ROBUST ESTIMATION FOR SPATIAL MODELS AND THE SKILL TEST FOR DISEASE DIAGNOSIS

A Thesis
Presented to
The Academic Faculty

by

Shu-Chuan Lin

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology
December 2008

ROBUST ESTIMATION FOR SPATIAL MODELS AND THE SKILL TEST FOR DISEASE DIAGNOSIS

Approved by:

Professor Paul Kvam,
Committee Chair
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Jye-Chyi Lu, Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Brani Vidakovic
The Wallace H. Coulter Department of Biomedical Engineering and H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Nicoleta Serban
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Yajun Mei
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: August 14, 2008

To myself,

and my family.

ACKNOWLEDGEMENTS

I would like to thank Dr. J.-C. Lu and Prof. Paul Kvam for advising me in my research. I thank my advisors for research funding, the ISyE department for teaching assistantships, and Prof. Francois Sainfort and the Merck Foundation for a three-year fellowship. Thanks also to my committee members:, Dr. Yajun Mei, Dr. Nicoleta Serban, and Dr. Brani Vidakovic.

Many thanks to my family for support, especially my dad, my mom, my sister, my brother, my sister-in-law, my uncle An, and my aunt Sandy.

Many thanks to Fred Zahrn for all his help, and to his parents, James and Cynthia Zahrn.

I appreciate the support of my friends Melike, I-Tang, Doug, Chen-Ju, Richard, Pei, Ying, Yi-Huei, Ya-Mei and many others.

Thank you very much, to any others I forgot to mention.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis focuses on (1) the statistical methodologies for the estimation of spatial data with outliers and (2) classification accuracy of disease diagnosis.

Chapter I, Robust Estimation for Spatial Markov Random Field Models:

Markov Random Field (MRF) models are useful in analyzing spatial lattice data collected from semiconductor device fabrication and printed circuit board manufacturing processes or agricultural field trials. When outliers are present in the data, classical parameter estimation techniques (e.g., least squares) can be inefficient and potentially mislead the analyst. This chapter extends the MRF model to accommodate outliers and proposes robust parameter estimation methods such as the robust M- and RA-estimates. Asymptotic distributions of the estimates with differentiable and non-differentiable robustifying function are derived. Extensive simulation studies explore robustness properties of the proposed methods in situations with various amounts of outliers in different patterns. Also provided are studies of analysis of grid data with and without the edge information. Three data sets taken from the literature illustrate advantages of the methods. The key contributions include proposing robust estimation methodology under the MRF models. Asymptotic properties of the robust estimators are carefully studied under different robustifying functions. We compare the efficiency of the robust estimators under different simulated situations. We analyze the examples by using robust estimators.

Chapter II, Extending the Skill Test for Disease Diagnosis:

For diagnostic tests, we present an extension to the skill plot introduced by Mozer and Briggs (2003). The method is motivated by diagnostic measures for osteoporosis

in a study. By restricting the area under the ROC curve (AUC) according to the skill statistic, we have an improved diagnostic test for practical applications by considering the misclassification costs. We also construct relationships, using the Koziol-Green model and mean-shift model, between the diseased group and the healthy group for improving the skill statistic. Asymptotic properties of the skill statistic are provided. Simulation studies compare the theoretical results and the estimates under various disease rates and misclassification costs. We apply the proposed method in classification of osteoporosis data. The key contributions include proposing the idea using the skillful region for the partial AUC (PAUC). We construct the empirical analog to the skill score and the skill score asymptotic properties. We give reasonable explanation for how to choose the skillful region of false positive rate for the PAUC given the disease rates and loss function. We analyze the osteoporosis example by applying our proposed methods.

CHAPTER I

ROBUST ESTIMATIONS FOR SPATIAL MARKOV RANDOM FIELD MODELS

1.1 Introduction

Modeling spatially correlated lattice data has many practical applications. For example, when analyzing data from industrial semiconductor die grids and electronic circuit layouts, medical and other image processing, epidemiology, astronomy, Earth science, or agricultural field trials, crop science, forestry and soil science, data are often collected in two-dimensional (2D) lattices or grids. Examples of spatial lattice data studies include Papadakis (1937), Atkinson (1969), Beaton and Tukey (1974), Besage (1974), Bartlett (1978), Martin (1979), Kempton and Howes (1981), Martin (1982), Besag and Kempton (1986), Gleeson and Cullis (1987), Zimmerman and Harville (1989, 1991), Chellappa and Jain (1993), Cressie (1993), Basu and Reinsel (1993, 1994), Mesenbrink et al. (1994), Hughes-Oliver et al. (1998), Besag and Higdon (1999), Grau (2000), Ojeda et al. (2002), Grau and Lu (2004), and Rue and Held (2005, Chapter 1) etc.

Two-dimensional time series models are commonly used to fit spatial lattice data. For examples, Bustos et al. (1984), and Basu and Reinsel (1993) explored properties of the spatial unilateral first-order autoregressive moving average (ARMA) models. Martin (1990, 1996) and Cullis and Gleeson (1991) used the separable ARMA models, mainly of first order, for modeling 2D lattice field trial data. Because the 2D time series models, for example AR(1)×AR(1) model, are more restricted (Grau 2000 and Ojeda et al. 2002) and the Markov Random Field (MRF) models describe the spatial data more naturally by neighboring structures (see Eq.(2) for details), this chapter

focuses on MRF models. Rue and Held (2005 Chapter 1) pointed out that MRF is probably known only to researchers in spatial statistics and image analysis. However, there exist other applications far beyond these two areas, e.g., time series analysis, analysis of longitudinal and survival data, semi-parametric statistics etc.

Besag (1974) introduced the MRF models for ecological applications. Usually the sites will be represented by points or regions in the Euclidean plane and will often be subject to a rigid lattice structure. MRF models have been widely used in spatial statistics (Besag and Kooperberg 1995) and Bayesian image analysis (Li 1995), where they are intended to describe interactions between random variables at fixed sites in Euclidean space. One of the common features in spatial experiments is the presence of systematic heterogeneity among the experimental units (Zimmerman and Harville 1991). Typically, the heterogeneity arises because there is correlation among neighboring units.

