicitly available, we can use the locally defined canonical variable as obtain from (9). The combined log-likelihood function is

$$\ell(\theta) = \ell_1(\theta) + \cdots + \ell_k(\theta)$$

and hence the maximum likelihood estimate of θ can be obtained. Therefore, the signed log-likelihood function r can be calculated from (12).

From (13), the rate of change of the score variable from the i^{th} investigation with respect to the corresponding canonical paramter at the observed data from the i^{th} investigation is

$$w_i = j^{1/2}_{i,\varphi_i\varphi_i}\left(\hat{\varphi}_i^0\right) j^{1/2}_{i,\varphi_i\varphi_i}\left(\varphi_i\left(\hat{\theta}^0\right)\right)\varphi_{i,\theta}\left(\hat{\theta}^0\right) \tag{14}$$

where

$$\varphi_{i,\theta}\left(\hat{\theta}^0\right) = \left.\frac{\mathrm{d}\varphi_i(\theta)}{\mathrm{d}\theta}\right|_{\hat{\theta}^0} \quad \text{and} \quad j_{i,\varphi_i\varphi_i}\left(\hat{\varphi}_i^0\right) = -\left.\frac{\mathrm{d}^2\ell_i(\theta)}{\mathrm{d}\varphi_i^2}\right|_{\hat{\varphi}_i^0}.$$

Hence, the combined canonical parameter is

$$\varphi = \varphi(\theta) = w_1\varphi_1(\theta) + \cdots + w_k\varphi_k(\theta). \tag{15}$$

The standardized maximum likelihood departure based on the combined canonical parameter can be calculated from (5). Thus, a new *p*-value can be obtained from the combined log-likelihood function and the combined canonical parameter using the Lugannani and Rice formula or the Barndorff-Nielsen formula.

5. Examples

In this section, we first revisit the rate of arrival problem discussed in Section 1 and show that the proposed method gives results that is quite different from the results obtained by the Fisher's way of combining *p*-values. Then simulation studies are performed to compare the accuracy of the proposed method with the Fisher's method for the rate of arrival problem. Moreover, two well-known models: scalar canonical exponential family model and normal mean model, are examined. It is shown that, theoretically, the proposed method gives the same results as obtained by the third-order method that was discussed in Fraser and Reid [5] and DiCiccio *et al.* [14], respectively.

5.1. Revisit the Rate of Arrival Problem

From the first investigation discussed in Section 1, the log-likelihood function for the Poisson model is

$$\ell_1(\theta) = -20\theta + 14\log(\theta)$$

where $\varphi_1 = \varphi_1(\theta) = \log(\theta)$ is the canonical parameter. We have

$$\hat{\varphi}_1^0 = \log(14/20) = -0.3567, \quad j_{1,\varphi_1\varphi_1}\left(\hat{\varphi}_1^0\right) = 14.$$

Moreover, from the second investigation discussed in Section 1, the log-likelihood function for the exponential model is

$$\ell_2(\theta) = \log(\theta) - 2\theta$$

where $\varphi_2 = \varphi_2(\theta) = \theta$ is the canonical parameter. We have

$$\hat{\varphi}_2^0 = 1/2 = 0.5, \quad j_{2,\varphi_2\varphi_2}\left(\hat{\varphi}_2^0\right) = 4.$$

The combined log-likelihood function is

$$\ell(\theta) = \ell_1(\theta) + \ell_2(\theta) = -22\theta + 15\log(\theta)$$

and we have

$$\hat{\theta}^0 = 15/22 = 0.6818, \quad j_{\theta\theta}\left(\hat{\theta}^0\right) = 32.2667.$$

Therefore,

$$\varphi_1\left(\hat{\theta}^0\right) = -0.3830, \quad \left.\frac{\mathrm{d}\varphi_1(\theta)}{\mathrm{d}\theta}\right|_{\hat{\theta}^0} = 1.4667, \quad j_{1,\varphi_1\varphi_1}\left(\hat{\varphi}_1^0\right) = 13.6364$$

$$\varphi_2\left(\hat{\theta}^0\right) = 0.6818, \quad \left.\frac{\mathrm{d}\varphi_2(\theta)}{\mathrm{d}\theta}\right|_{\hat{\theta}^0} = 1, \quad j_{2,\varphi_2\varphi_2}\left(\hat{\varphi}_1^0\right) = 2.1511$$

and from (17) we have $w_1 = 20.2650$ and $w_2 = 2.9333$. Thus, the combined locally defined canonical parameter is

$$\varphi(\theta) = 20.2650\log(\theta) + 2.9333\theta.$$

Hence, $r = -1.5844$ is obtained from (12) using the combined log-likelihood function. Since the signed log-likelihood ration statistic is asymptotically distributed as a standard normal distribution, the *p*-value obtained from the signed log-likelihood ratio method is 0.0565. It is well-known that the signed log-likelihood ratio method has only first order accuracy. From (8) using the combined locally defined canonical parameter, we have $q = -1.5124$. Finally, the *p*-value obtained by the Lugannani and Rice formula and by the Barndorff-Nielsen formula is 0.0600, which is less certain about the evidence that θ is greater than 1 as suggested by the result from Fisher's way of combining of *p*-values. Note that in literature, there are many detailed studies comparing the accuracy of the first order and third order methods (see Barndorff-Nielsen [4], Fraser [6], Jensen [7], Brazzale *et al*. [8], and DiCiccio *et al*. [14]). Thus, in this paper, we will not compare the signed log-likelihood ratio method and the proposed method.

Figure 1 plot $p(\theta) = P\left(\hat{\theta} \le \theta\right)$ obtained from Fisher's method, Lugannani and Rice method and Barndorff-Nielsen method. From the plot, it is clear that the two proposed methods give almost identical results, which are very different from the results obtained by the Fisher's method.

5.2. Simulation Study

Simulation studies are performed to compare the three methods discussed in this paper. We examine the rate of arrival problem that was discussed in Section 1. For each combination of (n,θ), we

1) generate x^0 from Poisson $(n*\theta)$, and y^0 from exponential (θ);

2) calculate *p*-values obtained by the three methods discussed in this paper;

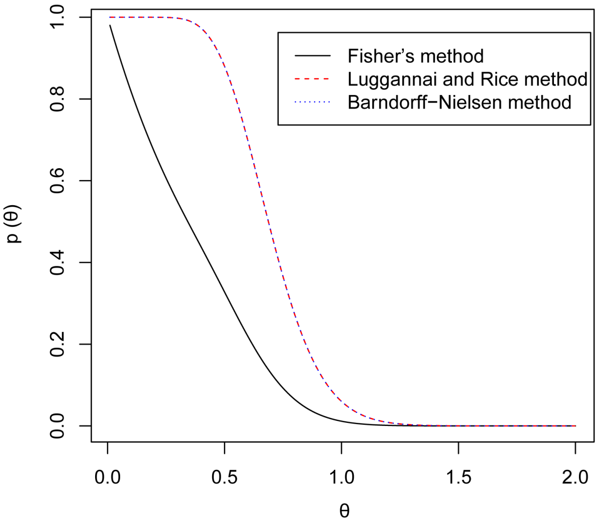

Figure 1. *p*-value function.

3) record if the *p*-value is less than a preset value α;

4) repeat this process $N = 10,000$ times.

Finally, report the proportion of *p*-values that is less than α and this value, sometimes, is referred to as the simulated Type I errors. For an accurate method, the result should be close to α. The simulated standard error of this process is $\sqrt{\alpha(1-\alpha)/N}$.

Table 1 recorded the simulated Type I errors obtained by the Fisher's method (Fisher), Lugannani and Rice method (LR) and Barndorff-Nielsen method (BN). Results from **Table 1** illustrated that the proposed methods are extremely accurate as they are all within 3 simulated standard errors. And the results by the Fisher's method are not satisfactory as they are way larger than the prescriped α values.

5.3. Scalar Canonical Exponential Family Model

Consider k independent investigations from canonical exponential family model with density

$$f_i(x;\theta) = \exp\{\theta t_i - K_i(\theta)\} h_i(x), \quad i = 1, \cdots, k$$

where θ is the scalar canonical parameter of interest and $t_i = t_i(x)$ is the minimal sufficient statistic for the i^{th} model.

From the above model, we have $\varphi_i = \varphi_i(\theta) = \theta$. The log-likelihood function and its corresponding derivatives are

$$\ell_i(\varphi_i) = \varphi_i t_i - K_i(\varphi_i)$$

$$\frac{\mathrm{d}\ell_i(\varphi_i)}{\mathrm{d}\varphi_i} = t_i - K_i^{(1)}(\varphi_i)$$

$$\frac{\mathrm{d}^2\ell_i(\varphi_i)}{\mathrm{d}\varphi_i^2} = -K_i^{(2)}(\varphi_i)$$

where $K_i^{(r)}(\varphi_i) = \dfrac{\mathrm{d}^r K_i(\varphi_i)}{\mathrm{d}\varphi_i^r}$. Hence $\hat{\varphi}_i$ has to satisfy $K_i^{(1)}(\hat{\varphi}_i) = t_i$, and the observed information evaluated at $\hat{\varphi}_i$ is $j_{i,\varphi_i\varphi_i}(\hat{\varphi}_i) = K_i^{(2)}(\hat{\varphi}_i)$. The combined log-likelihood function is

$$\ell(\theta) = \ell_i(\theta) + \cdots + \ell_k(\theta) = \theta \sum_{i=1}^{k} t_i - \sum_{i=1}^{k} K_i(\theta)$$

Table 1. Simulated Type I errors (based on 10,000 simulated sample).

		$\alpha = 0.10$			$\alpha = 0.05$			$\alpha = 0.01$		
n	θ	Fisher	LR	BN	Fisher	LR	BN	Fisher	LR	BN
5	0.1	0.2459	0.1084	0.1073	0.1231	0.0525	0.0521	0.0225	0.0099	0.0097
	1.0	0.3252	0.0992	0.0992	0.1908	0.0496	0.0496	0.0515	0.0123	0.0123
	2.0	0.3256	0.1025	0.1025	0.1961	0.0513	0.0513	0.0547	0.0112	0.0112
10	0.5	0.3318	0.1014	0.1014	0.1942	0.0490	0.0490	0.0513	0.0128	0.0128
	1.0	0.3325	0.1005	0.1005	0.1965	0.0530	0.0530	0.0574	0.0105	0.0105
	2.0	0.3269	0.1006	0.1006	0.1975	0.0513	0.0513	0.0562	0.0107	0.0107
20	1.0	0.3365	0.1000	0.1000	0.2018	0.0526	0.0526	0.0546	0.0098	0.0096
	2.0	0.3387	0.1064	0.1064	0.2027	0.0528	0.0528	0.0578	0.0109	0.0109
	5.0	0.3356	0.1048	0.1048	0.2037	0.0528	0.0528	0.0582	0.0111	0.0111

and the log-likelihood ratio statistic obtained from the combined log-likelihood function can be obtained from (12). Moreover, from (17), we have

$$w_i = j^{1/2}_{i,\varphi_i\varphi_i}\left(\hat{\varphi}_i\right) j^{1/2}_{i,\varphi_i\varphi_i}\left(\varphi_i\left(\hat{\theta}\right)\right)\varphi^{(1)}_{i,\varphi_i}\left(\hat{\theta}\right) = \left[K_i^{(2)}\left(\hat{\varphi}_i\right)K_i^{(2)}\left(\hat{\theta}\right)\right]^{1/2}$$

and hence the combined canonical parameter is

$$\varphi(\theta) = \left\{\sum_{i=1}^{k}\left[K_i^{(2)}\left(\hat{\varphi}_i\right)K_i^{(2)}\left(\hat{\theta}\right)\right]^{1/2}\right\}\theta.$$

The maximum likelihood departure in the combined canonical parameter space is

$$\varphi\left(\hat{\theta}\right) - \varphi(\theta) = \left\{\sum_{i=1}^{k}\left[K_i^{(2)}\left(\hat{\varphi}_i\right)K_i^{(2)}\left(\hat{\theta}\right)\right]^{1/2}\right\}\left(\hat{\theta}-\theta\right)$$

with the observed information evaluated at $\hat{\theta}$ being

$$j_{\varphi\varphi}\left(\hat{\theta}\right) = \left[\sum_{i=1}^{k}K_i^{(2)}\left(\hat{\theta}\right)\right]\left\{\sum_{i=1}^{k}\left[K_i^{(2)}\left(\hat{\varphi}_i\right)K_i^{(2)}\left(\hat{\theta}\right)\right]^{1/2}\right\}^{-2}$$

and thus,

$$q = \left(\hat{\theta}-\theta\right)\left[\sum_{i=1}^{k}K_i^{(2)}\left(\hat{\theta}\right)\right]^{1/2}$$

which is the same as directly applying the third-order method to the canonical exponential family model with θ being the canoncial parameter as discussed in Fraser and Reid [5].

5.4. Normal Mean Model

Consider k independent investigations from normal mean model with density

$$f_i\left(x_i;\theta\right) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(x_i-\theta\right)^2\right\},\quad i=1,\cdots,k$$

where θ is the mean parameter of interest. The pivotal quantity is $z_i\left(\theta,x_i\right) = x_i - \theta$. Hence, $\varphi_i = \varphi_i\left(\theta\right) = \theta - x_i$, and

$$\ell_i\left(\varphi_i\right) = -\frac{1}{2}\varphi_i^2$$

$$\frac{\mathrm{d}\ell_i\left(\varphi_i\right)}{\mathrm{d}\varphi_i} = -\varphi_i$$

$$\frac{\mathrm{d}^2\ell_i\left(\varphi_i\right)}{\mathrm{d}\varphi_i^2} = -1$$

with $\hat{\varphi}_i = 0$ and $j_{i,\varphi_i\varphi_i}\left(\hat{\varphi}_i\right) = 1$. The combined log-likelihood function is

$$\ell(\theta) = -\frac{1}{2}\sum_{i=1}^{k}\left(x_i-\theta\right)^2$$

with $\hat{\theta} = \overline{x}$ and $j_{\theta\theta}\left(\hat{\theta}\right) = k$. From (17), we have $w_i = 1$ and, therefore the combined canonical parameter is

$$\varphi(\theta) = \sum_{i=1}^{k}\left(\theta - x_i\right)$$

and $\varphi_\theta\left(\theta\right) = k$. Finally, from Equation (12), the signed log-likelihood ratio statistic is

$$r = \overline{x} - \theta$$

and the standardized maximum likelihood departure calculated in the locally defined canonical parameter scale can be obtained from Equation (8) and is

$$q = \sqrt{k}\left(\overline{x} - \theta\right).$$

These are exactly the same as those obtained in DiCiccio *et al*. [14].

6. Conclusion

In this paper, a method is proposed to obtain a *p*-value by combining the likelihood functions and the standardized maximum likelihood estimates departure calculated in the canonical parameter space of independent investigations for testing a scalar parameter of interest. It is shown that for the canonical exponential model and the normal mean model, the proposed method gives exactly the same results as using the joint likelihood function. Moreover, for the rate of arrival problem, the proposed method gives very different results from the results obtained by the Fisher's way of combining *p*-values. And simulation studies illustrate that the proposed method is extremely accurate.

Acknowledgements

This research was supported in part by and the National Sciences and Engineering Research Council of Canada.

References

[1] Fisher, R.A. (1925) Statistical Methods for Research Workers. Oliver and Boyd, Edinburg.

[2] Lugannani, R. and Rice, S. (1980) Saddlepoint Approximation for the Distribution of the Sum of Independent Random Variables. *Advances in Applied Probability*, **12**, 475-490. http://dx.doi.org/10.2307/1426607

[3] Barndorff-Nielsen, O.E. (1986) Inference on Full or Partial Parameters Based on the Standardized Log Likelihood Ratio. *Biometrika*, **73**, 307-322.

[4] Barndorff-Nielsen, O.E. (1991) Modified Signed Log-Likelihood Ratio. *Biometrika*, **78**, 557-563. http://dx.doi.org/10.1093/biomet/78.3.557

[5] Fraser, D.A.S. and Reid, N. (1995) Ancillaries and Third Order Significance. *Utilitas Mathematica*, **47**, 33-53.

[6] Fraser, D.A.S. (1990) Tail Probabilities from Observed Likelihoods. *Biometrika*, **77**, 65-76. http://dx.doi.org/10.1093/biomet/77.1.65

[7] Jensen, J.L. (1992) The Modified Signed Log Likelihood Statistic and Saddlepoint Approximations. *Biometrika*, **79**, 693-704. http://dx.doi.org/10.1093/biomet/79.4.693

[8] Brazzale, A.R., Davison, A.C. and Reid, N. (2007) Applied Asymptotics: Case Studies in Small-Sample Statistics. Cambridge University Press, New York. http://dx.doi.org/10.1017/CBO9780511611131

[9] Fraser, D.A.S. and Reid, N. (2001) Ancillary Information for Statistical Snference, Empirical Bayes and Likelihood Inference. Springer-Verlag, New York, 185-209. http://dx.doi.org/10.1007/978-1-4613-0141-7_12

[10] Reid, N. and Fraser, D.A.S. (2010) Mean Likelihood and Higher Order Inference. *Biometrika*, **97**, 159-170. http://dx.doi.org/10.1093/biomet/asq001

[11] Davison, A.C., Fraser, D.A.S. and Reid, N. (2006) Improved Likelihood Inference for Discrete Data. *Journal of the Royal Statistical Society Series B*, **68**, 495-508. http://dx.doi.org/10.1111/j.1467-9868.2006.00548.x

[12] Fraser, A.M., Fraser, D.A.S. and Fraser, M.J. (2010) Parameter Curvature Revisited and the Bayesian Frequentist Divergence. *Journal of Statistical Research*, **44**, 335-346.

[13] Jeffreys, H. (1946) An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London Series A*: *Mathematical and Physical Sciences*, **186**, 453-461. http://dx.doi.org/10.1098/rspa.1946.0056

[14] DiCiccio, T., Field, C. and Fraser, D.A.S. (1989) Approximations of Marginal Tail Probabilities and Inference for Scalar Parameters. *Biometrika*, **77**, 77-95. http://dx.doi.org/10.1093/biomet/77.1.77

Open Journal of Statistics, 2015, 5, 60-65
Published Online February 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.51008

Statistical Models for Forecasting Tourists' Arrival in Kenya

Albert Orwa Akuno[1], Michael Oduor Otieno[2], Charles Wambugu Mwangi[1], Lawrence Areba Bichanga[1]

[1]Department of Mathematics, Egerton University, Egerton, Kenya
[2]Department of Mathematics and Computer Science, Laikipia University, Nyahururu, Kenya
Email: orwaakuno@gmail.com, mkaili91@yahoo.com, charlesmwangi@gmail.com, lawareba@gmail.com

Received 24 January 2015; accepted 13 February 2015; published 16 February 2015

Abstract

In this paper, an attempt has been made to forecast tourists' arrival using statistical time series modeling techniques—Double Exponential Smoothing and the Auto-Regressive Integrated Moving Average (ARIMA). It is common knowledge that forecasting is very important in making future decisions such as ordering replenishment for an inventory system or increasing the capacity of the available staff in order to meet expected future service delivery. The methodology used is given in Section 2 and the results, discussion and conclusion are given in Section 3. When the forecasts from these models were validated, Double Exponential Smoothing model performed better than the ARIMA model.

Keywords

Exponential Smoothing, ARIMA Model, Tourists' Arrival Data

1. Introduction

Tourism is one of Kenya's major foreign exchange earners. This greatly depends on the arrival of various groups of tourists. The forecast of tourists' arrivals is important since it would enable the tourism related industries like airlines, hotels and other stakeholders to adequately prepare for any number of tourists at any future date. In this paper, an attempt has been made to forecast tourists' arrivals using statistical time series modeling techniques—Double Exponential Smoothing and Auto-Regressive Integrated Moving Average (ARIMA). [1] used the same models to forecast milk production in India. [2] used univariate SARIMA models to forecast tourists' demands in India.