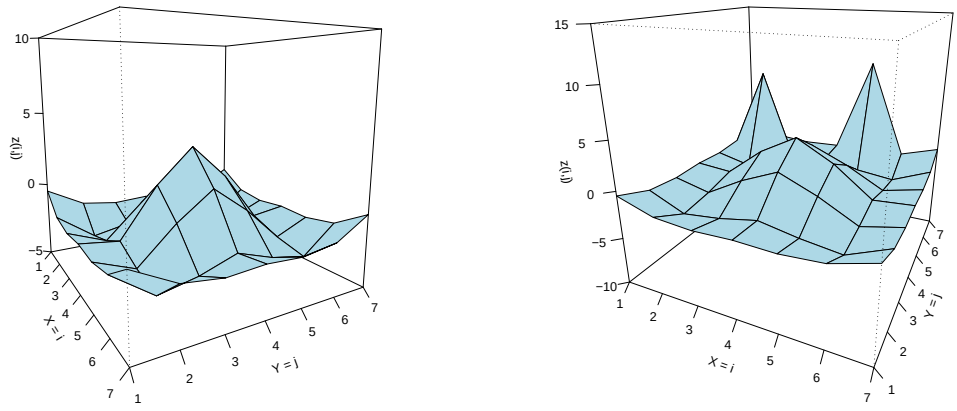
Spatial models usually require that the observations homogeneously follow a normal distribution (see Eq.(1)). This assumption is often violated in practice either due to a few observations that contaminate the normal distribution, or the generating process is not normal at all, e.g., a t-distribution (Little and Robin 2002). The inconsistent observations are usually referred to as outliers, which are caused either by measurement errors, recording mistakes, or by some events which make an observation or set of observations unusual. Outliers can occur randomly or in clusters. For example, in semiconductor manufacturing, procedures often contaminate randomly *dispersed outliers*; also, the problems in soldering process of printed circuit boards might cause a *cluster* of discordant observations (Grau and Lu 2004). Both randomly distributed or clustered outliers may or may not influence surrounding lattice observations depending on the physical process generating those outliers.

The possible physical outlier process has been described by two models: the heavy-tailed distribution model and the mixture distributions model. A main pattern of the

spatial stucture with the heavy-tailed distribution error term means there is a greater chance of outliers (Lange et al. 1989). Those outliers occur randomly. For the mixture distributions model, usually there are two distributions (For example, Fox 1972, Ojeda et al. 2002, Grau and Lu 2004). The two distributions come from two sources: the main pattern of the spatial structure with distribution error term and an additional outliers distribution. The mixture distributions model is practical for real life examples, such as the effect of dust on wafer thickness in semiconductor manufacturing.

Figure 1 shows the impact of those outliers with simulated 2D grid data. They are mock examples from electronics manufacturing processes, where presence of outliers made experimental study of treatment effects very difficult in developing new products (Mesenbrink et al. 1994, Davis et al. 1996, and Hughes-Oliver et al. 1998). Note that the mean effect was removed to clearly see the impact of outliers. Figure 2 presents additional grid data from Lee and Rawlings (1982). This 2D spatial data used five soybean plants growth after twenty days in growth chambers of various sizes, under various temperature conditions. Z_{ij} is the total leaf area of each plant, where i and j represent the position of the plant within the growth chamber. When modeling the spatial correlation with the first-order MRF model, two potential outliers at positions (3,10) and (6,3) could cause problems with estimating the mean function and neighbor parameters.

To overcome this, we need statistical methods that can deal with outlier situations. Lange et al. (1989) proposed the multivariate t maximum likelihood estimation ($t-MLE$) to solve a problem that had longer-than-normal tails. However, $t-MLE$ has two unfavorable properties: the many ways to estimate the degrees of freedom of $t-MLE$, and the requirement of complete neighborhood information. Therefore, the edge data from the grid data is not being used completely. The robust estimators are well known for dealing with outliers. Robust estimators have three



(a) (b)

Figure 1: Plots of (a) Clustered and (b) Randomized Outliers in 7×7 Grids.

main advantages over $t - MLE$. First, the robust estimators do not require complete neighborhood information. Therefore, the robust estimations use all the data. Second, the robust estimators do not have the distribution assumption, unlike likelihood estimation. Third, the robust estimators do not require the assumption of a symmetric distribution. We also establish robust methods for estimating neighbor parameters with mixture outlier distribution models. They are easy to implement and use full information from edge data. Further discussion of the efficiency of robust estimators is in the simulation analysis of Section 1.6.3.

The objective of this chapter is to develop robust methods for estimating parameters in the first-order MRF model with various outlier structures. Section 1.2 presents the underlying model and the techniques for parameter estimation. Section 1.3 develops robust estimation methods. Section 1.4 explores properties of the proposed methods. Section 1.5 presents simulations under four scenarios. Section 1.6 uses three data sets taken from the literature to illustrate the methods. Section 1.7 concludes this chapter and outlines future research.

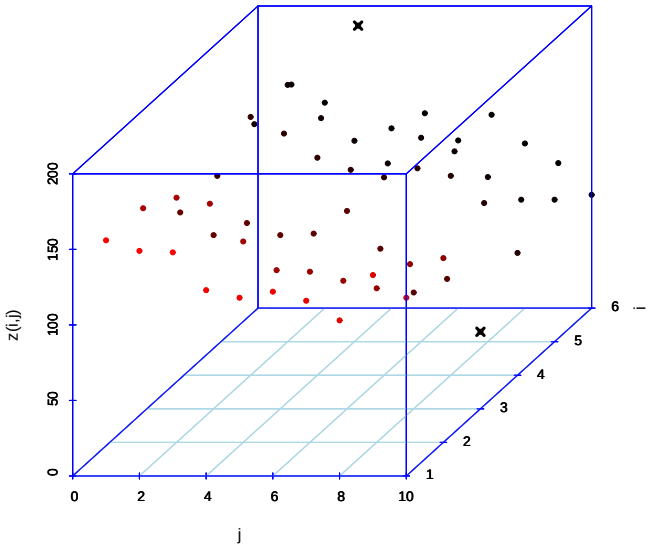


Figure 2: Plot of Lee and Rawlings Data (1982): **x** represents a potential outlier.

1.2 Modeling Spatially Correlated Grid Data

In this section, we present the spatially correlated data model in Section 1.2.1. Two possible outlier models in Section 1.2.2.

1.2.1 Spatial Lattice Data Model

For an $m \times n$ rectangular grid representing a spatial lattice the random variable X measured at site (i, j) is labelled as $X_{i,j}$. Let r represent the index for each of the mn sites. Note that the reasonable grid size will be $m, n \geq 3$. Then, X_r can be decomposed into the following components

$$X_r = \mu_r + \varepsilon_r, r = 1, \cdots, mn, \tag{1}$$

where μ_r represents the mean function, which could be a function of treatment effects or controllable variables, and ε_r represents the stochastic variation in the data. If data points on the grid are independent, then a common assumption is that $\varepsilon_r \sim i.i.d.$

$(0, \sigma^2)$, where *i.i.d.* stands for independent and identically distributed. However, if spatial dependence exists in the data, it can be represented by spatial models. Various autoregressive models have been developed for lattice data, including simultaneous autoregressive models (Whittle 1954) and conditional autoregressive models (Bartlett 1955, 1967, 1968; Besag 1974). This chapter focuses on a conditional autoregressive MRF models (Besag 1974). The first-order MRF model without mean function is given as:

$$X_{i,j} = \alpha_1 X_{i-1,j} + \alpha_2 X_{i+1,j} + \alpha_3 X_{i,j-1} + \alpha_4 X_{i,j+1} + \varepsilon_{ij}, i = 1, \ldots, m, j = 1, \ldots, n, \quad (2)$$

where ε_{ij} are *i.i.d.* with a common variance σ^2. The second-order MRF model considers both the neighbors and the corner data:

$$\begin{aligned} X_{i,j} &= \alpha_1 X_{i-1,j} + \alpha_2 X_{i+1,j} + \alpha_3 X_{i,j-1} + \alpha_4 X_{i,j+1} \\ &\quad + \beta_1 X_{i-1,j-1} + \beta_2 X_{i+1,j+1} + \beta_3 X_{i-1,j+1} + \beta_4 X_{i+1,j-1} + \varepsilon_{ij}. \end{aligned}$$

Since most of the insights for understanding the impact of outliers can be learned from the first-order model, this article studies the simpler first-order MRF model. Note that regardless of the possible different models generating the outlier(s) as discussed below, Eq.(2) is assumed in fitting the data. Least Squares (LS) and robust $M-$ and $RA-$methods are three different techniques studied for estimating parameters in Eq.(2). Note that our estimation methods developed in Section 3 do not require any distribution assumption on the errors. However, in simulation studies a distribution such as normal or t-distribution etc. is needed to generate data.

1.2.2 MRF Outlier Models

There are many ways for setting up the outlier generation models. In Eq.(2) is from a t-distribution with small degree of freedom such as one. See Lange et al. (1989) for an example of replacing the normal error distribution assumption with a t-distribution for making the maximum likelihood estimation more robust. However, this method does not have the control of outlier locations.

In order to study the efficiency of our proposed robust method, we need to set up/control the outlier location. We present the MRF additive outlier model. An example of outlier modeling is the following additive outliers (AOs) model as suggested in Fox (1972), Grau (2000) and Ojeda et al. (2002), and Grau and Lu (2004), which studied robust estimation issues in the spatial AR(1)×AR(1) model. Denote by Z_{ij} the random variables representing the data which might exhibit outlier(s). Here we construct the first-order MRF additive outlier model:

$$\begin{aligned} Z_{ij} &= X_{ij} + \eta_{ij} V_{ij} = \alpha_1 X_{i-1,j} + \alpha_2 X_{i+1,j} + \alpha_3 X_{i,j-1} + \alpha_4 X_{i,j+1} + \varepsilon_{ij} + \eta_{ij} V_{ij} \\ &= \alpha_1 X_{i-1,j} + \alpha_2 X_{i+1,j} + \alpha_3 X_{i,j-1} + \alpha_4 X_{i,j+1} + \varepsilon^*_{ij}, \end{aligned} \tag{3}$$

where X_{ij} can follow any spatial lattice model such as the separable MRF model in Eq.(2). The random variables V_{ij} are independent of X_{ij}, and are $i.i.d.$ with mean zero and variance τ^2. The constant η_{ij} is equal to one when an outlier occurs at site (i, j); it is equal to zero otherwise. $\varepsilon^*_{ij} = \varepsilon_{ij} + \eta_{ij} V_{ij}$. The main focus of this chapter is the robust estimation methods for dependence parameters $\boldsymbol{\theta}^T = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$. Four cases are studied in simulations: (1.) every η_{ij} is zero, i.e., no outlier exists, (2.) one outlier occurs in the middle of the lattice, (3.) outliers exist in 10% of the sites and these sites are all near the center with a cluster outlier pattern, and (4.) same as (3.), but the outliers are uniformly distributed in the lattice.

Since Eq.(3) is only used for generating data to test our robust estimation methods

and the underlying model for modelling the spatial lattice data is Eq.(2), treating η_{ij} as a known constant gives us better control in generating outliers with specific locations and patterns. Extension of the single outlier-location detection procedure given by Chang et al. (1988) from time series to spatial lattices and Neter et al. (1996) outlier diagnostics for regression model are possible. But these require much more work compared to our robust estimation methods due to the possibility of a large number of outliers, of many potential outlier locations, or of differences in the magnitudes of outliers. Thus, an outlier location detection study is beyond the scope of this chapter.

1.3 Robust Estimation

The LSE is commonly used in model parameter estimations in many statistical applications. It is well known that LSE is not resistant to outliers. This section proposes two new outlier resistant methods, robust $M-$ and $RA-$estimators, and compares their performance against the LS method under a spatially correlated model, the MRF model.