Then data on tourists' arrival in Kenya were obtained from the Ministry of East African Affairs, Commerce

How to cite this paper: Akuno, A.O., Otieno, M.O., Mwangi, C.W. and Bichanga, L.A. (2015) Statistical Models for Forecasting Tourists' Arrival in Kenya. *Open Journal of Statistics*, **5**, 60-65. http://dx.doi.org/10.4236/ojs.2015.51008

and Tourism, Department of Tourism. Tourists' arrival for the period 1995 to 2008 was used for model fitting, and data for the remaining periods from 2009 to 2012 were used for model validation. The analysis was carried out using R-language, Excel and Minitab version 16.1.1.

2. Methodology

2.1. Selection of Appropriate Smoothing Techniques

Once the presence of trend is detected in the data, smoothing of the time series data follows. Various smoothing techniques as discussed by [3] include; Simple Exponential Smoothing (SES), Double Exponential Smoothing (DES), Triple Exponential Smoothing (TES) and Adaptive Response Rate Simple Exponential Smoothing (ARRSES) which are briefly described below:

2.1.1. Simple Exponential Smoothing (SES)

For the series $Y_1, Y_2, \cdots, Y_t$, the forecast for the preceding value Y_{t+1}, say F_{t+1}, is based on the weights α and $1-\alpha$ to the recent observation Y_t and forecast F_t respectively, where α is the smoothing constant. The form of the model is

$$F_{t+1} = F_t + \alpha\left(Y_t - F_t\right). \tag{1}$$

The size of α used has a great influence on the forecast. The best value of α corresponding to the minimum mean square error (MSE) is usually used.

2.1.2. Double Exponential Smoothing (Holt's)

The form of the model is

$$\left.\begin{aligned} L_t &= \alpha Y_t + (1-\alpha)\left(L_{t-1} + b_{t-1}\right) \\ b_t &= \beta\left(L_t - L_{t-1}\right) + (1-\beta) b_{t-1} \\ F_{t+m} &= L_t + b_t m \end{aligned}\right\} \tag{2}$$

where L_t in the model is the level of the series at time t and b_t is the slope (Trend) of the series at time t, α and β $(= 0.1, 0.2, \cdots, 0.9)$ are the smoothing coefficient for level and smoothing coefficient for trend respectively. In order to fit the model, it is necessary to calculate the initial values of the level L_0 and the trend b_0. [4] suggests that the initial values can be obtained as $L_0 = Y_1$ and $b_0 = 0$, or $Y_2 - Y_1$ or $\left(Y_n - Y_1\right)/(n-1)$. In this paper, the initial values have been obtained as $L_0 = Y_1$ and $b_0 = \left(Y_n - Y_1\right)/(n-1)$.

The pair of α and β that gives a minimum Mean Square Error is preferred.

2.1.3. Triple Exponential Smoothing (Winter's)

When time series data exhibit seasonality, Triple Exponential Smoothing method is the most recommendable. It incorporates three smoothing equations; first for the level, second for trend and third for seasonality.

2.2. Auto-Regressive Integrated Moving Average (ARIMA Model)

2.2.1. Model Identification

According to Box and Jenkins two graphical procedures are used to access the correlation between the observations within a single time series data. According to [5], these devices are called an estimated autocorrelation functions and the estimated partial autocorrelation function. These two procedures measure statistical relationships within the time series data. Summarization of statistical correlation within the time series data is the other step in the identification. Box and Jenkins suggest a whole family of ARIMA models from which we may choose.

In choosing the model that seems appropriate we use the estimated ACF and PACF. This is due to the basic idea that every ARIMA model will have unique ACF and PACF associated with it. Thus we select the model whose theoretical ACF and PACF resembles the anticipated ACF and PACF of the time series data [6].

2.2.2. Estimation

An estimate of the coefficients of the model is obtained by modified least squares method or the maximum likelihood estimation method suitable to the time series data.

2.2.3. Diagnostic Checking

Diagnostic checks help to determine if the anticipated model is adequate. At this stage, an examination of the residuals from the fitted model is done and if it fails the diagnostic tests, it is rejected and we repeat the cycle until an appropriate model is achieved.

The ARIMA model is obtained by taking W_t as the first differenced time series, *i.e.* $d = 1$

$$(W_t - \mu) - \alpha_1 (W_{t-1} - \mu) - \cdots - \alpha_p (W_{t-p} - \mu) + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}. \tag{3}$$

Equation (3) is referred to as the ARIMA $(p,1,q)$.

Different combinations of AR and MA individually yield different ARIMA models [7]. The optimal model is obtained on the basis of minimum value of Akaike Information Criteria (AIC) given by

$$\text{AIC} = -2\log L + 2m \tag{4}$$

where $m = p + q$ and L is the likelihood function. The Root Mean Square Error (RMSE) and the Mean Absolute Percentage Error (MAPE) are used to evaluate the performance of the various models and are given below.

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - F_t}{Y_t}\right| \times 100 \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - F_t)^2} \tag{6}$$

where Y_t is the tourists' arrival in different years and F_t is the forecasted tourists' arrivals in the corresponding years and n is the number of years used as forecasting period.

3. Results and Discussions

3.1. Exponential Smoothing Model

Table 1 shows the yearly tourists' arrival in Kenya (in thousands) for the period 1995-2012. The time plot (**Figure 1**) revealed that there was increasing trend from the year 2002 to 2007. However, there was a sharp drop

Table 1. Data on tourists' arrival in Kenya for the period 1995 to 2012.

Sl. No.	Year	Observed tourists' arrival ('000)
1	1995	973.6
2	1996	1003.0
3	1997	1000.6
4	1998	894.3
5	1999	969.3
6	2000	1036.5
7	2001	993.6
8	2002	1001.5
9	2003	1146.2
10	2004	1360.7
11	2005	1479.0
12	2006	1600.7
13	2007	1816.8
14	2008	1203.2
15	2009	1490.4
16	2010	1609.1
17	2011	1822.9
18	2012	1873.8

Source: Ministry of East African affairs, Commerce and Tourism: Department of Tourism (Kenya).

in the number of tourists in the year 2008 followed by an increasing trend from the year 2009 to 2012. For smoothing the data, Holt's Double Exponential Smoothing was used. Various combinations of α and β both ranging from 0.1 to 0.9 with increments of 0.1 were tried and Mean Squared Error for the forecasts (54.186) and Mean Absolute Percentage Error (3.028) was least for $\alpha = 0.1$ and $\beta = 0.7$. The fitted model is therefore given by;

$$\left.\begin{aligned} L_t &= 0.1Y_t + 0.9\left(L_{t-1} + b_{t-1}\right) \\ b_t &= 0.7\left(L_t - L_{t-1}\right) + 0.3b_{t-1} \\ F_{t+m} &= L_t + b_t m \end{aligned}\right\} \tag{7}$$

where $m = 1,2,3$ and 4 the initial values for the level L_t and trend b_t are 973.6 and 17.66 respectively. **Table 2** shows the forecast of tourists' arrivals using the chosen double exponential smoothing model.

3.2. ARIMA Model

Figure 1 showed that the series was non-stationary since there was some trend component present. The data was made stationary by taking the first order difference $(d = 1)$. The time plot of the differenced data is shown in **Figure 2**.

Using R-language for different values of p and q, various ARIMA models were fitted and the best model was chosen on the basis of minimum value of the selection criteria, that is, Akaike Information Criteria (AIC) whose formula is given in Equation (4). In this way, ARIMA (1, 1, 1) was found to be the best model. The fitted model is given by

$$Y_t = 43.6319 - 0.9999\varepsilon_t + 0.4115Y_{t-1}. \tag{8}$$

The estimation of the model parameters was done by maximum likelihood estimation technique. The fitted model was then used to forecast tourists' arrival from 2009 to 2012. The forecast values are shown in **Table 3**.

Table 2. Forecast of tourists' arrival in Kenya using double exponential smoothing.

S. No.	Year	Observed tourists' arrival ('000)	Forecast of tourists' arrival
			Double exponential model
1	2009	1490.4	1560.936
2	2010	1609.1	1660.595
3	2011	1822.9	1760.254
4	2012	1873.8	1859.912

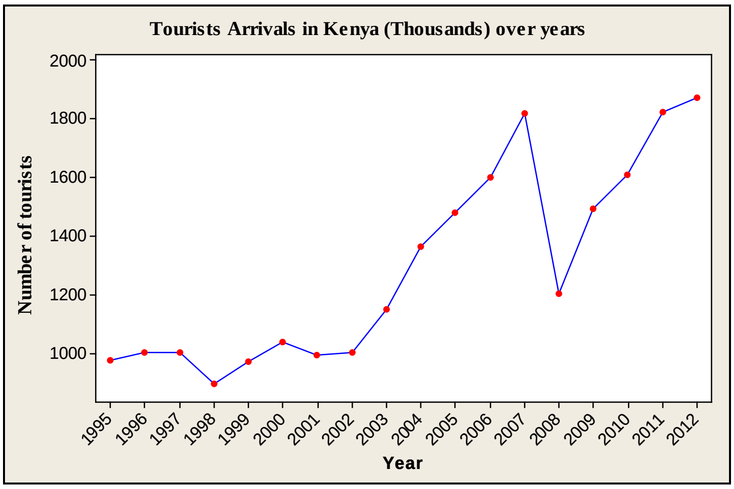

Figure 1. Time plot of tourists' arrival in Kenya between 1995 and 2009.

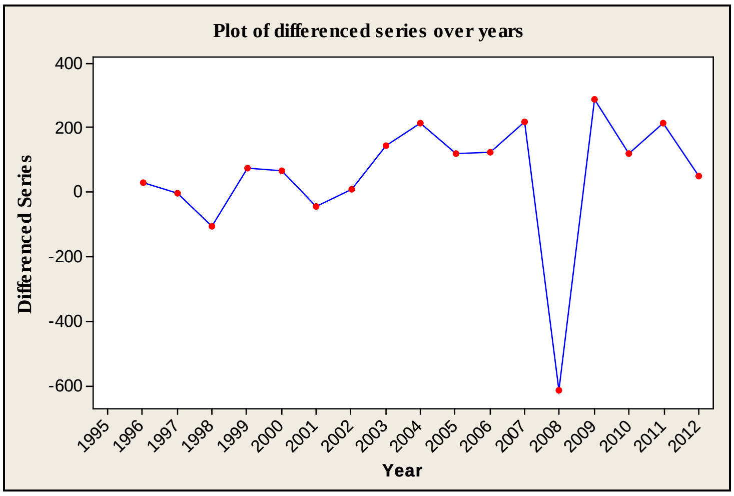

Figure 2. Plot of differenced tourists' arrival data.

Table 3. Forecast of tourists' arrival in Kenya ARIMA (1, 1, 1) models.

Sl. No.	Year	Observed tourists' arrival ('000)	Forecast of tourists' arrival
			ARIMA (1, 1, 1) model
1	2009	1600.7	1393.607
2	2010	1816.8	1497.643
3	2011	1203.2	1566.134
4	2012	1490.4	1619.997

Table 4. Forecast of tourists' arrival in Kenya using double exponential smoothing and ARIMA (1, 1, 1) models.

Sl. No.	Year	Observed tourists' arrival ('000)	Forecast of tourists' arrival	
			Double exponential model	ARIMA (1, 1, 1) model
1	2009	1490.4	1560.936	1393.607
2	2010	1609.1	1660.595	1497.643
3	2011	1822.9	1760.254	1566.134
4	2012	1873.8	1859.912	1619.997
MAPE			3.028	10.263
RMSE			54.186	195.023

3.3. Comparison and Conclusion of the Performance of the Two Models

Performance evaluation measures MAPE and the RMSE were obtained for the forecasted tourists' arrivals for the years 2009 to 2012.

The comparison of the two models based on MAPE and RMSE is as given in **Table 4**. Based on the results from the table, Double Exponential Smoothing model was the best to forecast tourists' arrival in Kenya as both its MAPE and RMSE values were least compared to those of ARIMA (1, 1, 1).

References

[1] Satya, P., Ramasubramanian, V. and Menta, S.C. (2007) Statistical Models for Forecasting Milk Production in India. *Journal of the Indian Society of Agricultural Statistics*, **61**, 80-83.

[2] Padhan, P.C. (2011) Forecasting International Tourists Footfalls in India: An Assortment of Competing Models. *International Journal of Business and Management*, **6**, 190-202.

[3] Gardener, E.S. (1985) Exponential Smoothing—The State of the Art. *Journal of Forecasting*, **4**, 1-28. http://dx.doi.org/10.1002/for.3980040103

[4] Jani, P.N. (2014) Business Statistics: Theories and Applications. PHI Learning Private Limited, Delhi.

[5] Box, G.E.P. and Jenkins, G.M. (1970) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

[6] Pankratz, A. (1983) Forecasting with Univariate Box-Jenkins Models: Concepts and Cases. John Wiley and Sons, New York. http://dx.doi.org/10.1002/9780470316566

[7] Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1998) Forecasting: Methods and Applications. John Wiley & Sons, New York.

Open Journal of Statistics, 2015, 5, 66-74
Published Online February 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.51009

Forecasting High-Frequency Long Memory Series with Long Periods Using the SARFIMA Model

Handong Li, Xunyu Ye

School of Government, Beijing Normal University, Beijing, China
Email: lhd@bnu.edu.cn, xunyuye@bnu.edu.cn

Received 25 January 2015; accepted 13 February 2015; published 17 February 2015

Abstract

This paper evaluates the efficiency of the SARFIMA model at forecasting high-frequency long memory series with especially long periods. Three other models, the ARFIMA, ARMA and PAR models, are also included to compare their forecasting performances with that of the SARFIMA model. For the artificial SARFIMA series, if the correct parameters are used for estimating and forecasting, the model performs as well as the other three models. However, if the parameters obtained by the WHI estimation are used, the performance of the SARFIMA model falls far behind that of the other models. For the empirical intraday volume series, the SARFIMA model produces the worst performance of all of the models, and the ARFIMA model performs best. The ARMA and PAR models perform very well both for the artificial series and for the intraday volume series. This result indicates that short memory models are competent in forecasting periodic long memory series.

Keywords

High-Frequency Financial Series, Long Memory, Long Periods, SARFIMA, Monte Carlo Simulation, Intraday Volume

1. Introduction

Recent years have witnessed a vast increase in the amount of high-frequency financial market data that are available. Using these data, practitioners are now able to manage their assets in greater detail. For example, the intraday volume series is often used in the Volume Weighted Average Price (VWAP) strategy to avoid a large reverse impact when executing large orders. Consequently, the econometrics of the high-frequency financial se-

How to cite this paper: Li, H.D. and Ye, X.Y. (2015) Forecasting High-Frequency Long Memory Series with Long Periods Using the SARFIMA Model. *Open Journal of Statistics*, **5**, 66-74. http://dx.doi.org/10.4236/ojs.2015.51009

ries receives wider attention in the academic field. As Engle (2005) summarizes, intraday financial series often contain periodic patterns and present a long horizon of strong dependence [1]. The autocorrelation function (ACF) of these series decays slowly and is particularly significant at the seasonal lags. These periods can be especially long when the sampling interval becomes short.

A number of works have been concerned with forecasting the periodic long memory series. They mainly focused on forecasting series with relatively short periods, namely twelve monthly periods or four seasonal periods in a year. On one hand, various long memory models have been used for forecasting this series. An autoregressive fractionally integrated moving average (ARFIMA) model [2] [3] was directly applied by Franses & Ooms (1997) in [4] to forecast quarterly UK inflation. Porter-Hudak (1990) suggested a seasonal autoregressive fractionally integrated moving average (SARFIMA) model to forecast monetary aggregates [5]. This model tries to remove the hyperbolic decay at the seasonal lags by including a seasonal fractional differencing filter $\left(1-B^s\right)^D$ in the ARFIMA model, where B is the backward shift operator, s is the given period and D is the seasonal differencing parameters. This model is later used in [6] for monthly river flows and in [7] for inflation rates. By introducing seasonal dummy variables to seasonally change the fractional differencing parameter in the ARFIMA model, Ref. [4] proposed a periodic ARFIMA (PARFIMA) model for forecasting periodic long memory series. On the other hand, short memory models, such as the autoregression (AR) model and the periodic autoregression (PAR) model, were also proven to be competent in handling this series.

However, no consistent conclusion has been made on the superiority of specific models for forecasting periodic long memory series. Franses & Ooms (1997) [4] tried the periodic PAR model, AR model, PARFIMA model and ARFIMA model to forecast the quarterly UK inflation, but found no significant difference between these models. Those authors did find that the PARFIMA model was generally outperformed by rival models. Porter-Hudak (1990) compared the SARFIMA model and the Airline model, and found that the former outperformed the latter [5]. Nasr & Trabelsi (2005) tried the PARFIMA, SARFIMA, PAR, and AR models in [7] to forecast inflation rates in four different countries, and showed that the long memory models, the PARFIMA model and the SARFIMA model, performed better than the short memory models in terms of information criteria and clean residuals.

This paper studies the performance of different models when forecasting high-frequency long memory series with long periods. In particular, we want to deduce whether the SARFIMA model is capable of forecasting this type of series, because the mechanism of the SARFIMA process fits the description of the periodic long memory series well. Artificial SARFIMA series are generated to test the performance of different forecasting models, including the ARFIMA, the SARFIMA, the AR and the PAR models. We are also interested in finding a suitable model for forecasting intraday volume series, which is a very useful series for VWAP trading. These four models will also be tried on this series for comparison.

This paper is organized as follows: Section 2 introduces the four forecasting models that will be tested. Section 3 studies the performances of the four models through Monte Carlo simulations. Section 4 uses these models to forecast the intraday volume in both the American and Chinese stock markets and then compares their performance. Section 5 presents our conclusion.

2. The Models

2.1. Long Memory Models

Two long memory models will be used in our study. The ARFIMA model ignores periodicity. The other, the SAFARMA model, includes periodicity.

If we assume that a simple fractional differencing operator can remove the high autocorrelation at both the seasonal lags and non-seasonal lags, we can use an ARFIMA model directly to help forecast a periodic long memory series. The ARFIMA model is defined as:

$$\phi(B)(1-B)^d\left(X_t-\mu\right)=\theta(B)\varepsilon_t$$

where μ is the mean of the process, ε_t is the white noise process, B is the backward shift operator, $0 < d < 0.5$ is the differencing parameters respectively, $\phi(B)$, $\theta(B)$ are the polynomial operators with orders p, q respectively.

$$\phi(B)=1-\phi_1B-\phi_2B^2-\cdots-\phi_pB^p,\quad \theta(B)=1-\theta_1B-\theta_2B^2-\cdots-\theta_qB^q$$

A full definition of the SARFIMA $(p,d,q)\times(P,D,Q)_s$ model is defined as:

$$\phi(B)\Phi(B)(1-B)^d\left(1-B^s\right)^D(X_t-\mu)=\theta(B)\Theta(B)\varepsilon_t$$

where $s\in\mathrm{N}$ is the seasonal period, $0<d<0.5$ and $0<D<0.5$ are the non-seasonal and seasonal differencing parameters respectively, with additional constraints $0<d+D<0.5$ to assure the stationary of the process, $\phi(B)$, $\theta(B)$ are the non-seasonal polynomial operators with orders *p*, *q* respectively, $\Phi(B)$, $\Theta(B)$ are the seasonal polynomial operators with orders *P*, *Q* respectively:

$$\Phi(B)=1-\Phi_1B-\Phi_2B^2-\cdots-\Phi_PB^P,\quad \Theta(B)=1-\Theta_1B-\Theta_2B^2-\cdots-\Theta_QB^Q.$$

For convenience, this paper is restricted to the SARFIMA model with $P=Q=0$, namely:

$$\phi(B)(1-B)^d\left(1-B^s\right)^D(X_t-\mu)=\theta(B)\varepsilon_t$$

and the ARFIMA model with $p=q=1$.

For the ARFIMA model, there are several methods chosen for its parameter estimation, including the Exact Maximum Likelihood method [8], WHI method [9] and Non-Linear Least Squares estimator [10]. For the SARFIMA model, Reisen *et al*. (2006) suggest a maximum likelihood method for its estimation [11]. However, this method is time-consuming when calculating the covariance matrix, especially for the high-frequency series with a relatively large sample size and when the AR coefficients, MA coefficients, seasonal and non-seasonal fractionally differencing parameters are all nonzero. Moreover, no further improvement for simplifying the procedure of this method, such as what Sowell (1991) does to improve the maximum likelihood estimation for the ARFIMA model, has yet been proposed. Bisognin & Lopes (2007) use the WHI method for the SARFIMA model's estimation in [12]. This method is simpler and faster in application. Because this paper discusses the forecasting of large sample high-frequency data using the SARFIMA model with nonzero AR and MA coefficients and non-seasonal fractionally differencing parameters, we use the WHI method for estimating the SARFIMA model.

For consistency, this paper also uses the WHI method to estimate the parameters of the ARFIMA. WHI is an approximated likelihood method. The discrete form of its likelihood function is given by:

$$L(\varsigma)=(2n)^{-1}\sum_{j=1}^{n-1}\left\{\ln f_X\left(\lambda_j,d\right)+\frac{I\left(\lambda_j\right)}{f_X\left(\lambda_j,\varsigma\right)}\right\}$$

where ς is the vector of unknown parameters $\left(d,D,\mu,\sigma_\varepsilon^2\right)'$, λ_j is the frequency, *n* is the sample size, $f\left(\lambda_j,\theta\right)$ is the spectral density function of X_t and

$$I\left(\lambda_j\right)=(2\pi n)^{-1}\left|\sum_{t=1}^{n}X_t\mathrm{e}^{it\lambda_j}\right|^2.$$

2.2. Short Memory Models

The two short memory models used in this paper are the ARMA model and the PAR model.

The ARMA model is formulated as:

$$\phi(B)(X_t-\mu)=\theta(B)\varepsilon_t.$$

By incorporating seasonal polynomial operators to the AR model, the PAR(*p*) model is defined as:

$$\phi(B)\Phi^s(B)(X_t-\mu)=\varepsilon_t$$

where *s* is the given period and $\Phi^s(B)$ is the seasonal polynomial operators with orders *P*:

$$\Phi^s(B)=1-\Phi_1B^s-\Phi_2B^{2s}-\cdots-\Phi_PB^{Ps}.$$

This paper is restricted with $p = q = P = 1$, namely an ARMA (1,1) and a PAR(1). The parameters of the AR and PAR models are estimated by non-linear least squares method.

3. Simulation Study

To test the performance of different models for forecasting a periodic long memory series, we generate the SARFIMA $(p=1,d,q=1)\times(P=0,D,Q=0)_s$ artificial series X_t with zero mean and unit variance:

$$\phi(B)(1-B)^d\left(1-B^s\right)^D X_t = \theta(B)\varepsilon_t$$

$n \in \{300,1100\}$ when $s = 48$, $n \in \{500,2000\}$ when $s = 78$, autoregressive and moving average parameters $(\phi_1,\theta_1) = (0.3,0.4)$ and fractional differencing parameters $(d,D) = (0.2,0.2)$. Periods $s \in \{48,78\}$ are used to correspond to the periods of the intraday volume series examined in the next section. Consequently, there are four types of this artificial series. The last two periods of each series are left for forecasting; the former data are used for estimation. Forecasts are undertaken one-step in advance. For example, for {*s* = 48, *n* = 300} series, the 1st to 204th real data are used for estimation so that we can obtain the 205th predictive data. Then, the 2nd to 205 real data are used for estimation, and we forecast the 206th predictive data. Under each sample size, 100 duplicated series are generated to investigate the overall forecasting performance of different models. **Figure 1** plots the examples of the last two periods of the four types of the artificial series. **Figure 2** shows the ACF of these four series.

The periodicity does not seem to be apparent for all of these series, but the ACF shows significant autocorrelations both at the seasonal and non-seasonal lags for the two series. The Augmented Dickey-Fuller unit root test and two semiparametric tests, GPH test and the Gaussian Semi-parametric (GSP) test [13] [14] are also undertaken to examine their stationary status and long memory. **Table 1** lists the ADF unit root test and the long

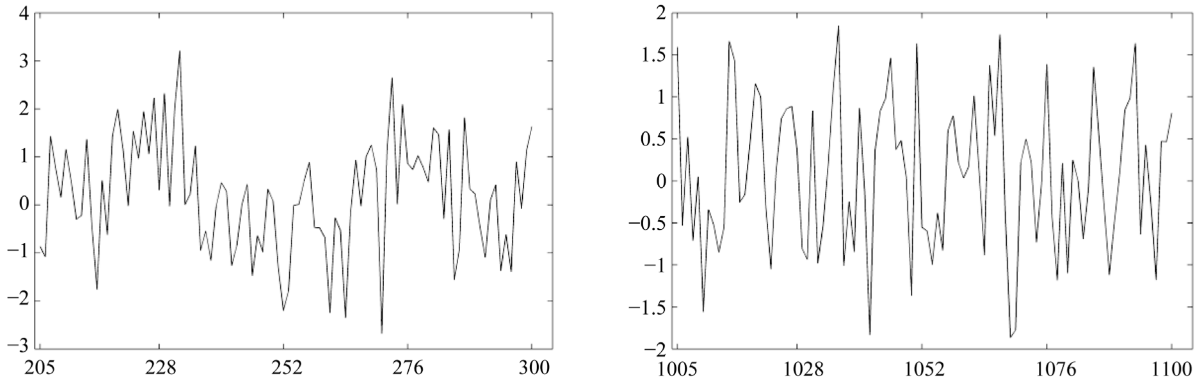

Figure 1. Data plots for the artificial series, the 205th to 300th data for *n* = 300 series (left), the 1005th to 1100th data for *n* = 1100 series (right).

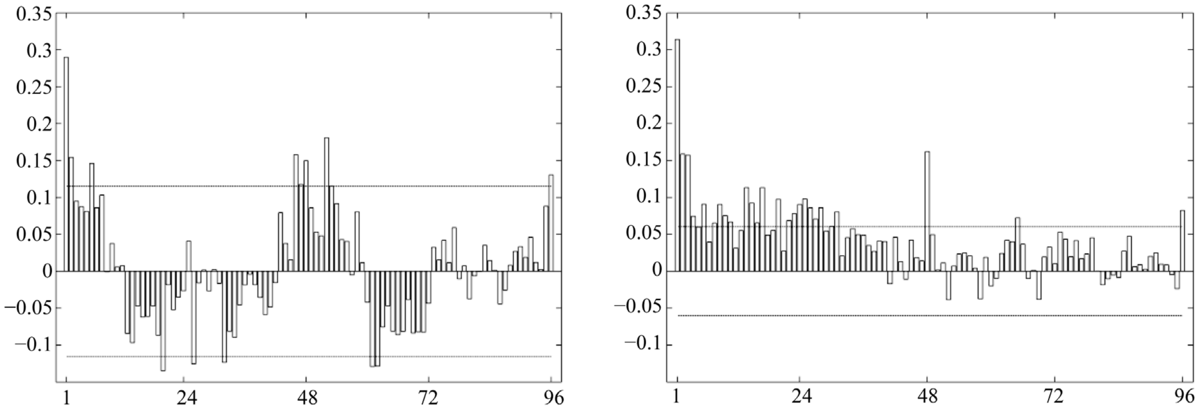

Figure 2. ACF for the artificial series, *n* = 300 series (left), *n* = 1100 series (right).

Table 1. ADF and long memory test results.

Test	$s = 48, n = 300$	$s = 48, n = 1100$	$s = 78, n = 500$	$s = 78, n = 2000$
ADF (*t*-statistics)	−12.7408*	−14.3546*	−15.3429*	−19.6710*
GPH (*d* value)	0.2264	0.2126	0.2435	0.2577
GSP (*d* value)	0.1798	0.2191	0.2657	0.2471

*Significant at 5% level.

memory test results. The ADF and long memory tests show that the two series are both stationary and with significant long memory.

For other duplicated series, their plots, ACF, stationarity and long memory properties are similar with the two series examined above. Due to space constraints, we provide only two examples here and do not elaborate on them.

The four models are then used to forecast the artificial series for testing their forecasting performances. Two sets of parameters are used for forecasting, especially for the SARFIMA model. One parameter is obtained by the WHI estimation. The other, as we already know the true parameters of the artificial SARFIMA series, is the set of parameters of the artificial series. Accordingly, we can take the performance of these two different parameter settings for the WHI estimation together to determine whether estimation bias would cause any negative effect. The statistical measure used in this paper for measuring forecasting accuracy is the mean squared error of the estimators (MSE), given by:

$$\text{MSE} = \sum_{t=1}^{k}\left(\hat{X}_t - X_t\right)^2$$

where k is the number of predicted data and X_t and $\hat{X}_t$ are the real and predictive value of the series respectively. For each type within the artificial series, we calculate the average MSE of the 100 duplicated series by:

$$\overline{\text{MSE}} = \sum_{i=1}^{n=100} \text{MSE} \Big/ 100.$$

Table 2 lists the average MSE of the four models for forecasting the four types of the artificial series. The SARFIMA model with known parameters is denoted as SARFIMA-known. The SARFIMA model with parameters estimated by WHI is denoted as SARFIMA-WHI. The averages of the MSE of these models for the 100 duplicated series are listed in the last row.

First, we can see from **Table 2** that the PAR model performs best at forecasting all types of the artificial series. Its MSE are the smallest for most duplicated series. This finding indicates that a periodic short memory model is competent at predicting an SARFIMA series. The SARFIMA model using known parameters also performs well, with its average MSE ranked second. The non-periodic models, namely the ARFIMA and ARMA models, perform slightly worse than the periodic models. This finding indicates that considering periodicity is beneficial for accurately forecasting the artificial SARFIMA series, but the differences between the performances of the PAR, ARMA, SARFIMA-known and ARFIMA models are not very large. The differences of their average MSE are within 0.06.

However, the performance of SARFIMA-WHI falls significantly behind that of other models. Most of its MSE are much larger than that of others. **Table 3** lists the average of the estimated parameters for the SARFIMA-WHI model.

We can see that the WHI method tends to underestimate both the seasonal and non-seasonal fractional differencing parameters d and D. This phenomenon is true especially for the $n = 300$ artificial series, for which the WHI method underestimate d and D by nearly 0.1. This finding indicates that the estimation bias is responsible for the loss of forecasting accuracy of the SARFIMA model using the WHI estimation.

4. Empirical Study

The intraday volume series is a very useful series for VWAP trading strategy, which splits and executes orders according to the predicted intraday volume distribution. Intraday volume series is a typical series that presents both periodicity and long memory. Here, we choose data gathered from the NASDAQ composite index and the

Table 2. The average MSEs for forecasting four types of the artificial series.

Type of the artificial series	ARFIMA	SARFIMA-known	SARFIMA-WHI	ARMA	PAR
$s = 48$, $n = 300$	1.0171	1.0123	1.2862	1.0404	0.9906
$s = 48$, $n = 1100$	1.0472	1.0223	1.3629	1.0630	1.0105
$s = 78$, $n = 500$	1.0370	1.0219	1.3460	1.0635	1.0117
$S = 78$, $n = 2000$	1.0585	1.0275	1.4514	1.0725	1.0144

Table 3. The average of the estimated parameters for the SARFIMA-WHI model.

Type of the artificial series	d	D	AR	MA
$s = 48$, $n = 300$	0.1354	0.0959	0.1078	0.1061
$s = 48$, $n = 1100$	0.1564	0.1626	−0.0304	0.1797
$s = 78$, $n = 500$	0.1822	0.1380	0.0292	0.1520
$s = 78$, $n = 2000$	0.1866	0.1823	0.1652	0.2838

Shanghai Stock Exchange 50 Index (SSE 50 Index)[1] to populate the sample for this description. We use two one-month samples. The SSE 50 Index ranges from January 4th to 31st 2011 in 5-minute intervals. The NASDAQ sample ranges from January 3rd to 31st 2011 in 5-minute intervals. Because the trading time for the Chinese stock market is 4 hours per day and for the American stock market is 6.5 hours per day, the total time of every trading day can be divided into 48 parts and 78 parts, respectively. Therefore, for 20 trading days, we obtained 960 and 1560 observed values from the Chinese market and American market, respectively. For each 5-minute interval, volume means the sum of all volumes traded within 5 minutes. **Figure 3** shows the plots and the ACF of the sample series.

The periodicity and slow decay of the intraday volume series seem to be much more apparent than the artificial SARFIMA series. The plots show that the sample of the intraday volume series of the SSE 50 Index and NASDAQ composite index fluctuate in a U-shape and presents an apparent 48 and 78 periods, respectively. The ACF of the series show a very slow decay in the autocorrelation function both at the seasonal and non-seasonal lags for the series.

Next, we apply the four models to a one-year sample of the SSE 50 Index and NASDAQ composite index intraday volume to investigate their forecasting performance. This is an in-sample forecast comparison. The parameters of the models are estimated every month. The forecast is undertaken one-step ahead, using the monthly fixed parameters and historical rolling five-day data to forecast the next data. **Table 4** and **Table 5** list the statistics of the mean, maximum value, minimum value, ADF *t*-statistics and fractional differencing parameters *d* estimated by GPH and GSP of the SSE 50 Index and NASDAQ composite index intraday volume for each month in 2011, respectively.

On average, the maximum value of each month is more than 4 times the mean value and more than 28 times the minimum value. This finding indicates a large deviation, partly due to the seasonal pattern. Although the ADF tests prove these series all to be stationary, most of the two semi-parametric estimators of the fractional differencing parameter, GPH and GSP, are near or above 0.5. These rates indicate that these stationary series have very strong long memories.

Applying the four models to the sample intraday volume series, we obtain the statistics of MSE of their forecasting, as listed in **Table 6** and **Table 7**.

The results of the two tables are fairly similar. For most monthly samples, the ARFIMA model performs best, indicating that a fractional differencing is beneficial for forecasting intraday volume series. Meanwhile, the non-periodic models, the ARFIMA and the ARMA, seem to be superior to the periodic models. This finding indicates that, for forecasting intraday volume, adding periodicity may be unnecessary or redundant in terms of forecasting accuracy. The worst performance belongs to the SARFIMA model, of which the MSE are highest for forecasting all monthly samples. We can conclude that although this model is considered to be theoretically suitable for modeling periodic long memory series, it does not actually work very well on our intraday volume

[1] The SSE 50 Index used consists of the 50 largest stocks of good liquidity and representativeness from the Shanghai security market according to an objective, scientific method.

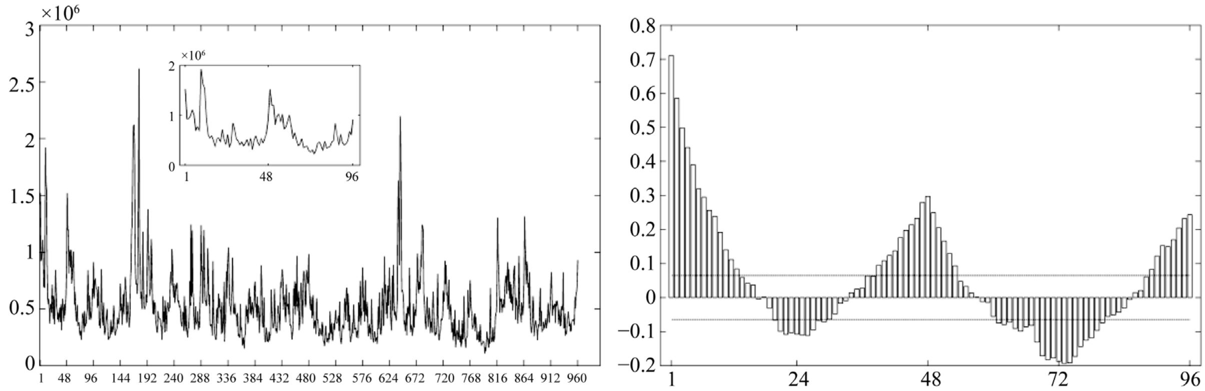

Figure 3. The plots and the ACF of the SSE 50 Index, January 4th to 31st 2011, 5-minute intervals.

Table 4. The statistics of the SSE 50 Index intraday volume for each month in 2011.

Month	Number of observations	Mean	Max	Min	ADF (*t*-statistics)	GPH (*d* value)	GSP (*d* value)
Jan.	960	517213	2617444	109180	−10.4575*	0.5885	0.5332
Feb.	720	578184	2032704	132651	−4.9658*	0.5653	0.5490
Mar.	1104	597276	2921257	134438	−8.7126*	0.5403	0.5735
Apr.	912	604845	3010937	175328	−8.6219*	0.5591	0.5880
May.	1008	315125	1052178	98968	−7.7034*	0.4277	0.4303
Jun.	1008	349801	1515968	85317	−6.6711*	0.4829	0.4470
Jul.	1008	412374	2153114	96716	−10.6532*	0.4949	0.5078
Aug.	1104	361702	2835370	83057	−9.1132*	0.5173	0.5241
Sep.	1008	269615	1563451	63966	−10.6051*	0.4033	0.4006
Oct.	768	386268	2983349	70697	−9.0214*	0.5066	0.4673
Nov.	1056	297883	1356324	66211	−6.5933*	0.5083	0.5431
Dec.	1056	247512	4117696	60367	−10.3629*	0.4056	0.4242
Average	-	-	-	-	-	0.5000	0.4990

*Significant at 5% level.

Table 5. The statistics of the NASDAQ intraday volume for each month in 2011.

Month	Number of observations	Mean	Max	Min	ADF	GPH	GSP
					(*t*-statistics)	(*d* value)	(*d* value)
Jan.	1560	55	195	0	−10.4543*	0.6201	0.6568
Feb.	1482	57	160	1	−8.3468*	0.5625	0.6194
Mar.	1794	67	206	3	−8.3348*	0.5373	0.6121
Apr.	1560	57	186	13	−8.9120*	0.5541	0.6630
May.	1638	62	179	5	−8.9724*	0.5117	0.6138
Jun.	1716	65	190	4	−8.5986*	0.4677	0.5749
Jul.	1560	66	183	1	−8.8776*	0.5265	0.5968
Aug.	1794	105	257	9	−6.2128*	0.6028	0.5980
Sep.	1638	97	225	29	−6.2058*	0.6183	0.6754
Oct.	1638	94	232	2	−5.9833*	0.6016	0.6858
Nov.	1638	78	186	1	−7.5528*	0.6128	0.6561
Dec.	1638	59	172	7	−7.2939*	0.5171	0.6365
Average	-	-	-	-	-	0.5610	0.6324

*Significant at 5% level.

Table 6. The MSE ($\times 10^{10}$) of the four models' forecasting results, SSE 50 Index.

Month	SARFIMA	ARFIMA	PAR	ARMA	Performance
Jan.	4.4344	2.9568	3.4543	3.1701	ARFIMA > ARMA > PAR > SARFIMA
Feb.	5.1260	3.5417	4.1332	3.8100	ARFIMA > ARMA > PAR > SARFIMA
Mar.	4.9637	3.6353	4.1220	3.8267	ARFIMA > ARMA > PAR > SARFIMA
Apr.	4.9465	3.6692	4.2262	3.8605	ARFIMA > ARMA > PAR > SARFIMA
May.	2.2025	1.0294	1.3206	1.0867	ARFIMA > ARMA > PAR > SARFIMA
Jun.	3.2072	1.5498	1.9334	1.6016	ARFIMA > ARMA > PAR > SARFIMA
Jul.	2.4091	1.6758	1.9425	1.8102	ARFIMA > ARMA > PAR > SARFIMA
Aug.	3.5710	1.7880	3.3168	3.6005	ARFIMA > PAR > ARMA > SARFIMA
Sep.	2.9869	1.7880	2.1075	1.9091	ARFIMA > ARMA > PAR > SARFIMA
Oct.	6.3777	3.4519	4.1498	3.6213	ARFIMA > ARMA > PAR > SARFIMA
Nov.	2.5193	1.1536	1.5092	1.1923	ARFIMA > ARMA > PAR > SARFIMA
Dec.	1.7676	0.8571	1.0541	1.0622	ARFIMA > PAR > ARMA > SARFIMA
Average	3.7093	2.2581	2.7725	2.5459	ARFIMA > ARMA > PAR > SARFIMA

In the column "Performance", ">" means to be superior to, e.g. ARFIMA > ARMA means the ARFIMA model is superior to the ARMA model for forecasting the corresponding months' intraday volume.

Table 7. The MSE ($\times 10^{2}$) of forecasting results, NASDAQ composite index.