All LSE, robust $M-$, and $RA-$estimators could estimate the neighbors parameters under the MRF models. In this section, we construct the general form to estimate the robust estimations. The outlier locations will effect the estimator's efficiency. We present an efficiency comparison for the simulation results of various situations of outliers in Section 1.5. In Section 1.3.3, we show the advantage of using LSE, robust $M-$, and $RA-$estimators by using the full information of data in contrast to maximum likelihood methods.

The first-order MRF model, Eq.(2), can be rewritten in the following matrix format:

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{X}^T = [x_{11}, x_{12}, \ldots, x_{1n}, x_{21}, \ldots, x_{mn}]$, $\boldsymbol{I}$ is a $mn \times mn$ identity matrix and $\boldsymbol{A}$ is a $mn \times mn$ coefficient matrix for neighboring data.

$$
\mathbf{A} = \begin{pmatrix}
0 & \alpha_4 & 0 & \cdots & & 0 & \alpha_2 & 0 & \cdots & & & & & & & 0 \\
\alpha_3 & 0 & \alpha_4 & 0 & \cdots & & 0 & \alpha_2 & 0 & \cdots & & & & & & 0 \\
0 & \alpha_3 & 0 & \alpha_4 & 0 & \cdots & & 0 & \alpha_2 & 0 & \cdots & & & & & 0 \\
 & & \ddots & & & & & & & & & & & & & \\
0 & \cdots & 0 & \alpha_3 & 0 & \alpha_4 & 0 & \cdots & & 0 & \alpha_2 & 0 & \cdots & & & 0 \\
0 & \cdots & & 0 & \alpha_3 & 0 & 0 & & \cdots & & 0 & \alpha_2 & 0 & \cdots & & 0 \\
\alpha_1 & 0 & \cdots & & & 0 & 0 & \alpha_4 & 0 & \cdots & & 0 & \alpha_2 & 0 & \cdots & 0 \\
0 & \alpha_1 & 0 & \cdots & & 0 & \alpha_3 & 0 & \alpha_4 & 0 & \cdots & & 0 & \alpha_2 & 0 & \cdots & 0 \\
 & & & & & \ddots & & & & & & & & & & \\
0 & \cdots & & & 0 & \alpha_1 & 0 & \cdots & 0 & \alpha_3 & 0 & 0 & \cdots & & & 0 & \alpha_2 \\
0 & \cdots & & & & 0 & \alpha_1 & 0 & \cdots & 0 & \alpha_3 & 0 & \alpha_4 & 0 & \cdots & & 0 \\
 & & & & & & & \ddots & & & & & & & & \\
0 & \cdots & & & & & & & & 0 & \alpha_1 & 0 & \cdots & 0 & \alpha_3 & 0 & \alpha_4 \\
0 & \cdots & & & & & & & & & 0 & \alpha_1 & 0 & \cdots & 0 & \alpha_3 & 0
\end{pmatrix}
$$

$$
\boldsymbol{\varepsilon} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{X}.
$$

Let $\boldsymbol{w_{i,j}}^T = (x_{i-1,j}, x_{i+1,j}, x_{i,j-1}, x_{i,j+1})$ be the observation vector and $\boldsymbol{\theta}^T = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ the parameter vector. Therefore, the residual of the first-order MRF outlier model is:

$$
\varepsilon_{ij} = X_{ij} - \boldsymbol{w_{i,j}}^T \boldsymbol{\theta}.
$$

The classical Least Squared estimator LSE is determined in the usual way by

minimizing the sum of the squared errors:

$$min_{\boldsymbol{\theta}} \sum_{i=1}^{m} \sum_{j=1}^{n} (\varepsilon_{i,j})^2. \tag{5}$$

We differentiate the sum of squared of residuals given in Eq.(5) with respect to the parameters. The LSEs, $\hat{\theta}_{LS}$, are determined as the solutions to the following equations:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} x_{i-1,j} \hat{\varepsilon}_{i,j_{LS}} = 0, \sum_{i=1}^{m} \sum_{j=1}^{n} x_{i+1,j} \hat{\varepsilon}_{i,j_{LS}} = 0,$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} x_{i,j-1} \hat{\varepsilon}_{i,j_{LS}} = 0, \sum_{i=1}^{m} \sum_{j=1}^{n} x_{i,j+1} \hat{\varepsilon}_{i,j_{LS}} = 0, \tag{6}$$

where the residual in these equations is:

$$\hat{\varepsilon}_{i,j_{LS}} = x_{ij} - \boldsymbol{w_{i,j}}^T \hat{\boldsymbol{\theta}}_{LS}.$$

1.3.1 Robust M-Estimator

The robust $M-$estimator $\boldsymbol{\theta}_M$ is found by:

$$min_{\boldsymbol{\theta}} \sum_{i=1}^{m} \sum_{j=1}^{n} \rho(\varepsilon_{i,j}), \tag{7}$$

where $\rho(\cdot)$ is a symmetric robustifying loss function. Take the first derivatives of this function with respect to the parameters $\boldsymbol{\theta}$. The robust $M-$estimator $\hat{\boldsymbol{\theta}}_M$ is a solution of the following estimating equations:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} x_{i-1,j} \psi(\hat{\varepsilon}_{i,j_M}) = 0, \sum_{i=1}^{m} \sum_{j=1}^{n} x_{i+1,j} \psi(\hat{\varepsilon}_{i,j_M}) = 0,$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} x_{i,j-1} \psi(\hat{\varepsilon}_{i,j_M}) = 0, \sum_{i=1}^{m} \sum_{j=1}^{n} x_{i,j+1} \psi(\hat{\varepsilon}_{i,j_M}) = 0, \tag{8}$$

where the residual is

$$\hat{\varepsilon}_{i,j_M} = x_{ij} - \mathbf{w_{i,j}}^{\mathbf{T}} \hat{\theta}_{\mathbf{M}}.$$