Month	SARFIMA	ARFIMA	PAR	ARMA	Performance
Jan.	2.9006	2.2891	2.5360	2.3673	ARFIMA > ARMA > PAR > SARFIMA
Feb.	2.8684	2.0828	2.3794	2.1539	ARFIMA > ARMA > PAR > SARFIMA
Mar.	3.4062	2.6044	2.8700	2.6131	ARFIMA > ARMA > PAR > SARFIMA
Apr.	2.3652	2.1415	2.2669	2.1756	ARFIMA > ARMA > PAR > SARFIMA
May.	3.4284	2.4680	2.8457	2.4884	ARFIMA > ARMA > PAR > SARFIMA
Jun.	4.0275	2.6247	3.0821	2.5896	ARMA > ARFIMA > PAR > SARFIMA
Jul.	4.5266	3.4209	3.9391	3.4774	ARFIMA > ARMA > PAR > SARFIMA
Aug.	6.7548	4.5915	5.2309	4.5433	ARMA > ARFIMA > PAR > SARFIMA
Sep.	4.7220	3.5422	3.9096	3.4678	ARMA > ARFIMA > PAR > SARFIMA
Oct.	3.4851	2.6291	2.8474	2.5316	ARMA > ARFIMA > PAR > SARFIMA
Nov.	4.1874	3.0506	3.4239	3.0508	ARFIMA > ARMA > PAR > SARFIMA
Dec.	2.7012	2.0851	2.2770	2.0745	ARMA > ARFIMA > PAR > SARFIMA
Average	3.7811	2.7942	3.1340	2.7944	ARFIMA > ARMA > PAR > SARFIMA

In the column "Performance", ">" means to be superior to, e.g. ARFIMA > ARMA means the ARFIMA model is superior to the ARMA model for forecasting the corresponding months' intraday volume.

samples. Additionally, the two short memory models, the ARMA and PAR models, perform slightly worse than the ARFIMA model, but much better than the SARFIMA model does. This finding indicates that the short memory models are also competent in forecasting intraday volume.

5. Conclusions

This paper evaluates the performance of the SARFIMA model at forecasting periodic long memory series, including the artificial SARFIMA series, the SSE 50 Index intraday volume series and NASDAQ composite index volume series. Three other models are also included in our study to compare their forecasting performances with that of the SARFIMA model.

For the artificial SARFIMA series, if we use the correct parameters for estimating and forecasting, it performs well relative to the other three models. However, if we use the parameters obtained by the WHI estimation, the forecasting performance of this model falls considerably behind other models. This phenomenon may be partly due to the estimation bias of the WHI estimation, which tends to underestimate both the seasonal and non-seasonal fractional differencing parameters. The PAR model performs best at forecasting all four artificial series. Meanwhile, the non-periodic models, namely the ARFIMA and ARMA models, do not perform as well as the periodic models. This outcome indicates that considering periodicity is beneficial for forecasting the artificial SARFIMA series.

For the intraday volume series, the ARFIMA model performs the best among all the models, indicating that fractional differencing is beneficial for forecasting the intraday volume series. For most monthly samples, the non-periodic models, the ARFIMA model and the ARMA model, seem to be superior to the periodic models. This outcome indicates that, for forecasting intraday volume, adding periodicity may be unnecessary or redundant in terms of forecasting accuracy. The SARFIMA model does not work very well on our intraday volume samples, exhibiting the worst performance among all the models used. In addition, the two short memory models, the ARMA and PAR models, also performed well compared to the SARFIMA model.

In summary, the SARFIMA was outperformed by other rival models in our study. Combining the results of the simulation and empirical study together, we conclude that the poor performance of the SARFIMA model may be caused by the inaccurate estimation obtained by the WHI method. The estimation method for this model still needs further improvement. Before more effective and more accurate estimation methods are proposed, we suggest that the SARFIMA model should be carefully applied when forecasting a high-frequency long memory time series with long periods.

References

[1] Engle, R. (2005) Handbook of Financial Econometrics. North Holland, Amsterdam.

[2] Granger, C.W.J. and Joyeux, R. (1980) An Introduction to Long-Memory Time Series Models and Fractional Differencing. *Journal of Time Series Analysis*, **1**, 15-29. http://dx.doi.org/10.1111/j.1467-9892.1980.tb00297.x

[3] Hosking, J.R.M. (1981) Fractional Differencing. *Biometrika*, **68**, 165-176. http://dx.doi.org/10.1093/biomet/68.1.165

[4] Franses, P.H. and Ooms, M. (1997) A Periodic Long-Memory Model for Quarterly UK Inflation. *International Journal of Forecasting*, **13**, 117-126. http://dx.doi.org/10.1016/S0169-2070(96)00715-7

[5] Porter-Hudak, S. (1990) An Application of the Seasonal Fractionally Differenced Model to the Monetary Aggregates. *American Statistical Association*, **85**, 338-344. http://dx.doi.org/10.1080/01621459.1990.10476206

[6] Ooms, M. and Franses, P.H. (2001) A Seasonal Periodic Long Memory Model for Monthly River Flows. *Environmental Modeling & Software*, **16**, 559-569. http://dx.doi.org/10.1016/S1364-8152(01)00025-1

[7] Nasr, A.B. and Trabelsi, A. (2005) Seasonal and Periodic Long Memory Models in the Inflation Rates. *European Financial Management Association* 2005 *Annual Meetings*, **31**.

[8] Sowell, F. (1992) Maximum Likelihood Estimation of Stationary Univariate Fractionally Integrated Time Series Models. *Journal of Econometrics*, **53**, 165-188. http://dx.doi.org/10.1016/0304-4076(92)90084-5

[9] Whittle, P. (1953) Estimation and Information in Stationary Time Series. *Arkiv för Matematik*, **2**, 423-434. http://dx.doi.org/10.1007/BF02590998

[10] Doornik, J.A. and Ooms, M. (2004) Inference and Forecasting for ARFIMA Models, with an Application to US and UK Inflation. *Studies in Nonlinear Dynamics and Econometrics*, **8**, 1218. http://dx.doi.org/10.2202/1558-3708.1218

[11] Reisen, V.A., Rodrigues, A.L. and Palma, W. (2006) Estimation of Seasonal Fractionally Integrated Processes. *Computational Statistics & Data Analysis*, **50**, 568-582. http://dx.doi.org/10.1016/j.csda.2004.08.004

[12] Bisognin, C. and Lopes, S.R.C. (2007) Estimating and Forecasting the Long-Memory Parameter in the Presence of Periodicity. *Journal of Forecasting*, **26**, 405-427. http://dx.doi.org/10.1002/for.1030

[13] Geweke, J. and Porter-Hudak, S. (1983) The Estimation and Application of Long Memory Time Series Models. *Journal of Time Series Analysis*, **4**, 221-238. http://dx.doi.org/10.1111/j.1467-9892.1983.tb00371.x

[14] Robinson, P.M. (1995) Log Periodogram Regression of Time Series with Long Range Dependence. *Annals of Statistics*, **23**, 1048-1072. http://dx.doi.org/10.1214/aos/1176324636

Open Journal of Statistics, 2015, 5, 75-89
Published Online February 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.51010

On the Maximum Likelihood and Least Squares Estimation for the Inverse Weibull Parameters with Progressively First-Failure Censoring

Amal Helu

Department of Mathematics, The University of Jordan, Amman, Jordan
Email: a.helu@ju.edu.jo

Received 5 February 2015; accepted 23 February 2015; published 26 February 2015

Abstract

In this article, we consider a new life test scheme called a progressively first-failure censoring scheme introduced by Wu and Kus [1]. Based on this type of censoring, the maximum likelihood, approximate maximum likelihood and the least squares method estimators for the unknown parameters of the inverse Weibull distribution are derived. A comparison between these estimators is provided by using extensive simulation and two criteria, namely, absolute bias and mean squared error. It is concluded that the estimators based on the least squares method are superior compared to the maximum likelihood and the approximate maximum likelihood estimators. Real life data example is provided to illustrate our proposed estimators.

Keywords

Inverse Weibull Distribution, Progressive First-Failure Censoring, Maximum Likelihood, Least Squares Method

1. Introduction

Let T follow $(\sim)$ a two-parameter Weibull distribution (α,β) with the probability density function (*pdf*)

$$f(t;\alpha,\beta)=\frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}\mathrm{e}^{-(\alpha/t)^{-\beta}},\quad t>0$$

How to cite this paper: Helu, A. (2015) On the Maximum Likelihood and Least Squares Estimation for the Inverse Weibull Parameters with Progressively First-Failure Censoring. *Open Journal of Statistics*, **5**, 75-89.
http://dx.doi.org/10.4236/ojs.2015.51010

then $X = \frac{1}{T}$ has an (IW) distribution with *pdf*

$$f(x;\alpha,\beta) = \alpha\beta(\alpha x)^{-\beta-1}\mathrm{e}^{-(\alpha x)^{-\beta}}, \quad x > 0 \tag{1}$$

where $\alpha > 0$ and $\beta > 0$, are the scale and shape parameters, respectively.

If $X \sim \text{IW}(\alpha,\beta)$, then the cumulative distribution function (*cdf*) of X is given by:

$$F(x;\alpha,\beta) = \mathrm{e}^{-(\alpha x)^{-\beta}}, \quad x > 0. \tag{2}$$

The IW distribution, also known as type 2 extreme value or the Frechet distribution (Johnson *et al*. [2]), has a long right tail compared to other known distributions. The hazard function of the IW distribution is similar to that of the log-normal and inverse gaussian distributions (Murthy *et al*. [3]). Carriere [4] used the IW distribution to model the mortality curve of a population. Keller and Kamath [5] suggested that this distribution was suitable to model the failure of the degradation phenomena of mechanical components of diesel engines such as pistons, crankshafts, and main bearings. Furthermore, Erto [6] showed that the IW distribution provided a good fit to several data such as the time to breakdown of an insulating fluid subjected to the action of a constant tension (Nelson [7]).

Several researches have been carried out on the IW distribution using classical and Bayesian approaches. For example, Calabria and Pulcini [8] obtained the maximum likelihood estimates (MLE) and least squares estimates of the parameters of the IW distribution. Calabria and Pulcini [9] considered the Bayesian approach to predict the ordered lifetimes in a future sample when those lifetimes are assumed to follow the IW distribution. Panaitescu *et al*. [10] developed the Bayesian and non-Bayesian analysis in the context of recording statistic values from a modified IW distribution. All these studies have been done based on a complete sample. However, there are many scenarios in life testing and reliability experiment when researchers can not obtain complete information on failure times for all the units in the experiment as in the case of accidental breakage of an experimental unit or if an individual under study drops out. Moreover, there are many situations in which the researcher intentionally removes units prior to their failure and this is due to the lack of funds and/or time constrains. Data obtained from such experiments are called censored data. Therefore, we consider estimation procedures based on censored samples.

The most common censoring schemes are type-I censoring in which the test ceases at a pre-fixed time, and type-II censoring that allows the experiment to be terminated at a predetermined number of failures. These methods do not allow the removal of active units during the experiment; therefore, the focus in the last few years has been on progressive censoring due to its flexibility that allows the experimenter to remove active units during the experiment. A progressively type-II censoring is a generalization of type-II censoring. Many authors have discussed inference under progressive censoring using various lifetime distributions, among others, Cohen [11], Mann [12], Wingo [13], Balakrishnan and Sandhu [14], Aggarwala and Balakrishnan [15], Balakrishnan and Asgharzadeh [16]. For a comprehensive recent review of progressive censoring, readers are referred to Balakrishnan [17].

Johnson [18] introduced the first-failure censoring plan where the experimenter could arrange k items into n sets, then all the $k \times n$ items were tested simultaneously until the first failure in each n set occurred. However, in situations where the lifetime of a product is high and test facilities are limited but test material is cheap. Balasooriya [19] modified Johnson [18] approach by testing each set one after the other until the first failure in each set occurred. This modified approach can save time and money.

However, due to certain situations such as loss of contact with the individuals under study or loss of experiment units as mentioned above, it is desirable for researchers to be able to remove sets before the final termination point. This situation leads to the area of progressive censoring.

Wu and Kus [1] wanted to improve the efficiency of the test by developing a new life test scheme, progressively first-failure censoring scheme, by combining the concept of first failure censoring with the progressive censoring. In this scheme sets with no failures can be removed from the test before the end of the experiment. Based on this scheme, Wu and Kus [1] derived the MLEs and constructed exact and approximate confidence intervals for the parameters of the Weibull distribution. Wu and Huang [20] developed the reliability sampling plans for the Weibull distribution. Soliman *et al*. ([21] [22]) derived Bayes and frequentist estimators for the parameters of Gompertz and Burr type XII distributions respectively. Hong *et al*. [23] used the same scheme to construct MLE for the lifetime performance index C_L based on progressively first-failure censoring from Weibull distribution. Ahmadi *et al*. [24] developed a confidence interval and ML estimator for C_L based on the progressive first-failure censored sample under the Weibull distribution when the shape parameter was known.

In this study and based on *m* progressively first-failure censored sample from IW model, we consider the problem of estimating the parameters of the model using the maximum likelihood, the approximate maximum likelihood and the least square estimators (LSE). Balakrishnan *et al*. [25] conducted inference on progressive type-II censored data for extreme value distribution. They derived the MLE and approximate values for the maximum likelihood estimators (AMLE) using the Taylor expansion. They also concluded that the MLEs and AMLEs were almost identical in terms of bias and variance. Kim *et al*. [26] derived the maximum likelihood and the Bayes estimates for the three-parameter exponentiated Weibull distribution for type-II progressively censored sample. Gusmao *et al*. [27] studied the properties of a mixture of two generalized IW distribution and derived the maximum likelihood estimator of the parameters of this mixture based on censored data.

This article unfolds as follows: In Section 2 we describe the formulation of a progressive first-failure censoring scheme as described by Wu and Kus [1]. The MLEs, approximate MLEs, and LSE methods for estimating the unknown parameters based on the progressive first-failure censoring scheme are derived in Section 3, 4 and 5 respectively. Simulation studies, results and conclusion are presented in Section 6. All methods that are discussed in this article are illustrated in Section 7 through a real life data set coming from highways pavement projects in Amman-Jordan during 2012.

2. A Progressive First-Failure Censoring Scheme

The progressive first-failure censoring can be described as follows: Given $m \le n$ and $\mathbf{R} = R_1, R_2, \cdots, R_m$ non-negative integers such that $n = m + \sum_{i=1}^{m} R_i$. Let *n* independent groups with k items within each group be placed on a life testing experiment and only m failures are completely observed. The censoring occurs progressively in m stages. At the time of the first failure $X_{1:m:n:k}$, R_1 random groups and the group with the observed failure are randomly removed. Similarly, at the time of the second failure $X_{2:m:n:k}$, R_2 random groups and the group with the observed failure are randomly removed and so on. Finally, at the time of the m-th failure all the remaining active groups (R_m) and the group with the observed failure are removed. Then $X_{1:m:n:k} < X_{2:m:n:k} < \cdots < X_{m:m:n:k}$ is the progressive first-failure censored order statistics.

The main advantage of this scheme is that it reduces the time in which more items are used but only *m* out of $k \times n$ items are observed. Moreover, it includes as special cases, the progressively type-II scheme (when *k* = 1), first-failure scheme (when $\mathbf{R} = (0, 0, \cdots, 0)$), conventional type II scheme (when *k* = 1 and $\mathbf{R} = (0, 0, \cdots, n - m)$ and the complete sample (when $k = 1$, $n = m$ and $\mathbf{R} = (0, 0, \cdots, 0)$). Furthermore, the progressively first-failure censored sample $X_{1:m:n:k} < X_{2:m:n:k} < \cdots < X_{m:m:n:k}$ can be considered as a progressively type-II censored sample from a population with distribution function $1 - (1 - F(x))^k$ (Wu and Kus [1]) which enables us to extend all the results on progressively type-II censored order statistics to progressively first-failure censored order statistics.

3. Maximum Likelihood Estimation

Suppose that n independent units are placed on a test. The ordered m failures are observed under the progressively first-failure.

Let $\mathbf{X} = (X_{1:m:n:k}, X_{2:m:n:k}, \cdots, X_{m:m:n:k})$ with $X_{1:m:n:k} < \cdots < X_{m:m:n:k}$ denote the progressively first-failure censored ordered statistics with the progressive censoring scheme $\mathbf{R}$ from a population with *pdf* and *cdf* given in Equations (1) and (2), respectively. For notation simplicity, we will write X_i for $X_{i:m:n:k}$. The likelihood function based on progressively first-failure censored sample (see Wu and Kus [1]) is given by:

$$L(\alpha, \beta; \mathbf{X}) = Ak^m \prod_{i=1}^{m} f(x_i; \alpha, \beta)\left[1 - F(x_i; \alpha, \beta)\right]^{k(R_i+1)-1} \quad (3)$$

where,

$$A = n\left(n-1-R_1\right)\left(n-2-R_1-R_2\right)\cdots\left(n-\sum_{i=1}^{m}\left(R_i+1\right)\right).$$

In accordance with (1), (2) and (3), the log-likelihood function of α and β based on progressively first-failure censored sample $\mathbf{X}$ becomes

$$\begin{aligned}\ln L\left(\alpha,\beta;\mathbf{X}\right) = \text{constant} + m\ln\left(\alpha\beta\right) - \left(\beta+1\right)\sum_{i=1}^{m}\ln\left(\alpha x_i\right) \\ -\sum_{i=1}^{m}\left(\alpha x_i\right)^{-\beta} + \sum_{i=1}^{m}\left(k\left(R_i+1\right)-1\right)\ln\left(1-\mathrm{e}^{-\left(\alpha x_i\right)^{-\beta}}\right).\end{aligned} \tag{4}$$

The MLEs of the parameters α and β can be obtained by deriving (4) with respect to α and β and equating the normal equations to 0 as follows:

$$\frac{\partial \ln L\left(\alpha,\beta;\mathbf{X}\right)}{\partial\alpha} = \frac{-m+\sum_{i=1}^{m}\left(\alpha x_i\right)^{-\beta} - \sum_{i=1}^{m}\frac{\left(k\left(R_i+1\right)-1\right)\left(\alpha x_i\right)^{-\beta}\mathrm{e}^{-\left(\alpha x_i\right)^{-\beta}}}{1-\mathrm{e}^{-\left(\alpha x_i\right)^{-\beta}}}}{\alpha} = 0 \tag{5}$$

$$\begin{aligned}\frac{\partial \ln L\left(\alpha,\beta;\mathbf{X}\right)}{\partial\beta} = \frac{m}{\beta} - \sum_{i=1}^{m}\ln\left(\alpha x_i\right)\left(1-\left(\alpha x_i\right)^{-\beta}\right) \\ -\sum_{i=1}^{m}\frac{\left(k\left(R_i+1\right)-1\right)\left(\alpha x_i\right)^{-\beta}\mathrm{e}^{-\left(\alpha x_i\right)^{-\beta}}\ln\left(\alpha x_i\right)}{1-\mathrm{e}^{-\left(\alpha x_i\right)^{-\beta}}} = 0.\end{aligned} \tag{6}$$

The MLEs are exist and unique (see Calabria and Pulcini [8] and Marusic *et al*. [28]). Notice that there are no explicit solutions to (5) and (6). Hence, numerical methods are applied to solve the required equations. The maximum likelihood estimation method based on progressively censored data has been studied extensively, but traditionally, the Newton Raphson (NR) method was utilized to obtain the MLEs (Ng *et al*. [29]). However, the MLEs via the NR method are very sensitive to their initial parameters estimation value. In this article we propose using the Expectation-Maximization (EM) algorithm for computing the MLEs.

4. Approximate Maximum Likelihood Estimation

Since the MLE does not provide explicit estimators for the shape and scale parameters of the IW distribution as mentioned before, we derive approximate MLE (AMLE) for the parameters α and β.

Balakrishnan ([30]-[34]) and Balakrishnan and Vardan [35] developed the AMLE procedure. This procedure depends on the Taylor expansion of the likelihood function when the *pdf* under consideration belongs to the location-scale families. However, the IW distribution does not have the location-scale structure required for the AMLE procedure, but if we consider the transformation $Y = -\ln X$, then $Y \sim$ extreme value distribution and this distribution has this feature.

The *pdf* and *cdf* of Y are given respectively by

$$h\left(y;\mu,\sigma\right) = \frac{1}{\sigma}\mathrm{e}^{\frac{y-\mu}{\sigma}-\mathrm{e}^{\frac{y-\mu}{\sigma}}}, \quad -\infty < y < \infty \tag{7}$$

and,

$$H\left(y;\mu,\sigma\right) = 1-\mathrm{e}^{-\mathrm{e}^{\frac{y-\mu}{\sigma}}} \tag{8}$$

where, $\mu = \ln\alpha$ and $\sigma = \frac{1}{\beta}$ are the location and scale parameters respectively.

Hence, the AMLE procedure can be used to estimate the parameters α and β of the IW distribution.

Let $Y = \left(Y_1, Y_2, \cdots, Y_m\right)$ with $Y_1 < Y_2 < \cdots < Y_m$ denotes a progressively first-failure censored sample from (7) and (8). Then the joint *pdf* based on the censored sample is given by:

$$L\left(y_1,y_2,\cdots,y_m;\mu,\sigma\right)=\frac{c}{\sigma^m}\prod_{i=1}^{m} g\left(\frac{y_i-\mu}{\sigma}\right)\left(1-G\left(\frac{y_i-\mu}{\sigma}\right)\right)^{\left(k\left(R_i+1\right)-1\right)} \tag{9}$$

where,

$$c=n\left(n-R_1-1\right)\cdots\left(n-R_1-R_2-\cdots-R_{m-1}-m+1\right).$$

If $z_i=\dfrac{y_i-\mu}{\sigma}$, then (9) can be written as

$$L\left(z_1,z_2,\cdots,z_m;\mu,\sigma\right)=\frac{c}{\sigma^m}\prod_{i=1}^{m} g\left(z_i\right)\left(1-G\left(z_i\right)\right)^{\left(k\left(R_i+1\right)-1\right)} \tag{10}$$

with log-likelihood equation

$$\ln L\left(z_1,\cdots,z_m;\mu,\sigma\right)=c-m\ln\sigma+\sum_{i=1}^{m}\ln g\left(z_i\right)+\sum_{i=1}^{m}\left(k\left(R_i+1\right)-1\right)\ln\left(1-G\left(z_i\right)\right). \tag{11}$$

Taking derivatives with respect to μ and σ and equating them to zero, gives

$$\frac{\partial \ln L\left(z_i;\mu,\sigma\right)}{\partial\sigma}=-\frac{m}{\sigma}-\sum_{i=1}^{m}\frac{g'\left(z_i\right)}{g\left(z_i\right)}z_i+\sum_{i=1}^{m}\left(k\left(R_i+1\right)-1\right)z_i\frac{g\left(z_i\right)}{1-G\left(z_i\right)}=0 \tag{12}$$

$$\frac{\partial \ln L\left(z_i;\mu,\sigma\right)}{\partial\mu}=-\sum_{i=1}^{m}\frac{g'\left(z_i\right)}{g\left(z_i\right)}+\sum_{i=1}^{m}\left(k\left(R_i+1\right)-1\right)\frac{g\left(z_i\right)}{1-G\left(z_i\right)}=0. \tag{13}$$

Because of the presence of the terms $\dfrac{g\left(z_i\right)}{1-G\left(z_i\right)}$ and $\dfrac{g'\left(z_i\right)}{g\left(z_i\right)}$, Equations (12) and (13) do not have explicit solution. Hence, we consider a first-order Taylor approximation to $\Delta_1=\dfrac{g'\left(z_i\right)}{g\left(z_i\right)}$ and $\Delta_2=\dfrac{g\left(z_i\right)}{1-G\left(z_i\right)}$ around $\nu_{i:m:n}=E\left[Z_{i:m:n}\right]$ (see Balakrishnan and Aggarwala [36]; for reasoning).

From Balakrishnan and Aggarwala [36], if $u_{i:m:n}$ $i=1,2,\cdots,m$ denote a progressively first-failure censored sample from Uniform $(0,1)$ with censoring scheme $\left(R_1,\cdots,R_m\right)$, then $V_i=\dfrac{1-U_{m-i+1:m:n}}{1-U_{m-i:m:n}}$, $i=1,2,\cdots,m$ are statistically independent random variables from Beta $\left(i+\sum_{l=m-j+1}^{m}\left(k\left(R_l+1\right)-1\right),1\right)$ with

$$U_{i:m:n}=1-\prod_{l=m-i+1}^{m}V_l,\quad i=1,2,\cdots,m \tag{14}$$

and,

$$E\left[U_{i:m:n}\right]=1-\prod_{j=m-i+1}^{m}E\left(V_j\right)=1-\prod_{j=m-i+1}^{m}\frac{j+\sum_{l=m-j+1}^{m}\left(k\left(R_l+1\right)-1\right)}{1+j+\sum_{l=m-j+1}^{m}\left(k\left(R_l+1\right)-1\right)}. \tag{15}$$

The approximation is around $\nu_{i:m:n}=\ln\left(-\ln\left(1-E\left[U_{i:m:n}\right]\right)\right)$ upon expanding Δ_1 and Δ_2 around the point $\nu_{i:m:n}$ and keeping only the first two terms, we get

$$\Delta_1\left(z_i\right)\cong\Delta_1\left(\nu_i\right)+\Delta_1'\left(\nu_i\right)\left(z_i-\nu_i\right)=\varepsilon_i-\beta_i z_i \tag{16}$$

where,

$$\varepsilon_i=\Delta_1\left(\nu_i\right)-\nu_i\Delta_1'\left(\nu_i\right)=\left(1-\mathrm{e}^{\nu_i}\right)+\nu_i\mathrm{e}^{\nu_i},$$

$$\beta_i=-\Delta_1'\left(\nu_i\right)=\mathrm{e}^{\nu_i},\qquad\qquad i=1,2,\cdots,m$$

and,

$$\Delta_2(z_i) \cong \Delta_2(\nu_i) + \Delta_2'(\nu_i)(z_i - \nu_i) = \gamma_i + \delta_i z_i \tag{17}$$

where,

$$\gamma_i = \Delta_2(\nu_i) - \nu_i \Delta_2'(\nu_i) = \mathrm{e}^{\nu_i} - \nu_i \mathrm{e}^{\nu_i},$$
$$\delta_i = \Delta_2'(\nu_i) = \mathrm{e}^{\nu_i}, \qquad\qquad i = 1, 2, \cdots, m.$$

Plugging (16) and (17) in (12) and (13) we get

$$\begin{aligned}\frac{\partial \ln L(z_i;\mu,\sigma)}{\partial \sigma} &\cong -m - \sum_{i=1}^{m}(\varepsilon_i - \beta_i z_i) z_i + \sum_{i=1}^{m} R_i^* z_i (\gamma_i + \delta_i z_i) = 0 \\ &= -m - \sum_{i=1}^{m}(\varepsilon_i - R_i^*\gamma_i) z_i + \sum_{i=1}^{m}(R_i^*\delta_i + \beta_i) z_i^2 = 0 \\ &= -m - \sum_{i=1}^{m}(\varepsilon_i - R_i^*\gamma_i)\frac{y_i - \mu}{\sigma} + \sum_{i=1}^{m}(R_i^*\delta_i + \beta_i)\frac{(y_i-\mu)^2}{\sigma^2} = 0\end{aligned} \tag{18}$$

and,

$$\begin{aligned}\frac{\partial \ln L(z_i;\mu,\sigma)}{\partial \mu} &\cong -\sum_{i=1}^{m}(\varepsilon_i - \beta_i z_i) + \sum_{i=1}^{m} R_i^* (\gamma_i + \delta_i z_i) = 0 \\ &= -\sum_{i=1}^{m}(\varepsilon_i - R_i^*\gamma_i) + \sum_{i=1}^{m}(R_i^*\delta_i + \beta_i) z_i = 0 \\ &= -\sum_{i=1}^{m}(\varepsilon_i - R_i^*\gamma_i) + \sum_{i=1}^{m}(R_i^*\delta_i + \beta_i)\frac{y_i - \mu}{\sigma} = 0\end{aligned} \tag{19}$$

where, $R_i^* = k(R_i + 1) - 1$. Equations (18) and (19) can be rewritten as

$$0 = m\sigma^2 + A\sigma + B \tag{20}$$

$$\mu = D + C\sigma \tag{21}$$

where,

$$A = \sum_{i=1}^{m}(\varepsilon_i - R_i^*\gamma_i)(y_i - D); \qquad B = -\sum_{i=1}^{m}(R_i^*\delta_i + \beta_i)(y_i - D)^2 \le 0;$$

$$C = \frac{-\sum_{i=1}^{m}(\varepsilon_i - R_i^*\gamma_i)}{\sum_{i=1}^{m}(R_i^*\delta_i + \beta_i)}; \qquad D = \frac{\sum_{i=1}^{m}(R_i^*\delta_i + \beta_i) y_i}{\sum_{i=1}^{m}(R_i^*\delta_i + \beta_i)}.$$

The solutions to (20) and (21) yield the AMLEs

$$\hat{\sigma}_{\mathrm{AMLE}} = \frac{-A + \sqrt{(A)^2 - 4mB}}{2m}, \quad \hat{\mu}_{\mathrm{AMLE}} = D + C\hat{\sigma}_{\mathrm{AMLE}}. \tag{22}$$

One of the drawbacks of the AMLEs is that they are biased. Moreover, the exact bias of $\hat{\mu}_{\mathrm{AMLE}}$ and $\hat{\sigma}_{\mathrm{AMLE}}$ can not be theoretically computed because of the intractability encountered of $\sqrt{B^2 - 4AC}$. On the other hand, the AMLEs provide an excellent starting value for the iterative solution of the likelihood equations.

5. Least Squares Estimation Method

The LS method which is originally suggested by Swain *et al.* [37] is computationally easier to handle, it provides simple closed form solutions for estimates (Hossain and Zimmer [38]). In addition it can also be used quite effectively to estimate the shape and scale parameters of the IW distribution. Finally Marusic *et al.* [28] showed that the least squares estimators (LSE) for estimating the parameters of the IW distribution did exist.

In this section we will discuss the least squares method for estimating α and β using the set up in Section 1; that is, $X_1 < X_2 < \cdots < X_m$ are progressively first-failure censored sample from the IW distribution with censoring scheme $(R_1, \cdots, R_m)$. The LS method is a combination of parametric (F) and non-parametric $(\hat{F})$ distribution functions. It depends on the choice of $\hat{F}$ which should be as effective as possible. In our study we use $\hat{F}$ which is proposed by Montanari and Cacciari [39] as a non-parametric *cdf* for progressive type-II censored sample.

$$\hat{F}_{X_i}(x) = \frac{J_i - 0.5}{n + 0.25} \tag{23}$$

where,

$$J_i = J_{i-1} + \Delta, \quad i = 1, 2, \cdots, m \text{ and } J_0 = 0$$

and,

$$\Delta = \frac{n + 1 - J_{i-1}}{1 + \sum_{w=i}^{m} k(R_w + 1)}.$$

For the parametric *cdf* $F(x)$, Balakrishnan and Aggarwala [36] and Kim and Han [40], proposed

$$F_{X_j}(x) = 1 - l_{j-1} \sum_{i=1}^{j} \frac{a_{i,j}}{r_i} \left(1 - F^*(x)\right)^{r_i}, \quad j = 1, 2, \cdots, m \tag{24}$$

where,

$$F^*(x) = 1 - \left(1 - F(x)\right)^k, \quad l_{j-1} = \prod_{w=1}^{j} r_w, \quad r_i = m - i + 1 + \sum_{w=1}^{m} \left(k(R_w + 1) - 1\right),$$

$$a_{i,j} = \prod_{\substack{w=1 \\ w \neq i}}^{j} \frac{1}{r_w - r_i}, \qquad 1 \le i \le j \le m.$$

The procedure attempts to minimize the following function with respect to α and β

$$\sum_{j=1}^{m} \left(F_{X_j}(x) - \hat{F}_{X_i}(x)\right)^2. \tag{25}$$

The LSE estimates of α and β are denoted by $\hat{\alpha}_{\text{LSE}}$ and $\hat{\beta}_{\text{LSE}}$ respectively.

6. Simulation Study

The purpose of the simulation study is to compare the performance of the MLE, AMLE and LSE estimates based on progressively first-failure censored samples generated from the IW distribution with $(\alpha, \beta) = (0.1, 3.0)$, $(0.5, 3.0)$, $(0.9, 3.0)$, $(1.5, 2.5)$, $(2.5, 2.5)$, $(4.0, 4.0)$, using different combinations of *n*, *m*, *k* and different censoring schemes $\mathbf{R} = (R_1, \cdots, R_m)$. The data are simulated using Balakrishnan and Aggarwala [36] algorithm based on the fact that progressively first-failure censored sample with distribution $F(x)$ can be viewed as a progressively type II censored sample from a population with distribution function $1 - \left(1 - F(x)\right)^k$.

We obtain the MLEs of α and β by solving the nonlinear Equations (5) and (6), in which the AMLEs are used as starting values of the MLE iterations. The AMLE and LSE are computed using (22) and (25) respectively. The criteria used for comparing all the above estimates are the absolute bias (ABias) and the mean squared error (MSE). Suppose $\hat{\theta}_i \; (= \alpha_i, \beta_i)$ is the estimate of $\theta \; (= \alpha, \beta)$ for the *i*-th simulated data set, then the ABias and MSE are computed as follows:

1) $\text{ABias} = \frac{1}{7000} \sum_{i=1}^{7000} \left|\hat{\theta}_i - \theta\right|.$

2) $\text{MSE} = \frac{1}{7000} \sum_{i=1}^{7000} \left(\hat{\theta}_i - \theta\right)^2.$

6.1. Data Analysis and Comparison Study

Due to the large number of tables, only part of them is reported. Results are summarized in **Tables 1-4** provided at the end of this section as follows:

- **Table 1** and **Table 2** provide the ABias and MSE values for the estimates of α .
- **Table 3** and **Table 4** provide the ABias and MSE values for the estimates of β .

Throughout this section we will refer to $\mathbf{R}=(n-m,0,\cdots,0)$ by L_{n-m}, $\mathbf{R}=(a,\cdots,a,0,\cdots,0)$ by $L_{a,\cdots,a}$ (where $\sum a=n-m$), $\mathbf{R}=(0,\cdots,0,n-m)$ by $\mathbf{R}^{*}_{n-m}$, $\mathbf{R}=(0,\cdots,0,a,\cdots,a)$ by $\mathbf{R}^{*}_{a,\cdots,a}$, $\mathbf{R}=(0,\cdots,0,n-m,0,\cdots,0)$ by C_{n-m}, and finally $\mathbf{R}=(0,\cdots,0,a,\cdots,a,0,\cdots,0)$ by $C_{a,\cdots,a}$. Moreover, we will refer to schemes L_{n-m}, $\mathbf{R}^{*}_{n-m}$, and C_{n-m} by group-1, similarly we will refer to the schemes $L_{a,\cdots,a}$, $\mathbf{R}^{*}_{a,\cdots,a}$, and $C_{a,\cdots,a}$ by group-2. A summary of the results is provided below.

Table 1. Bias and MSE (parentheses) of $\hat{\alpha}_{(\cdot)}$ when $(\alpha,\beta)=(0.1,3)$.

			$k = 1$			$k = 3$			$k = 5$		
n	m	Scheme	MLE	AMLE	LSE	MLE	AMLE	LSE	MLE	AMLE	LSE
20	5	(15,0,0,0,0)	0.0022 (0.0009)	0.0107 (0.0012)	0.0048 (0.0010)	0.0051 (0.0006)	0.0107 (0.0013)	0.0073 (0.0007)	0.0062 (0.0006)	0.0101 (0.0019)	0.0107 (0.0007)
		(0,0,15,0,0)	0.0025 (0.0008)	0.0114 (0.0010)	0.0081 (0.0009)	0.0067 (0.0007)	0.0157 (0.0016)	0.0084 (0.0009)	0.0082 (0.0007)	0.0172 (0.0023)	0.0219 (0.0014)
		(0,0,0,0,15)	0.0019 (0.0006)	0.0084 (0.0010)	0.0015 (0.0007)	0.0077 (0.0006)	0.0130 (0.0031)	0.0043 (0.0007)	0.0098 (0.0008)	0.0140 (0.0049)	0.0095 (0.0007)
		(3,3,3,3,3)	0.0039 (0.0007)	0.0116 (0.0012)	0.0067 (0.0007)	0.0081 (0.0007)	0.0145 (0.0029)	0.0063 (0.0008)	0.0097 (0.0008)	0.0151 (0.0042)	0.0320 (0.0013)
	15	(5,...,0,0)	0.0003 (0.0007)	0.0062 (0.0007)	0.0011 (0.0007)	0.0031 (0.0004)	0.0077 (0.0008)	0.0065 (0.0005)	0.0042 (0.0004)	0.0080 (0.0012)	0.0076 (0.0005)
		(1,1,1,1,1,0,...,0)	0.0001 (0.0006)	0.0060 (0.0007)	0.0015 (0.0007)	0.0033 (0.0004)	0.0087 (0.0009)	0.0067 (0.0005)	0.0045 (0.0004)	0.0094 (0.0013)	0.0072 (0.0005)
		(0,..,0,5,0,...,0)	0.0001 (0.0006)	0.0061 (0.0007)	0.0032 (0.0006)	0.0036 (0.0004)	0.0099 (0.0010)	0.0071 (0.0005)	0.0049 (0.0004)	0.0112 (0.0014)	0.0077 (0.0005)
		$\left(0^{*5},1^{*5},0^{*5}\right)$	0.0001 (0.0006)	0.0061 (0.0007)	0.0033 (0.0006)	0.0036 (0.0004)	0.0099 (0.0010)	0.0072 (0.0005)	0.0049 (0.0004)	0.0112 (0.0014)	0.0078 (0.0005)
		(0,...,0,5)	0.0000 (0.0006)	0.0043 (0.0006)	0.0010 (0.0006)	0.0037 (0.0004)	0.0081 (0.0013)	0.0037 (0.0004)	0.0052 (0.0004)	0.0091 (0.0020)	0.0043 (0.0005)
		(0,...,0,1,1,1,1,1)	0.0002 (0.0006)	0.0052 (0.0007)	0.0034 (0.0006)	0.0037 (0.0004)	0.0091 (0.0012)	0.0051 (0.0004)	0.0052 (0.0004)	0.0104 (0.0018)	0.0080 (0.0005)
50	20	(30,0,............,0)	0.0016 (0.0005)	0.0063 (0.0005)	0.0038 (0.0005)	0.0024 (0.0003)	0.0049 (0.0007)	0.0040 (0.0004)	0.0029 (0.0003)	0.0040 (0.0010)	0.0034 (0.0004)
		(3,3,....,3,0...,0)	0.0017 (0.0004)	0.0064 (0.0005)	0.0040 (0.0004)	0.0033 (0.0003)	0.0075 (0.0008)	0.0046 (0.0004)	0.0039 (0.0004)	0.0078 (0.0011)	0.0039 (0.0006)
		(0,...0,30,0,...,0)	0.0019 (0.0004)	0.0068 (0.0005)	0.0059 (0.0005)	0.0037 (0.0004)	0.0091 (0.0009)	0.0049 (0.0005)	0.0043 (0.0004)	0.0100 (0.0012)	0.0034 (0.0005)
		$\left(0^{*5},3^{*10},0^{*5}\right)$	0.0018 (0.0004)	0.0073 (0.0005)	0.0072 (0.0005)	0.0038 (0.0004)	0.0099 (0.0010)	0.0052 (0.0005)	0.0045 (0.0004)	0.0109 (0.0014)	0.0039 (0.0005)
		(0............,0,30)	0.0016 (0.0003)	0.0049 (0.0003)	0.0024 (0.0003)	0.0044 (0.0004)	0.0070 (0.0021)	0.0005 (0.0004)	0.0055 (0.0004)	0.0074 (0.0031)	0.0020 (0.0003)
		(0,...,0,3,3,...,3)	0.0014 (0.0003)	0.0061 (0.0005)	0.0036 (0.0003)	0.0042 (0.0003)	0.0087 (0.0017)	0.0018 (0.0004)	0.0156 (0.0037)	0.0283 (0.0221)	0.0067 (0.0042)
50	30	(20,0,............,0)	0.0002 (0.0003)	0.0037 (0.0004)	0.0017 (0.0004)	0.0014 (0.0002)	0.0035 (0.0004)	0.0036 (0.0002)	0.0019 (0.0002)	0.0032 (0.0006)	0.0030 (0.0002)
		(2,2,....,2,0...,0)	0.0004 (0.0003)	0.0035 (0.0003)	0.0024 (0.0003)	0.0017 (0.0002)	0.0046 (0.0005)	0.0037 (0.0003)	0.0022 (0.0002)	0.0049 (0.0007)	0.0027 (0.0002)
		(0,...0,20,0,...,0)	0.0004 (0.0003)	0.0037 (0.0004)	0.0027 (0.0003)	0.0020 (0.0002)	0.0057 (0.0005)	0.0041 (0.0003)	0.0026 (0.0002)	0.0064 (0.0008)	0.0028 (0.0003)
		$\left(0^{*10},2^{*10},0^{*10}\right)$	0.0004 (0.0003)	0.0038 (0.0003)	0.0030 (0.0003)	0.0020 (0.0002)	0.0058 (0.0006)	0.0042 (0.0003)	0.0026 (0.0002)	0.0066 (0.0008)	0.0029 (0.0003)
		(0............,0,20)	0.0002 (0.0002)	0.0024 (0.0003)	0.0008 (0.0003)	0.0022 (0.0002)	0.0042 (0.0009)	0.0004 (0.0002)	0.0029 (0.0002)	0.0047 (0.0014)	0.0005 (0.0003)
		(0,...,0,2,2,...,2)	0.0002 (0.0003)	0.0031 (0.0004)	0.0009 (0.0003)	0.0021 (0.0002)	0.0052 (0.0008)	0.0011 (0.0002)	0.0028 (0.0002)	0.0058 (0.0012)	0.0007 (0.0003)

Table 2. Bias and MSE (parentheses) of $\hat{\alpha}_{(\cdot)}$ when $(\alpha,\beta)=(1.5,2.5)$.