$\psi(\cdot) = \rho'(\cdot)$ is a bounded function with $t\psi(t) \geq 0$ and usually $\psi'(0) = 1$. The $\psi-$function commonly used in the literature (e.g., Bustos and Yohai 1986) is from the Huber family given by:

$$\psi_{H,c}(r) = \text{sign}(r) \min(\frac{|r|}{\hat{\xi}}, c) * \hat{\xi}, \tag{9}$$

where the value of the tuning constant c depends on how one defines "outliers". A smaller value of c neglects the effect of borderline outliers, and a larger value negates the effect of only the most serious outliers. Experience indicates that appropriate values of c will be between 1.0 and 1.5. The standard deviation (*s.d.*), $\hat{\xi}$, can be computed using a robust estimate of the scale parameter. For example, the following estimator is commonly used in the one-dimensional time series robust estimation literature, e.g., Denby and Martin (1979) and Bustos and Yohai (1986):

$$\hat{\xi} = med(|\hat{\varepsilon}_{i,j_M}|)/0.6745, i = 1, \ldots, m, j = 1, \ldots, n. \tag{10}$$

1.3.2 Robust *RA*-Estimator

The robust Residual Autocovariance estimator (robust RA-estimator) is formed by replacing the above equation content by their robustifing functions:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} \phi(x_{i-1,j}, \hat{\varepsilon}_{i,j_{RA}}) = 0, \sum_{i=1}^{m}\sum_{j=1}^{n} \phi(x_{i+1,j}, \hat{\varepsilon}_{i,j_{RA}}) = 0,$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} \phi(x_{i,j-1}, \hat{\varepsilon}_{i,j_{RA}}) = 0, \sum_{i=1}^{m}\sum_{j=1}^{n} \phi(x_{i,j+1}, \hat{\varepsilon}_{i,j_{RA}}) = 0, \tag{11}$$

where one commonly used example of a robustifying function $\phi : R^2 \rightarrow R$ is the Mallows function, $\phi(\mu, \nu) = \psi(\mu)\psi(\nu)$. Here, $\psi : R \rightarrow R$ is the Huber function

defined in Eq.(9). The residual is:

$$\hat{\varepsilon}_{i,j_{RA}} = x_{ij} - \boldsymbol{w_{i,j}}^T \hat{\boldsymbol{\theta}}_{\boldsymbol{RA}}. \tag{12}$$

Focusing on the Mallows function, computation of the robust RA-estimator, $\hat{\theta}_{RA}$, can be done using the least squares procedure iteratively. Since the LSE and robust M-estimator are obtained from solving similar equations like Eq.(12), where $\phi(\mu, \nu) = \mu\nu$ in the LSE, and $\phi(\mu, \nu) = \mu\psi(\nu)$ in the M-estimator, only the more general RA-estimator is needed for studying the large sample properties of the three procedures. Notice that the Huber ψ-family in Eq.(9) contains an absolute value function, which is non-differentiable. Thus, deriving the asymptotic properties require careful treatment as discussed in Section 1.4.

1.3.3 Maximum Likelihood Methods

Under the assumption of the distribution, the maximum likelihood (ML) methods are other commonly used methods in model parameter estimations in many statistical applications. It is well known that ML estimators (and LSE) are not resistant to outliers.

Lange et al. (1989) proposed the $t-ML$ estimators which are flexible in situations with outliers because of the heavy-tailed property of t-distribution. However, the ML methods need full information of the neighborhood data. Thus, this case is more conservative because it utilizes only part of the data. The edge part of Eq.(2) (X_{ij} with $i = 1$, and m, $j = 1$ and n) is not included in the estimation procedure, because it misses some of the first-order neighbors:

$$X_{i,j} = \alpha_1 X_{i-1,j} + \alpha_2 X_{i+1,j} + \alpha_3 X_{i,j-1} + \alpha_4 X_{i,j+1} + \varepsilon_{ij}, i = 2, \ldots, m-1, j = 2, \ldots, n-1,$$

In the small grid size will lose more information than in a larger grid size. See Figure 3 for details. In a 7×7 grid, it loses about 49% information, (7×7-5×5)/(7×7)×100%.

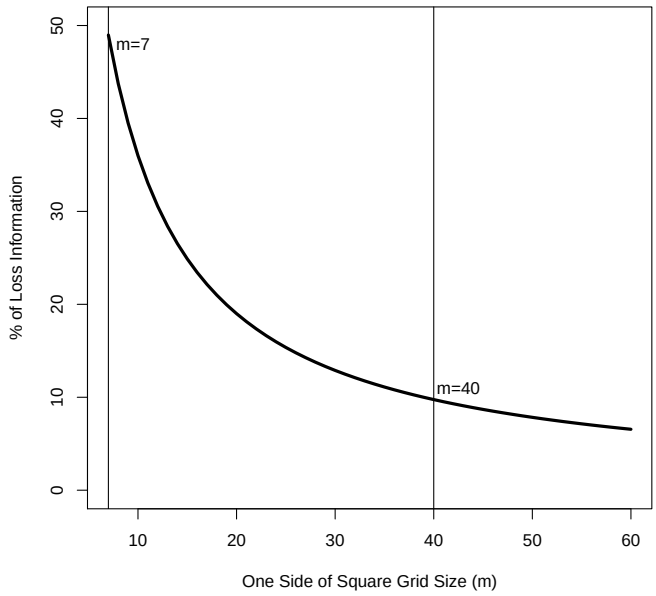


Figure 3: Plot of Percentage Loss Information of the Squared Grid Sizes.

When the grid-size is large (e.g., 40×40), it loss about 10 % information and the impact of discarding edge-data is limited. We compare robust $M-$ and $RA-$estimators with the LS method, NML method and $t-ML$ method when the model is a spatial MRF model with additive outliers and the randomly occurring outlier situation with complete neighborhood simulation study, Section 1.5.3.

Both normal distribution and $t-$ distribution are symmetric. The performance of NML method and $t-ML$ will be better under the condition of the symmetric distribution. LSE, robust $M-$estimators, and robust $RA-$estimators did not assume the symmetric distribution. This is another advantage of using the statistics: LSE, robust $M-$estimators, robust $RA-$estimators than $NMLE$ and $t-MLE$. Therefore, a skew distribution, for example Lognormal distribution , simulations are studied in Section 1.5.2.3 and 1.5.3.3.