			$k=1$			$k=3$			$k=5$		
n	*m*	Scheme	MLE	AMLE	LSE	MLE	AMLE	LSE	MLE	AMLE	LSE
20	5	(15,0,0,0,0)	0.0148 **(0.0341)**	0.0619 **(0.0355)**	0.0253 **(0.0333)**	0.0320 **(0.0231)**	0.0615 **(0.0459)**	0.0421 **(0.0292)**	0.0385 **(0.0235)**	0.0565 **(0.0683)**	0.1276 **(0.0488)**
		(0,0,15,0,0)	0.0164 **(0.0276)**	0.0662 **(0.0340)**	0.0465 **(0.0319)**	0.0414 **(0.0245)**	0.0906 **(0.0563)**	0.0359 **(0.0287)**	0.0505 **(0.0262)**	0.0982 **(0.0799)**	0.1871 **(0.0438)**
		(0,0,0,0,15)	0.0126 **(0.0234)**	0.0476 **(0.0320)**	0.0091 **(0.0249)**	0.0473 **(0.0244)**	0.0714 **(0.1139)**	0.0239 **(0.0262)**	0.0607 **(0.0276)**	0.0741 **(0.1794)**	0.0553 **(0.0263)**
		(3,3,3,3,3)	0.0248 **(0.0238)**	0.0668 **(0.0431)**	0.0385 **(0.0254)**	0.0502 **(0.0250)**	0.0809 **(0.1029)**	0.0484 **(0.0315)**	0.0601 **(0.0286)**	0.0814 **(0.1534)**	0.0625 **(0.0320)**
	15	(5,...,0,0)	0.0014 **(0.0251)**	0.0355 **(0.0249)**	0.0168 **(0.0245)**	0.0193 **(0.0147)**	0.0446 **(0.0303)**	0.0381 **(0.0173)**	0.0258 **(0.0155)**	0.0454 **(0.0445)**	0.0420 **(0.0190)**
		(1,1,1,1,1,0,...,0)	0.0019 **(0.0228)**	0.0347 **(0.0241)**	0.0079 **(0.0236)**	0.0209 **(0.0149)**	0.0500 **(0.0311)**	0.0391 **(0.0176)**	0.0277 **(0.0158)**	0.0539 **(0.0449)**	0.0442 **(0.0190)**
		(0,..,0,5,0,...,0)	0.0014 **(0.0212)**	0.0350 **(0.0234)**	0.0181 **(0.0216)**	0.0226 **(0.0156)**	0.0572 **(0.0422)**	0.0416 **(0.0187)**	0.0302 **(0.0161)**	0.0641 **(0.0497)**	0.0449 **(0.0191)**
		$\left(0^{*5},1^{*5},0^{*5}\right)$	0.0013 **(0.0211)**	0.0352 **(0.0234)**	0.0184 **(0.0214)**	0.0227 **(0.0155)**	0.0573 **(0.0455)**	0.0419 **(0.0186)**	0.0303 **(0.0161)**	0.0643 **(0.0503)**	0.0457 **(0.0194)**
		(0,...,0,5)	0.0003 **(0.0201)**	0.0247 **(0.0232)**	0.0010 **(0.0202)**	0.0221 **(0.0151)**	0.0456 **(0.0314)**	0.0212 **(0.0159)**	0.0322 **(0.0158)**	0.0505 **(0.0447)**	0.0247 **(0.0170)**
		(0,...,0,1,1,1,1,1)	0.0008 **(0.0203)**	0.0296 **(0.0233)**	0.0112 **(0.0202)**	0.0223 **(0.0152)**	0.0523 **(0.0318)**	0.0299 **(0.0160)**	0.0318 **(0.0159)**	0.0584 **(0.0449)**	0.0270 **(0.0176)**
50	20	(30,0,............,0)	0.0105 **(0.0190)**	0.0365 **(0.0184)**	0.0135 **(0.0177)**	0.0154 **(0.0122)**	0.0283 **(0.0233)**	0.0230 **(0.0151)**	0.0179 **(0.0117)**	0.0220 **(0.0344)**	0.0294 **(0.0138)**
		(3,3,....,3,0...,0)	0.0112 **(0.0143)**	0.0373 **(0.0174)**	0.0232 **(0.0155)**	0.0205 **(0.0125)**	0.0433 **(0.0279)**	0.0267 **(0.0159)**	0.0238 **(0.0130)**	0.0445 **(0.0392)**	0.0198 **(0.0146)**
		(0,...0,30,0,...,0)	0.0120 **(0.0134)**	0.0395 **(0.0179)**	0.0345 **(0.0170)**	0.0228 **(0.0129)**	0.0527 **(0.0312)**	0.0282 **(0.0170)**	0.0268 **(0.0138)**	0.0573 **(0.0435)**	0.0195 **(0.0223)**
		$\left(0^{*5},3^{*10},0^{*5}\right)$	0.0115 **(0.0127)**	0.0428 **(0.0186)**	0.0423 **(0.0170)**	0.0236 **(0.0128)**	0.0572 **(0.0358)**	0.0300 **(0.0180)**	0.0281 **(0.0140)**	0.0620 **(0.0508)**	0.0221 **(0.0211)**
		(0............,0,30)	0.0093 **(0.0111)**	0.0276 **(0.0134)**	0.0057 **(0.0120)**	0.0274 **(0.0128)**	0.0380 **(0.0237)**	0.0092 **(0.0129)**	0.0339 **(0.0155)**	0.0383 **(0.0346)**	0.0105 **(0.0125)**
		(0,...,0,3,3,...,3)	0.0105 **(0.0115)**	0.0351 **(0.0137)**	0.0207 **(0.0123)**	0.0273 **(0.0128)**	0.0406 **(0.0257)**	0.0117 **(0.0148)**	0.0338 **(0.0154)**	0.0414 **(0.0349)**	0.0116 **(0.0130)**
50	30	(20,0,............,0)	0.0018 **(0.0124)**	0.0217 **(0.0129)**	0.0052 **(0.0135)**	0.0089 **(0.0076)**	0.0202 **(0.0157)**	0.0209 **(0.0087)**	0.0116 **(0.0078)**	0.0176 **(0.0230)**	0.0177 **(0.0089)**
		(2,2,....,2,0...,0)	0.0028 **(0.0110)**	0.0202 **(0.0121)**	0.0057 **(0.0115)**	0.0108 **(0.0077)**	0.0267 **(0.0166)**	0.0218 **(0.0091)**	0.0137 **(0.0082)**	0.0282 **(0.0235)**	0.0159 **(0.0093)**
		(0,...0,20,0,...,0)	0.0030 **(0.0099)**	0.0214 **(0.0127)**	0.0155 **(0.0106)**	0.0124 **(0.0081)**	0.0329 **(0.0191)**	0.0240 **(0.0101)**	0.0158 **(0.0086)**	0.0369 **(0.0274)**	0.0164 **(0.0102)**
		$\left(0^{*10},2^{*10},0^{*10}\right)$	0.0028 **(0.0098)**	0.0219 **(0.0130)**	0.0175 **(0.0103)**	0.0125 **(0.0081)**	0.0339 **(0.0199)**	0.0246 **(0.0103)**	0.0160 **(0.0086)**	0.0379 **(0.0287)**	0.0170 **(0.0103)**
		(0............,0,20)	0.0015 **(0.0089)**	0.0136 **(0.0119)**	0.0038 **(0.0091)**	0.0134 **(0.0081)**	0.0236 **(0.0158)**	0.0026 **(0.0080)**	0.0179 **(0.0087)**	0.0255 **(0.0237)**	0.0026 **(0.0082)**
		(0,...,0,2,2,...,2)	0.0018 **(0.0090)**	0.0158 **(0.0119)**	0.0043 **(0.0092)**	0.0132 **(0.0082)**	0.0295 **(0.0160)**	0.0062 **(0.0086)**	0.0176 **(0.0086)**	0.0325 **(0.0239)**	0.0046 **(0.0104)**

Table 3. Bias and MSE (parentheses) of $\hat{\beta}_{(\cdot)}$ when $(\alpha,\beta)=(0.1,3.0)$.

			$k=1$			$k=3$			$k=5$		
n	m	Scheme	MLE	AMLE	LSE	MLE	AMLE	LSE	MLE	AMLE	LSE
20	5	(15,0,0,0,0)	0.3952 (0.8406)	0.2509 (0.8384)	0.2277 (0.8371)	0.3398 (0.6624)	0.2265 (0.6705)	0.0096 (0.5997)	0.3297 (0.6258)	0.2229 (0.6688)	0.0704 (0.5040)
		(0,0,15,0,0)	0.4901 (1.0275)	0.4286 (0.6892)	0.1824 (0.9365)	0.4350 (0.8549)	0.3974 (0.6069)	0.1780 (0.8316)	0.4234 (0.8154)	0.3930 (0.6049)	0.1698 (0.7361)
		(0,0,0,0,15)	0.4672 (0.9443)	0.4485 (1.0773)	0.0993 (0.7441)	0.5305 (1.0838)	0.5904 (1.5607)	0.3799 (1.0991)	0.5388 (1.1071)	0.5898 (1.5602)	0.0127 (1.0111)
		(3,3,3,3,3)	0.5226 (1.0838)	0.6008 (1.5919)	0.1362 (0.8806)	0.4521 (0.8823)	0.4351 (1.0524)	0.3536 (0.8638)	0.4523 (0.8785)	0.4337 (1.0516)	0.3031 (0.6174)
	15	(5,...,0,0)	0.2727 (0.5034)	0.2086 (0.4627)	0.1796 (0.5299)	0.2536 (0.4192)	0.1985 (0.4616)	0.0559 (0.4236)	0.2463 (0.4010)	0.1935 (0.4581)	0.1240 (0.4043)
		(1,1,1,1,1,0,...,0)	0.2870 (0.5211)	0.2396 (0.4554)	0.1744 (0.5403)	0.2659 (0.4350)	0.2286 (0.4551)	0.0976 (0.4444)	0.2551 (0.4146)	0.2240 (0.4509)	0.1324 (0.4760)
		(0,..,0,5,0,...,0)	0.3039 (0.5416)	0.2927 (0.5030)	0.1746 (0.5919)	0.2877 (0.4690)	0.2803 (0.5028)	0.1236 (0.4981)	0.2780 (0.4515)	0.2757 (0.4983)	0.1535 (0.4500)
		$\left(0^{*5},1^{*5},0^{*5}\right)$	0.3034 (0.5400)	0.2947 (0.5099)	0.1775 (0.5911)	0.2883 (0.4696)	0.2824 (0.5097)	0.1228 (0.4989)	0.2788 (0.4529)	0.2777 (0.5051)	0.1504 (0.4504)
		(0,...,0,5)	0.3018 (0.5252)	0.3218 (0.6662)	0.0040 (0.4452)	0.3025 (0.4892)	0.3198 (0.6346)	0.1992 (0.4717)	0.2986 (0.4850)	0.3161 (0.6295)	0.1554 (0.4616)
		(0,...,0,1,1,1,1,1)	0.3028 (0.5292)	0.3232 (0.6320)	0.0827 (0.4675)	0.3060 (0.4937)	0.3214 (0.6698)	0.2616 (0.4977)	0.3036 (0.4930)	0.3180 (0.6647)	0.1797 (0.5039)
50	20	(30,0,............,0)	0.2215 (0.3003)	0.0733 (0.2724)	0.1641 (0.3699)	0.1421 (0.2401)	0.0633 (0.2676)	0.0463 (0.2582)	0.1401 (0.2317)	0.0646 (0.2701)	0.0443 (0.2257)
		(3,3,.....,3,0...,0)	0.3028 (0.5292)	0.3232 (0.2604)	0.0827 (0.4675)	0.1770 (0.2761)	0.1319 (0.2403)	0.0542 (0.2996)	0.1726 (0.2650)	0.1322 (0.2428)	0.1270 (0.2779)
		(0,...0,30,0,...,0)	0.2230 (0.3678)	0.1910 (0.6320)	0.2068 (0.4477)	0.1966 (0.3043)	0.1780 (0.2555)	0.0786 (0.3403)	0.1921 (0.2934)	0.1781 (0.2579)	0.1607 (0.3168)
		$\left(0^{*5},3^{*10},0^{*5}\right)$	0.2257 (0.3664)	0.2167 (0.3062)	0.2153 (0.4950)	0.2062 (0.3176)	0.2024 (0.2995)	0.0846 (0.3798)	0.2040 (0.3125)	0.2027 (0.3029)	0.1610 (0.3526)
		(0.............,0,30)	0.2351 (0.3702)	0.2655 (0.4825)	0.0336 (0.3275)	0.3012 (0.3582)	0.2795 (0.4750)	0.2470 (0.3607)	0.2389 (0.4000)	0.2633 (0.4818)	0.2199 (0.3850)
		(0,...,0,3,3,...,3)	0.2427 (0.3793)	0.2834 (0.5672)	0.0433 (0.3725)	0.2484 (0.3829)	0.2802 (0.5587)	0.3203 (0.4757)	0.2550 (0.3693)	0.2829 (0.5671)	0.2466 (0.5135)
50	30	(20,0,............,0)	0.1484 (0.2006)	0.0675 (0.1995)	0.1197 (0.2753)	0.1074 (0.1665)	0.0636 (0.1927)	0.0070 (0.1934)	0.1063 (0.1603)	0.0639 (0.1928)	0.0785 (0.1773)
		(2,2,.....,2,0...,0)	0.1779 (0.2055)	0.0992 (0.1738)	0.1279 (0.2730)	0.1176 (0.1731)	0.0969 (0.1707)	0.0408 (0.1987)	0.1149 (0.1648)	0.0968 (0.1711)	0.1132 (0.1805)
		(0,...0,20,0,...,0)	0.1661 (0.2208)	0.1423 (0.1901)	0.1524 (0.3087)	0.1323 (0.1913)	0.1310 (0.1849)	0.0563 (0.2292)	0.1381 (0.1842)	0.1309 (0.1853)	0.1302 (0.2095)
		$\left(0^{*10},2^{*10},0^{*10}\right)$	0.1760 (0.2164)	0.1376 (0.1894)	0.1408 (0.3137)	0.1341 (0.1933)	0.1362 (0.1930)	0.0560 (0.2363)	0.1373 (0.1871)	0.1359 (0.1934)	0.1324 (0.2153)
		(0.............,0,20)	0.1523 (0.2137)	0.1705 (0.3002)	0.0090 (0.2088)	0.1629 (0.2057)	0.1656 (0.2762)	0.1459 (0.2124)	0.2128 (0.2072)	0.1665 (0.2773)	0.1480 (0.2230)
		(0,...,0,2,2,...,2)	0.1771 (0.2107)	0.1609 (0.2795)	0.0554 (0.2017)	0.1488 (0.2090)	0.1663 (0.2988)	0.2209 (0.2393)	0.2567 (0.2122)	0.1674 (0.3001)	0.1518 (0.2590)

Table 4. Bias and MSE (parentheses) of $\hat{\beta}_{(\cdot)}$ when $(\alpha,\beta)=(1.5,2.5)$.

			$k=1$			$k=3$			$k=5$		
n	m	Scheme	MLE	AMLE	LSE	MLE	AMLE	LSE	MLE	AMLE	LSE
20	5	(15,0,0,0,0)	0.3239 (0.5727)	0.2049 (0.5684)	0.1923 (0.6001)	0.2831 (0.4600)	0.1887 (0.5603)	0.0080 (0.4164)	0.2747 (0.4346)	0.1858 (0.5590)	0.0106 (0.3500)
		(0,0,15,0,0)	0.4024 (0.7003)	0.3524 (0.4719)	0.1562 (0.6383)	0.3625 (0.5937)	0.3312 (0.4656)	0.1483 (0.5775)	0.3528 (0.5662)	0.3275 (0.4645)	0.1415 (0.5112)
		(0,0,0,0,15)	0.4295 (0.7370)	0.4944 (1.0855)	0.0784 (0.5102)	0.4421 (0.7526)	0.4920 (1.0838)	0.3166 (0.7633)	0.4490 (0.7688)	0.4915 (1.0835)	0.0586 (0.7021)
		(3,3,3,3,3)	0.3839 (0.6434)	0.3686 (0.7339)	0.1096 (0.5713)	0.3768 (0.6127)	0.3626 (0.7308)	0.2946 (0.5999)	0.3769 (0.6101)	0.3614 (0.7303)	0.2526 (0.4288)
	15	(5,...,0,0)	0.2325 (0.3490)	0.1761 (0.3233)	0.1459 (0.3698)	0.2113 (0.2911)	0.1654 (0.3206)	0.0466 (0.2941)	0.2074 (0.2786)	0.1635 (0.3201)	0.1139 (0.2821)
		(1,1,1,1,1,0,...,0)	0.2446 (0.3615)	0.2016 (0.3187)	0.1416 (0.3773)	0.2216 (0.3021)	0.1905 (0.3160)	0.0813 (0.3086)	0.2168 (0.2884)	0.1885 (0.3156)	0.1257 (0.3321)
		(0,..,0,5,0,...,0)	0.2587 (0.3758)	0.2457 (0.3520)	0.1418 (0.4135)	0.2398 (0.3257)	0.2336 (0.3492)	0.1030 (0.3459)	0.2358 (0.3142)	0.2315 (0.3488)	0.1315 (0.3141)
		$(0^{*5},1^{*5},0^{*5})$	0.2583 (0.3747)	0.2474 (0.3568)	0.1443 (0.4133)	0.2402 (0.3261)	0.2354 (0.3540)	0.1023 (0.3464)	0.2366 (0.3151)	0.2332 (0.3535)	0.1289 (0.3147)
		(0,...,0,5)	0.2566 (0.3643)	0.2705 (0.4660)	0.0014 (0.3093)	0.2550 (0.3428)	0.2679 (0.4652)	0.2180 (0.3456)	0.2571 (0.3429)	0.2673 (0.4650)	0.1540 (0.3484)
		(0,...,0,1,1,1,1,1)	0.2575 (0.3671)	0.2715 (0.4421)	0.0651 (0.3259)	0.2521 (0.3397)	0.2665 (0.4407)	0.1660 (0.3276)	0.2529 (0.3374)	0.2656 (0.4404)	0.1332 (0.3203)
50	20	(30,0,............,0)	0.1351 (0.2050)	0.0589 (0.1871)	0.1336 (0.2514)	0.1184 (0.1668)	0.0527 (0.1858)	0.0386 (0.1793)	0.1151 (0.1580)	0.0517 (0.1856)	0.0388 (0.1533)
		(3,3,....,3,0...,0)	0.1681 (0.2329)	0.1178 (0.1682)	0.1577 (0.2745)	0.1475 (0.1918)	0.1099 (0.1669)	0.0452 (0.2081)	0.1421 (0.1802)	0.1086 (0.1666)	0.1037 (0.1886)
		(0,...0,30,0,...,0)	0.1839 (0.2503)	0.1576 (0.1789)	0.1749 (0.3040)	0.1638 (0.2113)	0.1484 (0.1774)	0.0655 (0.2363)	0.1583 (0.1996)	0.1469 (0.1772)	0.1318 (0.2152)
		$(0^{*5},3^{*10},0^{*5})$	0.1861 (0.2491)	0.1786 (0.2099)	0.1828 (0.3361)	0.1719 (0.2206)	0.1686 (0.2080)	0.0705 (0.2638)	0.1679 (0.2122)	0.1670 (0.2077)	0.1312 (0.2392)
		(0............,0,30)	0.2001 (0.2576)	0.2339 (0.3881)	0.0304 (0.2217)	0.2070 (0.2659)	0.2335 (0.3880)	0.2069 (0.3303)	0.2103 (0.2721)	0.2335 (0.3879)	0.2090 (0.3532)
		(0,...,0,3,3,...,3)	0.1937 (0.2513)	0.2189 (0.3302)	0.0346 (0.2549)	0.2058 (0.2641)	0.2329 (0.3821)	0.2010 (0.3070)	0.2089 (0.2698)	0.2328 (0.3821)	0.2025 (0.3303)
50	30	(20,0,............,0)	0.1402 (0.1362)	0.0572 (0.1344)	0.0973 (0.1925)	0.0895 (0.1157)	0.0530 (0.1338)	0.0058 (0.1343)	0.0883 (0.1111)	0.0523 (0.1337)	0.0653 (0.1230)
		(2,2,....,2,0...,0)	0.1418 (0.1434)	0.0853 (0.1190)	0.1087 (0.1929)	0.0980 (0.1202)	0.0808 (0.1185)	0.0340 (0.1380)	0.0955 (0.1142)	0.0800 (0.1184)	0.0950 (0.1253)
		(0,...0,20,0,...,0)	0.1471 (0.1515)	0.1146 (0.1289)	0.1193 (0.2173)	0.1102 (0.1328)	0.1092 (0.1284)	0.0469 (0.1591)	0.1080 (0.1276)	0.1083 (0.1283)	0.1049 (0.1451)
		$(0^{*10},2^{*10},0^{*10})$	0.1524 (0.1509)	0.1191 (0.1346)	0.1195 (0.2243)	0.1117 (0.1342)	0.1135 (0.1340)	0.0467 (0.1641)	0.1143 (0.1297)	0.1126 (0.1340)	0.1098 (0.1494)
		(0............,0,20)	0.1210 (0.1470)	0.1389 (0.2075)	0.0051 (0.1429)	0.1840 (0.1451)	0.1386 (0.2075)	0.1241 (0.1662)	0.2132 (0.1470)	0.1385 (0.2075)	0.1259 (0.1799)
		(0,...,0,2,2,...,2)	0.1206 (0.1475)	0.1392 (0.1920)	0.0466 (0.1442)	0.1216 (0.1429)	0.1380 (0.1918)	0.1158 (0.1475)	0.1769 (0.1435)	0.1378 (0.1918)	0.1229 (0.1548)

6.1.1. Scale Parameter α

- For progressively first-failure censoring $(k = 3\&5)$ we can easily notice that L_{n-m} is the most efficient scheme in terms of ABias and MSE values for MLE and AMLE, while scheme $\mathbf{R}^*_{n-m}$ is the most efficient scheme for LSE. On the other hand when $k = 1$, that is the progressively type-II censoring, scheme $\mathbf{R}^*_{n-m}$ is the most efficient for all estimates namely MLE, AMLE and LSE.
- Notice that when α is small (<1), the MSE values are almost identical for all the estimates regardless of the different schemes and the different values of k, this indicates that the estimates of α are sensitive to the choice of α.
- In general, LSE and MLE have comparable ABias and MSE values, which makes LSE estimates very good competitors to the MLE estimates.

6.1.2. Shape Parameter β

- When $k = 3\&5$ scheme L_{n-m} is the most efficient scheme in terms of ABias and MSE values for MLE and LSE.
- For progressively type-II censoring $(k = 1)$ scheme L_{n-m} is the most efficient in terms of ABias and MSE values for MLE while scheme $\mathbf{R}^*_{n-m}$ is the most efficient for LSE.
- As for the AMLE estimates, we notice that the scheme L_{n-m} is the most efficient in terms of ABias whereas scheme $L_{a,\cdots,a}$ is the most efficient in terms of MSE values for all values of k.
- In addition, $\hat{\beta}_{\text{LSE}}$ generally has the smallest ABias while $\hat{\beta}_{\text{MLE}}$ has the smallest MSE values.

6.2. Conclusions and Recommendations

In the past few years, progressive censoring has received a great attention by many researchers. This is due to its advantages in reducing the cost and time of the tests. Moreover, the availability of high speed computing resources enhances the focus on progressive censoring. In this article, we have considered the MLE, approximate MLE and LSE to estimate the unknown parameters of the IW distribution when data under consideration are progressively first-failure censoring.

It is out of question that all estimates are affected by the choice of k, and our goal is to compare the three methods namely MLE, AMLE and LSE and decide which is the most efficient for estimating α and β. It is important to point out the following:

- The results for group-1 and group-2 are very similar with slight edge improvement in favor of group-1.
- ABias and MSE values decrease as the effective sample proportion m/n increases for fixed $k\,\&\,n$ and for all estimates of α and β.
- In general, progressively first-failure censoring (*i.e.* $k = 3\&5$) is more efficient compared to progressive type-II censoring $(k = 1)$ in terms of ABias and MSE values. This is true for MLE and LSE estimates.
- **Table 1** and **Table 2** clearly show that the MSE values for LSE and MLE are almost identical and their ABias is comparable. Moreover, **Table 3** and **Table 4** show the similarity in performance between LSE and MLE for estimating β. Keep in mind that LSE formula is simple and easy to implement compared to the formula of the MLE.

Based on this, we highly recommend using LSE method and progressively first-failure censoring scheme for estimating the parameters of the IW distribution.

7. Real Life Data

In this example, we consider a real life data set to illustrate the proposed method and verify how our estimators work in practice. The validity of the IW model is checked using Kolmogrov-Smirnov $(K-S)$ test, as well as Anderson-Darling $(A-D)$ and chi-square tests. The data set for this application came from a real highway construction project in Amman/Jordan supervised by the Greater Amman Municipality and executed by a local contractor in 2012 (http://www.ammancity.gov.jo/en/gam/index.asp). The data consist of 64 readings that demonstrate the percentage of asphalt content in hot mix asphalt specimens sampled from the mentioned project above. Percentage of asphalt content is one of the main elements of a hot mix asphalt sample characteristics that has a direct effect on the quality and durability of the pavement. That is why this data is used in this example.

$(4.45,4.82)$	$(4.69,4.79)$	$(4.95,4.87)$	$(4.29,4.70)$	$(4.87,4.54)$	$(4.87,4.73)$
$(4.86,4.26)$	$(4.29,4.54)$	$(4.72,4.62)$	$(4.54,4.73)$	$(4.52,4.74)$	$(4.58,4.93)$
$(4.98,4.28)$	$(4.61,4.35)$	$(4.65,4.85)$	$(4.70,4.70)$	$(4.87,4.98)$	$(4.46,4.66)$
$(4.87,4.44)$	$(4.86,4.60)$	$(4.77,4.58)$	$(4.82,5.08)$	$(4.73,4.62)$	$(5.11,4.89)$
$(4.84,4.76)$	$(5.04,4.88)$	$(4.75,4.74)$	$(4.80,4.77)$	$(4.72,4.72)$	$(4.77,4.53)$
$(4.51,4.59)$	$(4.70,4.82)$				

We fit the IW distribution based on $\alpha = 0.209$ and $\beta = 29.083$. We observe that $K - S = 0.0864$ with $p_{\text{value}} = 0.8177$, $A - D = 0.3621$ and chi-square distance = 0.6468 with a corresponding $p_{\text{value}} = 0.98576$. This indicates that the IW model provides a good fit. The initial estimates for the MLEs are chosen by using pseudo complete estimates of the MLEs. We group the data into 32 sets with 2 items in each. We modify the data to consider four types of censoring as follows:

Case	Type of censoring	Censoring scheme
1	Complete data set	$k = 1,\ R = (0,0,\cdots,0)$
2	First failure censoring	$k = 2,\ R = (0,0,\cdots,0)$
3	Progressive type-II	$k = 1,\ R = (0^{*39},24)$
4	Progressive first-failure censoring	$k = 2,\ R = (12,0,\cdots,0)$

The modified data sets are provided in **Table 5**. The evaluated Hessian matrix to guarantee the uniqueness of the MLEs is presented in **Table 6**. Finally, the estimates of α and β based on different estimation methods are provided in **Table 7**.

Table 5. Progressive first-failure censored samples for the percentage of asphalt content in hot mix samples.

Case	n	m	Censored data											
1	64	64	4.26	4.28	4.29	4.29	4.35	4.44	4.45	4.46	4.51	4.52	4.53	4.54
			4.54	4.54	4.58	4.58	4.59	4.60	4.61	4.62	4.62	4.65	4.66	4.69
			4.70	4.70	4.70	4.70	4.72	4.72	4.72	4.73	4.73	4.73	4.74	4.74
			4.75	4.76	4.77	4.77	4.77	4.79	4.80	4.82	4.82	4.82	4.84	4.85
			4.86	4.86	4.87	4.87	4.87	4.87	4.87	4.88	4.89	4.93	4.95	4.98
			4.98	5.04	5.08	5.11								
2	32	32	4.45	4.69	4.87	4.29	4.54	4.73	4.26	4.29	4.62	4.54	4.52	4.58
			4.28	4.35	4.65	4.70	4.87	4.46	4.44	4.60	4.58	4.82	4.62	4.89
			4.76	4.88	4.74	4.77	4.72	4.53	4.51	4.70				
3	64	40	4.26	4.28	4.29	4.29	4.35	4.44	4.45	4.46	4.52	4.54	4.54	4.54
			4.58	4.60	4.61	4.62	4.65	4.66	4.69	4.70	4.70	4.70	4.72	4.73
			4.73	4.74	4.79	4.82	4.85	4.86	4.86	4.87	4.87	4.87	4.87	4.87
			4.93	4.95	4.98	4.98								
4	32	20	4.35	4.44	4.45	4.46	4.51	4.53	4.58	4.60	4.62	4.65	4.70	4.70
			4.72	4.74	4.76	4.77	4.82	4.87	4.88	4.89				

Table 6. The eigne-values and the determinant of the Hessian matrix for each of the four data sets.

Case 1		Case 2		Case 3		Case 4	
Eigen values	$\lvert\text{Hessian}\rvert$	Eigen values	$\lvert\text{Hessian}\rvert$	Eigen values	$\lvert\text{Hessian}\rvert$	Eigen values	$\lvert\text{Hessian}\rvert$
$\begin{pmatrix} -0.220 \\ -70083 \end{pmatrix}$	1541.56	$\begin{pmatrix} -0.186 \\ -41514 \end{pmatrix}$	7719.1	$\begin{pmatrix} -0.127 \\ -36313 \end{pmatrix}$	4626.0	$\begin{pmatrix} -0.083 \\ -44025 \end{pmatrix}$	3650.6

Table 7. The corresponding estimates.

Method	$\hat{\alpha}_{\text{MLE}}$	$\hat{\alpha}_{\text{AMLE}}$	$\hat{\alpha}_{\text{LSE}}$	$\hat{\beta}_{\text{MLE}}$	$\hat{\beta}_{\text{AMLE}}$	$\hat{\beta}_{\text{LSE}}$
Case 1	0.2172	0.2037	0.2169	22.5324	40.9753	27.7880
Case 2	0.2159	0.2172	0.2121	18.2934	33.4149	06.6950
Case 3	0.2174	0.2168	0.2196	22.5052	31.7096	26.1792
Case 4	0.2137	0.2133	0.2193	22.5059	39.1295	24.2470

It is quite clear that all the estimates for the scale parameter (α) are quite close to each other. It is of great importance to notice through this analysis that the estimates based on progressively first-failure are comparable with the values of the estimates based on progressively type-II censored samples and they are very close to those of the complete data set. Although $\hat{\beta}_{\text{AMLE}}$ is higher than $\hat{\beta}_{\text{MLE}}$ and $\hat{\beta}_{\text{LSE}}$, it is however comparable with its value when data is complete. Moreover, in this case $\hat{\beta}_{\text{AMLE}}$ is the closest to the complete case.

References

[1] Wu, S. and Kus, C. (2009) On Estimation Based on Progressive First-Failure Censored Sampling. *Computational Statistics and Data Analysis*, **53**, 3659-3670. http://dx.doi.org/10.1016/j.csda.2009.03.010

[2] Johnson, N., Kotz, S. and Balakrishnan, N. (1994) Continuous Univariate Distribution. Vol. 1, John Wiley and Sons, New York.

[3] Murthy, D.N.P., Bulmer, M. and Eccleston, J.A. (2004) Weibul Model Selection for Reliability Modeling. *Reliability Engineering & System Safety*, **86**, 257-267. http://dx.doi.org/10.1016/j.ress.2004.01.014

[4] Carriere, J. (1992) Parametric Models for Life Tables. *Transactions of Society of Actuaries*, **44**, 77-100.

[5] Keller, A.Z. and Kamath, A.R.R. (1982) Alternative Reliability Models for Mechanical Systems. *Proceeding of the 3rd International Conference on Reliability and Maintainability*, 411-415.

[6] Erto, P. (1986) Properties and Identification of the Inverse Weibull: Unknown or Just Forgotten. *Quality and Reliability Engineering International*, **9**, 383-385.

[7] Nelson, W. (1982) Applied Life Data Analysis. Wiley, New York. http://dx.doi.org/10.1002/0471725234

[8] Calabria, R. and Pulcini, G. (1990) On the Maximum Likelihood and Least Squares Estimation in the Inverse Weibull Distribution. *Statistics Applicata*, **2**, 53-66.

[9] Calabria, R. and Pulcini, G. (1994) Bayes 2-Sample Prediction for the Inverse Weibull Distribution. *Communications in Statistics—Theory & Methods*, **23**, 1811-1824. http://dx.doi.org/10.1080/03610929408831356

[10] Panaitescu, E., Popescu, P.G., Cozma, P. and Popa, M. (2010) Bayesian and Non-Bayesian Estimators Using Record Statistics of the Modified-Inverse Weibull Distribution. *Proceedings of the Romanian Academy*, *Series A*, **11**, 224-231.

[11] Cohen, A.C. (1963) Progressively Censored Samples in Life Testing. *Technometrics*, **5**, 327-339. http://dx.doi.org/10.1080/00401706.1963.10490102

[12] Mann, N.R. (1971) Best Linear Invariant Estimation for Weibull Parameters under Progressive Censoring. *Technometrics*, **13**, 521-533. http://dx.doi.org/10.1080/00401706.1971.10488815

[13] Wingo, D.R. (1993) Maximum Likelihood Estimation of Burr XII Distribution Parameters under Type II Censoring. *Microelectronics Reliability*, **33**, 1251-1257. http://dx.doi.org/10.1016/0026-2714(93)90126-J

[14] Balakrishnan, N. and Sandhu, A. (1995) A Simple Simulational Algorithm for Generating Progressive Type-II Censored Samples. *American Statistician*, **49**, 229-230.

[15] Aggarwala, R. and Balakrishnan, N. (1999) Maximum Likelihood Estimation of the Laplace Parameters Based on Progressive Type-II Censored Samples. In: Balakrishnan, N., Ed., *Advances in Methods and Applications of Probability and Statistics*, Gordan and Breach Publisher, Newark.

[16] Balakrishnan, N. and Asgharzadeh, A. (2005) Inference for the Scaled Half-Logistic Distribution Based on Progressively Type II Censored Samples. *Communications in Statistics—Theory and Methods*, **34**, 73-87. http://dx.doi.org/10.1081/STA-200045814

[17] Balakrishnan, N. (2007) Progressive Censoring Methodology: An Appraisal (with Discussion). *TEST*, **16**, 211-259. http://dx.doi.org/10.1007/s11749-007-0061-y

[18] Johnson, N. (1964) Theory and Technique of Variation Research. Elsevier, Amsterdam.

[19] Balasooriya, U. (1995) Failure-Censored Reliability Sampling Plans for the Exponential Distribution. *Journal of Sta-*

tistical Computation and Simulation, **52**, 337-349. http://dx.doi.org/10.1080/00949659508811684

[20] Wu, S. and Huang, S. (2012) Progressively First-Failure Censored Reliability Sampling Plans with Cost Constraint. *Computational Statistics & Data Analysis*, **56**, 2018-2030. http://dx.doi.org/10.1016/j.csda.2011.12.008

[21] Soliman, A., Abd-Ellah, A., Abou-Elheggag, N. and Abd-Elmougod, G. (2012) Estimation of the Parameters of Life for Gompertz Distribution Using Progressive First-Failure Censored Data. *Computational Statistics & Data Analysis*, **56**, 2471-2485. http://dx.doi.org/10.1016/j.csda.2012.01.025

[22] Soliman, A., Abd-Ellah, A., Abou-Elheggag, N. and Modhesh, A. (2012) Estimation from Burr Type XII Distribution Using Progressive First-Failure Censored Data. *Journal of Statistical Computation and Simulation*, **73**, 887-898.

[23] Hong, C., Lee, W. and Wu, J. (2012) Computational Procedure of Performance Assessment of Lifetime Index of Products for the Weibull Distribution with the Progressive First-Failure Censored Sampling Plan. *Journal of Applied Mathematics*, **2012**, Article ID: 717184.

[24] Ahmadi, M.V., Doostparast, M. and Ahmadi, J. (2013) Estimating the Lifetime Performance Index with Weibull Distribution Based on Progressively First-Failure Censoring Scheme. *Journal of Computational and Applied Mathematics*, **239**, 93-102. http://dx.doi.org/10.1016/j.cam.2012.09.006

[25] Balakrishnan, N., Kannan, N., Lin, C.T. and Ng, H.K.T. (2003) Point and Interval Estimation for Gaussian Distribution Based on Progressively Type-II Censored Samples. *IEEE Transactions on Reliability*, **52**, 90-95. http://dx.doi.org/10.1109/TR.2002.805786

[26] Kim, C., Jung, J. and Chung, Y. (2009) Bayesian Estimation for the Exponentiated Weibull Model under Type II Progressive Censoring. *Statistical Papers*, **51**, 375-387.

[27] Gusmao, F.R.S., Ortega, E.M.M. and Cordeiro, G.M. (2009) The Generalized Inverse Weibull Distribution. *Statistical Papers*, **52**, 271-273.

[28] Marusic, M., Markovic, D. and Jukic, D. (2010) Least Squares Fitting the Three-Parameter Inverse Weibull Density. *Mathematical Communications*, **15**, 539-553.

[29] Ng, H.K.T., Chan, P.S. and Balakrishnan, N. (2002) Estimation of Parameters from Progressively Censored Data Using EM Algorithm. *Computational Statistics & Data Analysis*, **39**, 371-386. http://dx.doi.org/10.1016/S0167-9473(01)00091-3

[30] Balakrishnan, N. (1989) Approximate MLE of the Scale Parameter of the Rayleigh Distribution with Censoring. *IEEE Transactions on Reliability*, **38**, 355-357. http://dx.doi.org/10.1109/24.44181

[31] Balakrishnan, N. (1989) Approximate Maximum Likelihood Estimation of the Mean and Standard Deviation of the Normal Distribution Based on Type-II Censored Samples. *Journal of Statistical Computation and Simulation*, **32**, 137-148. http://dx.doi.org/10.1080/00949658908811170

[32] Balakrishnan, N. (1990) Maximum Likelihood Estimation Based on Complete and Type-II Censored Samples in the Logistic Distribution. Marcel Dekker, New York.