1.4 Asymptotic Properties of the Robust RA-Estimator

Because the first-order MRF model has 4 nearest neighbors, they have some correlative structure. Therefore, the asymptotic properties of the nearest neighbors will be multi-dimensional. Asymptotic properties of the robust RA-estimators can be derived in a manner similar to a one-dimensional robust M-estimator (Serfling 1980) and extended to a multivariate central limit theorem (Eicker 1966). The general form of the asymptotic properties of the robust RA-estimator is in Section 1.4.1, if the robust function is (1) continuously differentiable or (2) monotone, bounded and continuous.

When the robust ψ-function is (1) not continuously differentiable and (2) not monotone, bounded and continuous, then the theorem in Section 1.4.1 can not be directly used. We give a solution for this situation in Section 1.4.2.

In Section 1.4.3, we discuss the identical assumption of the first-order MRF model.

1.4.1 General Asymptotic Properties of the Robust RA-Estimator

Let $\boldsymbol{\theta}$ be the true vector parameter. When ϕ and ψ are bounded and continuously differentiable functions such as Mallow's ϕ function and Tukey's bisquare ($\psi_{B,c_2}(x) = x(1 - x^2/c_2^2)^2$), the regular asymptotic theory such as Jureckova and Sen (1996; Chapter 5: Asymptotic Representations for M-estimators) is applicable and the following theorem gives a general statement of the asymptotic distribution of the robust RA-estimate.

Theorem 1 In the first-order MRF model, the error terms ($\varepsilon_{i,j}, i = 1, \ldots, m, j = 1, \ldots, n, m > 3, n > 3$) are independent and identically distributed ($i.i.d.$) random variables from a continuous distribution F. The robust RA-estimator uses the Mallows function: $\phi(u, v) = \psi(u)\psi(v)$. The ϕ function is differentiable. $N = mn$. As m and n go to infinity (N goes to infinity)

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{RA} - \theta) \sim \boldsymbol{N}_4[\mathbf{0}, \Sigma_{\boldsymbol{\theta_{RA}}}],$$

where covariance matrix $\Sigma_{\boldsymbol{\theta_{RA}}}$ is in the following proof. The robust RA-estimator, $\hat{\boldsymbol{\theta}}_{RA}$, is a consistent estimator of $\boldsymbol{\theta}$.

In order to prove this Theorem, one lemma will be proved on the asymptotic properties of a robust RA-estimator. Then, we will extend Lemma 1 to asymptotic properties of multivariate robust RA-estimators.

Lemma 1: In the first-order MRF model, the error terms $(\varepsilon_{i,j}, i = 1, \ldots, m, j = 1, \ldots, n, m > 3, n > 3)$ are independent and identical distribution $(i.i.d.)$ (F). F is a continuous distribution. The robust RA-estimator with the Mallows function: $\phi(u, v) = \psi(u)\psi(v)$. The ϕ function is differentiable. $N = mn$. As m and n go to infinity, the asymptotic normality of the robust $RA-$estimator, $\hat{\alpha}_{1RA}$, for the parameter α_1, is:

$$\hat{\alpha}_{1RA} \sim N(\alpha_1, N^{-1} Var_{RA}(\alpha_1)),$$

where

$$Var_{RA}(\alpha_1) = -\frac{\int_{-\infty}^{\infty} \phi_{1RA}^2(x, \varepsilon) dF(\varepsilon)}{\left[\frac{d \int_{-\infty}^{\infty} \phi_{1RA}(x,\varepsilon) dF(\varepsilon)}{d\alpha_1}\right]^2}.$$

where $\phi_{1RA}(x, \varepsilon) = \psi(x)\psi(\varepsilon_{RA})$.

Proof of Lemma 1: To prove this lemma, we use the results from Huber (1964) or Serfling (1980, page 249) of the asymptotic normality of the robust M-estimator. Then we apply the results to the robust $RA-$estimator.

We have a sample $Y_1, Y_2, \ldots, Y_n$ from a distribution F. The domain of distribution F is $(-\infty, \infty)$. The parameter of interest is θ.

$$\min_{\theta} \sum_{i=1}^{n} \rho(Y_i),$$

where $\rho(\cdot)$ is a symmetric robustifying loss function. Take the first derivative of this function with respect to the parameter θ. $\psi(\cdot) = \rho'(\cdot)$ is a bounded function with

$y\psi(y) \geq 0$ and usually $\psi'(0) = 1$.

θ may be estimated by the robust $M-$estimator, θ_M, by: $\sum_{i=1}^{n} \psi(Y_i) = 0$.

Huber (1964) and Serfling (1980) proved, under suitable regularity conditions, and for ψ function is differentiable, the asymptotic normality of the robust $M-$estimator.

$$\hat{\theta}_M \sim N(\theta, n^{-1} Var_M(\theta)),$$

where typically

$$Var_M(\theta) = -\frac{\int_{-\infty}^{\infty} \psi_M^2(y) dF(y)}{\left[\frac{d \int_{-\infty}^{\infty} \psi_M(y) dF(y)}{d\theta}\right]^2}$$

where d is the first derivative.

Apply the asymptotic normality of Huber (1964) or Serfling (1980) to the first-order MRF model. ε_{ij} are independent and identical distribution ($i.i.d.$) from a distribution F, where $i = 1, \ldots, m, j = 1, \ldots, n$. $\boldsymbol{\theta}^T = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$. For a univariate parameter α_1, the robust $M-$estimate, $\hat{\alpha}_{1M}$, is a solution of the following estimating equation:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} x_{i-1,j} \psi(\hat{\varepsilon}_{i,j_M}) = \sum_{i=1}^{m} \sum_{j=1}^{n} \psi_{1M}(\hat{\varepsilon}_{ij}) = 0,$$

where, $\psi_{1M}(\hat{\varepsilon}_{ij}) = x_{i-1,j} \psi(\hat{\varepsilon}_{i,j_M})$.