[33] Balakrishnan, N. (1990) Approximate Maximum Likelihood Estimation for a Generalized Logistic Distribution. *Journal of Statistical Planning and Inference*, **26**, 221-236. http://dx.doi.org/10.1016/0378-3758(90)90127-G

[34] Balakrishnan, N. (1990) On the Maximum Likelihood Estimation of the Location and Scale Parameters of Exponential Distribution Based on Multiply Type-II Censored Samples. *Journal of Applied Statistics*, **17**, 55-61. http://dx.doi.org/10.1080/757582647

[35] Balakrishnan, N. and Varden, J. (1991) Approximate MLEs for the Location and Scale Parameters of the Extreme Value Distribution with Censoring. *IEEE Transactions on Reliability*, **40**, 146-151. http://dx.doi.org/10.1109/24.87115

[36] Balakrishnan, N. and Aggarwala, R. (2000) Progressive Censoring: Theory, Methods and Applications. Birkhäuser, Boston. http://dx.doi.org/10.1007/978-1-4612-1334-5

[37] Swain, J., Venkatraman, S. and Wilson, J. (1988) Least Squares Estimation of Distribution Functions in Johnson's Translation System. *Journal of Statistical Computation and Simulation*, **29**, 271-297. http://dx.doi.org/10.1080/00949658808811068

[38] Hossain, A. and Zimmer, W. (2003) Comparison of Estimation Methods for the Weibull Parameters: Complete and Censored Samples. *Journal of Statistical Computation and Simulation*, **73**, 145-153. http://dx.doi.org/10.1080/00949650215730

[39] Montanari, G.C. and Cacciari, M. (1988) Progressively Censored Aging Tests on XLPE-Insulated Cable Models. *IEEE Transactions on Electrical Insulation*, **23**, 365-372. http://dx.doi.org/10.1109/14.2376

[40] Kim, C. and Han, K. (2010) Estimation of the Scale Parameter of the Half-Logistic Distribution under Progressively Type-II Censored Sample. *Statistical Papers*, **51**, 375-387. http://dx.doi.org/10.1007/s00362-009-0197-9

Open Journal of Statistics, 2015, 5, 90-95
Published Online February 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.51011

Comparison of Uniform and Kernel Gaussian Weight Matrix in Generalized Spatial Panel Data Model

Tuti Purwaningsih, Erfiani, Anik Djuraidah

Departement of Statistics, Graduate School of Bogor Agricultural University, Bogor, Indonesia
Email: purwaningsiht@yahoo.com

Received 7 February 2015; accepted 25 February 2015; published 28 February 2015

Abstract

Panel data combine cross-section data and time series data. If the cross-section is locations, there is a need to check the correlation among locations. ρ and λ are parameters in generalized spatial model to cover effect of correlation between locations. Value of ρ or λ will influence the goodness of fit model, so it is important to make parameter estimation. The effect of another location is covered by making contiguity matrix until it gets spatial weighted matrix (W). There are some types of W—uniform W, binary W, kernel Gaussian W and some W from real case of economics condition or transportation condition from locations. This study is aimed to compare uniform W and kernel Gaussian W in spatial panel data model using RMSE value. The result of analysis showed that uniform weight had RMSE value less than kernel Gaussian model. Uniform W had stabil value for all the combinations.

Keywords

Component, Uniform Weight, Kernel Gaussian Weight, Generalized Spatial Panel Data Model

1. Introduction

Panel data analysis combines cross-section data and time series data, in sampling when the data are taken from different locations. It's commonly found that the observation value at one location depends on observation value in another location. In the other name, there is spatial correlation between the observations, which is spatial dependence. Spatial dependence in this study is covered by generalized spatial model which is focussed on dependence between locations and errors [1]. If there is spatial influence but not involved in model so error assumption that between observations must be independent will not fulfilled. So the model will be in bad condi-

How to cite this paper: Purwaningsih, T., Erfiani and Djuraidah, A. (2015) Comparison of Uniform and Kernel Gaussian Weight Matrix in Generalized Spatial Panel Data Model. *Open Journal of Statistics*, **5**, 90-95.
http://dx.doi.org/10.4236/ojs.2015.51011

tion, for that need, a model that involves spatial influence in the analysis panel data will be mentioned as Spatial Panel Data Model.

Some recent literature of spatial cross-section data is Spatial Ordinal Logistic Regression by Aidi and Purwaningsih [2], and Geographically Weighted Regression [3]. Some of the recent literature of Spatial Panel Data is forecasting with spatial panel data [3] and spatial panel models [4]. For accomodating spatial dependence in the model, there is spatial weighted matrix (W) that is an important component to calculate the spatial correlation between locations. Spatial parameter in generalized spatial panel data model, is known as ρ or λ. There are some types of W—uniform W, binary W, inverse distance W and some W from real cases of economics condition or transportation condition from the area. This research is aimed to compare uniform W and kernel Gaussian W in generalized spatial panel data model using RMSE value which is obtained from simulation.

2. Literature Review

2.1. Data Panel Analysis

Data used in the panel data modelisa combination of cross section and time-series data. Crossection data is data collected at one time of many units of observation, then time-series data is data collected over time to an observation. If each unit has a number of observations a cross individuals in the same period of time series, it is calleda balanced panel data. Conversely, if each individual unit has a number of observations a cross different period of time series, it is called an unbalanced panel data (unbalanced panel data).

In general, panel data regression model is expressed as follows:

$$y_{it} = \alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta} + u_{it} \quad i = 1,2,\cdots,N \ ; \ t = 1,2,\cdots,T \tag{1}$$

with i is an index for crossection data and *t* is index of time series. α is a constant value, $\boldsymbol{\beta}$ is a vector of size $K \times 1$, with K specifies the number of explanatory variables. Then y_{it} is the response to the individual cross-*i* for all time period stand $\boldsymbol{x}_{it}$ are sized $K \times 1$ vector for observation *i*-th individual cross and all time periods *t* and u_{it} is the residual/error [5].

Residual components of the direction of the regression model in Equation (1) can be defined as follows:

$$u_{it} = \mu_i + \varepsilon_{it} \tag{2}$$

where μ_i is an individual-specific effect that is not observed, and ε_{it} is a remnant of crossection-*i* and time series-*t* [5].

2.2. Spatial Weighted Matrix (*W*)

Spatial weighted matrix is basically a matrix that describes the relationship between regions and obtained by distance or neighbourhood information. Diagonal of the matrix is generally filled with zero value. Since the weighting matrix shows the relationship between the overall observation, the dimension of this matrix is $N \times N$ [6]. There are several approaches that can be done to show the spatial relationship between the location, including the concept of intersection (contiguity). There are three types of intersection, namely Rook Contiguity, Bishinop Contiguity and Queen Contiguity [6].

After determining the spatial weighting matrix to be used, further normalization in the spatial weighting matrix. In general, the matrix used for normalization normalization row (row-normalize). This means that the matrix is transformed so that the sum of each row of the matrix becomes equal to one. There are other alternatives in the normalization of this matrix is to normalize the columns of the matrix so that the sum of each column in the weighting matrix be equal to one. Also, it can also perform normalization by dividing the elements of the weighting matrix with the largest characteristic root of the matrix ([6] [7]).

There are several types of Spatial Weight (W): binary *W*, uniform *W*, inverse distance *W* (non uniform weight) and some *W* from real case of economics condition or transportation condition from the area. Binary weight matrix has values 0 and 1 in off-diagonal entries; uniform weight is determined by the number of sites surrounding a certain site in ℓ-th spatial order; and non-uniform weight gives unequal weight for different sites. The element of the uniform weight matrix is formulated as,

$$W_{ij} = \begin{cases} \dfrac{1}{n_i^{(l)}}, & j \text{ is neighbor of } i \text{ in } l\text{-th order} \\ 0, & \text{others} \end{cases} \tag{3}$$

$n_i^{(l)}$ is the number of neighbor locations with site-*i* in ℓ-th order. The non-uniform weight may become uniform weight when some conditions are met. One method in building non-uniform weight is based on inverse distance. The weight matrix of spatial lag k is based on the inverse weights $1/(1+d_{ij})$ for sites i and j whose Euclidean distance d_{ij} lies within a fixed distance range, and otherwise is weight zero. Kernel Gaussian Weight follow this formulla:

$$w_j(i) = \exp\left[-1/2\left(d_{ij}/b\right)^2\right] \tag{4}$$

with d isdistance between location i and j, then b is *bandwith* which is a parameter for smoothing function.

2.3. Generalized Spatial Panel Data Model

Generalized spatial model expressed in the following equation:

$$y_{it} = \rho \sum_{j=1}^{N} w_{ij} y_{jt} + \mathbf{x}'_{it}\boldsymbol{\beta} + \mu_i + \phi_{it} \quad \text{dengan} \quad \phi_{it} = \lambda \sum_{j=1}^{N} w_{ij}\phi_{it} + \varepsilon_{it} \tag{5}$$

where ρ is spatial autoregressive coefficient, w_{ij} is elements of the spatial weighted matrix which has been normalized (W) and λ is spatial autocorrelation between error [7].

3. Methodology

Data used in this study was gotten from simulation using generalized spatial panel data model as Equation (5) with initiation of some parameter. Simulation was done use R program. The following step is used to generate the spatial data panel which is consist of index *n* and *t*. In dexnindicates the number of locations and indextindicates the number of period in each locations. Here is the proccess:

1) Determining the number of locations to be simulated is $N = 3$, $N = 9$ and $N = 25$.

2) Makes 3 types of map location on step 1.

3) Creating a binary spatial weighted matrix based on the concept of queen contiguity of each type of map locations. In this step, to map the 3 locations it will form a 3 × 3 matrix, 9 locations will form a 9 × 9 matrix and 25 locations form a 25 × 25 matrix.

4) Creating spatial uniform weighted matrix based on the concept of queen contiguity of each type of map locations.

5) Making weighted matrix kernel Gaussian based on the concept of distance. To make this matrix, previously researchers randomize the centroid points of each location. After setting centroid points, then measure the distance between centroids and used it as a reference to build kernel Gaussian *W*. Gaussian kernel *W* as follows:

$$w_j(i) = \exp\left[-\frac{1}{2}\left(d_{ij}/b\right)^2\right] \text{ [3].}$$

6) Specifies the number of time periods to be simulated is $T = 3$, $T = 6$, $T = 12$ and $T = 24$.

7) Generating the data Y and X based on generalized spatial panel data models follows Equation (5).

8) Cronecker multiplication between matrix identtity of time periods and *W*, then get new matrix named *IW*.

9) Multiply matrix *IW* and Y to obtain vector WY.

10) Build a spatial panel data models and get the value of RMSE.

11) Repeat steps 7)-9) until 1000 replications for each combination on types of W, N, T, ρ and λ. Description:

Types of *W*: *W* binary, *W* uniform and Gaussian kernel *W*;

Types of N: 3, 9 and 25 locations;

Types of T: 3, 6, 12 and 36 series;

Types of $\rho = 0.3$, 0.5, 0.8 and $\lambda = 0.3$, 0.5, 0.8.

12) Get the RMSE value for all of 1000 replicationsoh each combination between *W*, N, ρ and λ.

13) Determine the best *W* based on the smallest RMSE for all combinations.

4. Results and Discussions

Simulation generate data for vector *Y* as dependent variable and *X* matrix as independent variable. *Y* and *X* is generate with parameter initiation. After doing simulation, we can get RMSE for each combinations and proccessing it, then we can calculate RMSE for each *W*, *N*, *T*, ρ and λ. Here is the result. With the result in **Table 1** then continued to figure it into graphs in order to look the comparison easily.

Table 1. Value of RMSE resulted from simulation for all the combinations (*W*, *N*, *T*, ρ and λ).

***W* types**	**Location types**	**Periods types**	**Generalized spatial panel data model**			**Average RMSE**	**Average RMSE**
			$\rho = 0.3$, $\lambda = 0.3$	**$\rho = 0.5$, $\lambda = 0.5$**	**$\rho = 0.8$, $\lambda = 0.8$**		
Uniform *W*	*N* = 3	*T* = 3	1.076	1.23	2.06	1.771	1.634
		T = 6	1.223	1.387	2.684		
		T = 12	1.251	1.464	2.957		
		T = 36	1.296	1.524	3.099		
		Average	1.211	1.401	2.7		
	N = 9	*T* = 3	1.293	1.365	1.775	1.578	
		T = 6	1.341	1.401	1.976		
		T = 12	1.357	1.429	2.054		
		T = 36	1.362	1.448	2.139		
		Average	1.338	1.411	1.986		
	N = 25	*T* = 3	1.383	1.433	1.755	1.553	
		T = 6	1.397	1.446	1.812		
		T = 12	1.403	1.467	1.843		
		T = 36	1.407	1.409	1.877		
		Average	1.398	1.439	1.822		
Kernel Gaussian *W*	*N* = 3	*T* = 3	1.137	1.137	1.137	1.748	1.809
		T = 6	1.352	1.352	1.806		
		T = 12	1.405	2.971	2.014		
		T = 36	1.461	3.098	2.11		
		Average	1.339	2.14	1.767		
	N = 9	*T* = 3	2.101	1.115	1.056	1.243	
		T = 6	1.353	1.138	1.097		
		T = 12	1.255	1.15	1.106		
		T = 36	1.261	1.161	1.119		
		Average	1.493	1.141	1.095		
	N = 25	*T* = 3	1.49	1.282	1.168	2.436	
		T = 6	5.705	1.286	1.169		
		T = 12	6.004	1.293	1.177		
		T = 36	6.19	1.294	1.179		
		Average	4.847	1.289	1.173		

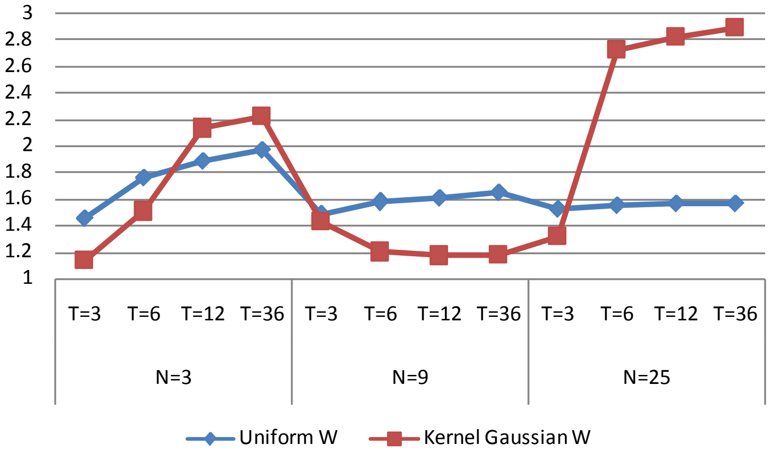

Figure 1. Comparison of RMSE between uniform *W* and kernel Gaussian *W* for all combinations.

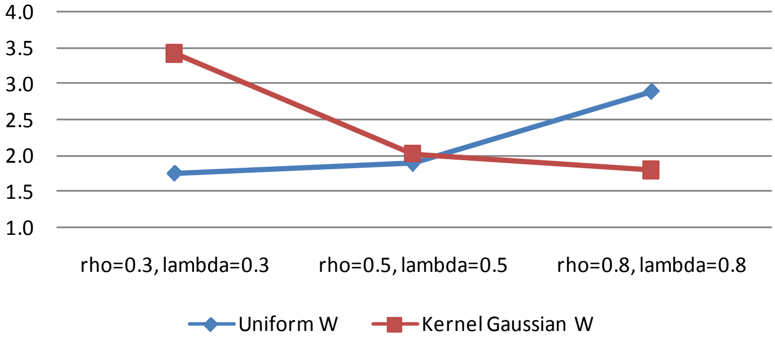

Figure 2. Comparison RMSE each *W* for each parameter.

Based on **Figure 1** can be said that uniform *W* has smaller RMSE than kernel Gaussian *W* for $T = 12$, $T = 36$ on location $N = 3$, then for $T = 6$, 12, 36 on location $N = 25$ and the remaining combinations, kernel Gaussian is higher. If we look the level of stabilization, uniform *W* is better than kernel Gaussian *W*. We can look ats the graph in blue line as uniform *W*, it has value only in range 1, 4 until 2 then kernel Gaussian *W* has range from 1 - 3. So can be concluded that uniform *W* is better than kernel Gaussian *W*.

Based on **Figure 2**, we can look that average RMSE of uniform *W* is smaller in $\rho = 0.3$, $\lambda = 0.3$ and $\rho = 0.5$, $\lambda = 0.5$ while kernel Gaussian *W* is smaller only in $\rho = 0.8$, $\lambda = 0.8$.

5. Conclusion

After looking at the result, it can be concluded that uniform *W* is better than kernel Gaussian *W* almost for all combinations of *N* and *T*. Then uniform *W* is better in ρ and λ in small value until medium (less than 0.5).

Acknowledgements

The first, authors would like to thankful to Allah SWT, my parents, lecturer and all of friends. This research was supported by private funds.

References

[1] Anselin, L., Gallo, J. and Jayet, H. (2008) The Econometrics of Panel Data. Springer, Berlin.

[2] Aidi, M.N. and Purwaningsih, T. (2012) Modelling Spatial Ordinal Logistic Regression and the Principal Component to Predict Poverty Status of Districts in Java Island. *International Journal of Statistics and Application*, **3**, 1-8.

[3] Fotheringham, A.S., Brunsdon, C. and Chartlon, M. (2002) Geographically Weighted Regression, the Analysis of Spa-

tially Varying Relationships. John Wiley and Sons, Ltd., Hoboken.

[4] Elhorst, J.P. (2011) Spatial Panel Models. Regional Science and Urban Econometric.

[5] Baltagi, B.H. (2005) Econometrics Analysis of Panel Data. 3rd Edition, John Wiley and Sons, Ltd., England.

[6] Dubin, R. (2009) Spatial Weights. In: Fotheringham, A.S. and Rogerson, P.A., Eds., *Handbook of Spatial Analysis*, Sage Publications, London. http://dx.doi.org/10.4135/9780857020130.n8

[7] Elhorst, J.P. (2010) Spatial Panel Data Models. In: Fischer, M.M. and Getis, A., Eds., *Handbook of Applied Spatial Analysis*, Springer, New York. http://dx.doi.org/10.1007/978-3-642-03647-7_19

ISSN: 2161-718X Vol. 5, No. 1, February 2015
Scientific Research Publishing
Open Journal of Statistics
www.scirp.org/journal/ojs

Advances in Alzheimer's Disease
Advances in Chemical Engineering and Science
Journal of Biomaterials and Nanobiotechnology
Journal of Electromagnetic Analysis and Applications
Journal of Modern Physics
Journal of Service Science and Management
Open Journal of Mass Spectrometry
Advances in Aging Research
Atmospheric and Climate Sciences
Journal of Electronics Testing and Thermal Control
Journal of Environmental Protection
Journal of Software Engineering and Applications
Journal of Water Resource and Protection
Open Journal of Microphysics
Psychology
JBiSE
Surgical Science
Natural Science
Advances in Applied Sociology
Art and Design Review
Advances in Reproductive Sciences
Creative Education
Health
Low Carbon Economy
Open Journal of Business and Management
Spectral Analysis Reviews
Advances in Bioscience and Biotechnology
Advances in Entomology
Applied Mathematics
Engineering
Intelligent Control and Automation
Modern Economy
Smart Grid and Renewable Energy
Open Journal of Chromatography
Advances in Biological Chemistry
Advances in Historical Studies
American Journal of Plant Sciences
Energy and Power Engineering
Communications, Network and System Sciences
Modeling and Numerical Simulation of Material Science
Open Journal of Accounting
Soft

Scientific
Research
Publishing

Website: http://www.scirp.org
Subscription: sub@scirp.org
Advertisement: service@scirp.org