If ψ function is differentiable, we prove the asymptotic normality of the Huber robust $M-$estimator for α_1:

$$\hat{\alpha}_{1M} \sim N(\alpha_1, N^{-1} Var_M(\alpha_1)).$$

where

$$Var_M(\alpha_1) = -\frac{\int_{-\infty}^{\infty} \psi_{1M}^2(\varepsilon) dF(\varepsilon)}{\left[\frac{\partial \int_{-\infty}^{\infty} \psi_{1M}(\varepsilon) dF(\varepsilon)}{\partial \alpha_1}\right]^2}$$

and ∂ is the first partial derivative.

The robust RA-estimator with the Mallows function $\phi(u,v) = \psi(u)\psi(v)$ is a special case for the robust $M-$estimator. Assume that the ϕ function is differentiable. $N = mn$. As m and n go to infinity, the asymptotic normality of the robust $RA-$estimator, $\hat{\alpha}_{1RA}$, for the parameter α_1 is:

$$\hat{\alpha}_{1RA} \sim N(\alpha_1, N^{-1}Var_{RA}(\alpha_1)).$$

where

$$Var_{RA}(\alpha_1) = -\frac{\int_{-\infty}^{\infty} \phi_{1RA}^2(x,\varepsilon)dF(\varepsilon)}{\left[\frac{\partial \int_{-\infty}^{\infty} \phi_{1RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_1}\right]^2}.$$

where $\phi_{1RA}(x,\varepsilon) = \psi(x)\psi(\varepsilon_{RA})$. Lemma 1 is proved.

Proof of Theorem 1: From Lemma 1, we have the asymptotic normality of a univariate robust RA-estimator. Here we will extend the results of the Lemma 1 to a multivariate central limit theorem (CLT) (Eicker 1966).

Eicker (1966) proved the multivariate CLT for q-dimensional linear forms of q-vectors with covariance structure. Let F be a distribution function of a random variable with zero mean and positive, finite variance. Assume the sample size is n. The $\boldsymbol{x} = (x_1, \ldots, x_q)^T$ is the random variables. If the following three conditions are simultaneously satisfied:

(I) $\max_j a_j^T(n)(A_n A_n^{-1})a_j(n) \longrightarrow 0$, where $j = 1, \ldots, q$, and T denotes the transpose.
$A_n = (a_1(n), a_2(n), \ldots, a_q(n))$. A_n is a $q \times n$ matrix. $Min_n \texttt{rank}(A_n) = q$.

(II) $\sup_{\boldsymbol{x}} \int_{|x_i|>c} x_i^2 dF(x_i) \longrightarrow 0$ as $c \longrightarrow \infty$.

(III) $\inf_{\boldsymbol{x}} \int x_i^2 dF(x_i) > 0$.

The covariance matrix of the vector A_n is $B_n^2 = A_n^T \Sigma A_n$, and Σ is the diag$(\sigma_1^2, \ldots, \sigma_q^2)$. Then $\boldsymbol{x} = A_n\zeta$ (or $\zeta = B_n^{-1}\boldsymbol{x}$, and B_n^{-1} is the well-defined inverse), ζ converges in distribution to $N_q(0, \Sigma)$.

There are 4 nearest neighbor parameters, $\boldsymbol{\theta}^T = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, in the first-order MRF model. Use Lemma 1 and the multivariate CLT. As m and n go to infinity (N goes to infinity)

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{RA} - \theta) \sim \boldsymbol{N}_4[\mathbf{0}, \Sigma_{\boldsymbol{\theta RA}}],$$

where $\Sigma_{\boldsymbol{\theta RA}} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}$

$$\sigma_{11} = Var_{RA}(\alpha_1) = -\frac{\int_{-\infty}^{\infty} \phi_{1RA}^2(x,\varepsilon)dF(\varepsilon)}{\left[\frac{\partial \int_{-\infty}^{\infty} \phi_{1RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_1}\right]^2}.$$

where $\phi_{1RA}(x,\varepsilon) = \psi_1(x)\psi(\varepsilon_{RA})$, $\psi_1(x_{i,j}) = \psi(x_{i-1,j})$, and $\phi_{1RA}(x_{ij}, \hat{\varepsilon}_{ij}) = \psi_1(x_{i,j})\psi(\hat{\varepsilon}_{i,j_{RA}}) = \psi(x_{i-1,j})\psi(\hat{\varepsilon}_{i,j_{RA}})$.

$$\sigma_{12} = \sigma_{21} = COV_{RA}(\alpha_1, \alpha_2) = -\frac{\int_{-\infty}^{\infty} \phi_{1RA}(x,\varepsilon)\phi_{2RA}(x,\varepsilon)dF(\varepsilon)}{\frac{\partial \int_{-\infty}^{\infty} \phi_{1RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_1} \frac{\partial \int_{-\infty}^{\infty} \phi_{2RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_2}}.$$

where $\phi_{2RA}(x,\varepsilon) = \psi_2(x)\psi(\varepsilon_{RA})$, $\psi_2(x_{i,j}) = \psi(x_{i+1,j})$, and $\phi_{2RA}(x_{ij}, \hat{\varepsilon}_{ij}) = \psi_2(x_{i,j})\psi(\hat{\varepsilon}_{i,j_{RA}}) = \psi(x_{i+1,j})\psi(\hat{\varepsilon}_{i,j_{RA}})$.

$$\sigma_{13} = \sigma_{31} = COV_{RA}(\alpha_1, \alpha_3) = -\frac{\int_{-\infty}^{\infty} \phi_{1RA}(x,\varepsilon)\phi_{3RA}(x,\varepsilon)dF(\varepsilon)}{\frac{\partial \int_{-\infty}^{\infty} \phi_{1RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_1} \frac{\partial \int_{-\infty}^{\infty} \phi_{3RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_3}}.$$

where $\phi_{3RA}(x,\varepsilon) = \psi_3(x)\psi(\varepsilon_{RA})$, $\psi_3(x_{i,j}) = \psi(x_{i,j-1})$, and $\phi_{3RA}(x_{ij}, \hat{\varepsilon}_{ij}) = \psi_3(x_{i,j})\psi(\hat{\varepsilon}_{i,j_{RA}}) = \psi(x_{i,j-1})\psi(\hat{\varepsilon}_{i,j_{RA}})$.

$$\sigma_{14} = \sigma_{41} = COV_{RA}(\alpha_1, \alpha_4) = -\frac{\int_{-\infty}^{\infty} \phi_{1RA}(x,\varepsilon)\phi_{4RA}(x,\varepsilon)dF(\varepsilon)}{\frac{\partial \int_{-\infty}^{\infty} \phi_{1RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_1} \frac{\partial \int_{-\infty}^{\infty} \phi_{4RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_4}}.$$

where $\phi_{4RA}(x,\varepsilon) = \psi_4(x)\psi(\varepsilon_{RA})$, $\psi_4(x_{i,j}) = \psi(x_{i,j+1})$, and $\phi_{4RA}(x_{ij}, \hat{\varepsilon}_{ij}) = \psi_4(x_{i,j})\psi(\hat{\varepsilon}_{i,j_{RA}}) = \psi(x_{i,j+1})\psi(\hat{\varepsilon}_{i,j_{RA}})$.

$$\sigma_{22} = Var_{RA}(\alpha_2) = -\frac{\int_{-\infty}^{\infty} \phi_{2RA}^2(x,\varepsilon)dF(\varepsilon)}{\left[\frac{\partial \int_{-\infty}^{\infty} \phi_{2RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_2}\right]^2}.$$

$$\sigma_{23} = \sigma_{32} = COV_{RA}(\alpha_2,\alpha_3) = -\frac{\int_{-\infty}^{\infty} \phi_{2RA}(x,\varepsilon)\phi_{3RA}(x,\varepsilon)dF(\varepsilon)}{\frac{\partial \int_{-\infty}^{\infty} \phi_{2RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_2}\frac{\partial \int_{-\infty}^{\infty} \phi_{3RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_3}}.$$

$$\sigma_{24} = \sigma_{42} = COV_{RA}(\alpha_2,\alpha_4) = -\frac{\int_{-\infty}^{\infty} \phi_{2RA}(x,\varepsilon)\phi_{4RA}(x,\varepsilon)dF(\varepsilon)}{\frac{\partial \int_{-\infty}^{\infty} \phi_{2RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_2}\frac{\partial \int_{-\infty}^{\infty} \phi_{4RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_4}}.$$

$$\sigma_{33} = Var_{RA}(\alpha_3) = -\frac{\int_{-\infty}^{\infty} \phi_{3RA}^2(x,\varepsilon)dF(\varepsilon)}{\left[\frac{\partial \int_{-\infty}^{\infty} \phi_{3RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_3}\right]^2}.$$

$$\sigma_{34} = \sigma_{43} = COV_{RA}(\alpha_2,\alpha_4) = -\frac{\int_{-\infty}^{\infty} \phi_{3RA}(x,\varepsilon)\phi_{4RA}(x,\varepsilon)dF(\varepsilon)}{\frac{\partial \int_{-\infty}^{\infty} \phi_{3RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_3}\frac{\partial \int_{-\infty}^{\infty} \phi_{4RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_4}}.$$

$$\sigma_{44} = Var_{RA}(\alpha_4) = -\frac{\int_{-\infty}^{\infty} \phi_{4RA}^2(x,\varepsilon)dF(\varepsilon)}{\left[\frac{\partial \int_{-\infty}^{\infty} \phi_{4RA}(x,\varepsilon)dF(\varepsilon)}{\partial \alpha_4}\right]^2}.$$

When N goes to infinity, every element of $\Sigma_{\boldsymbol{\theta_{RA}}}/N$ goes to zero. This proves $\hat{\boldsymbol{\theta}}_{RA}$ are consistent estimators of $\boldsymbol{\theta}$. Theorem 1 is proved.

1.4.2 A Special Case of the Asymptotic Properties of the Robust *RA*-Estimator

When the robust ψ-function is not continuously differentiable—as is the case for a Huber ψ-function, where there exist three non-differentiable points (see Eq.(9) for details)—the limiting matrix does not exist. Since the error distribution is continuous and the number of points where the ψ-function is not continuously differentiable is finite, the limiting matrix is still definable. As suggested in the derivation of the asymptotic normality of the ψ-function in Serfling (1980, page 253), one can separate the expectation operation into three parts to satisfy the monotone property for the differentiable components. Therefore, the asymptotic normal distribution given below still holds in Huber's case.

Theorem 2 In the first-order MRF model, the error terms ($\varepsilon_{i,j}, i = 1, \ldots, m, j = 1, \ldots, n, m > 3, n > 3$) are *i.i.d.* random variables from a continuous distribution F. The robust *RA*-estimator with the Mallows function: $\phi(u, v) = \psi(u)\psi(v)$. The ϕ function is monotone or is bounded and continuous. $N = mn$. As m and n go to infinity (N goes to infinity)

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{RA} - \theta) \sim \boldsymbol{N}_4[\mathbf{0}, \Sigma^*_{\boldsymbol{\theta_{RA}}}],$$

where covariance matrix $\Sigma^*_{\boldsymbol{\theta_{RA}}}$ has similar structure of the covariance matrix in Theorem 1. The robust *RA*-estimator, $\hat{\boldsymbol{\theta}}_{RA}$, is a consistent estimator of $\boldsymbol{\theta}$.

Proof of Theorem 2: We start from the asymptotic normal distribution of the univariate robust M-estimator, the univariate robust *RA*-estimator, and then extend to multivariate robust M-estimators.

Recall from Section 1.3.2, our ψ function is a Huber function (Eq.(9)), ψ function is bounded and continuous and monotone. There exist two non-differentiable points. The Huber ψ function is differentiable by separating into three parts: $(-\infty, -c_1), (-c_1, c_1)$, and (c_1, ∞), where $c_1 = c * \xi$, c is the tuning constant of a Huber function and $\xi = med(|\varepsilon_{i,j_{RA}}|)/0.6745$, $i = 1, \ldots, m, j = 1, \ldots, n$.

ε_{ij} are *i.i.d.* random variables from a distribution F, where $i = 1, \ldots, m, j = 1, \ldots, n$. $\boldsymbol{\theta}^T = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$. For α_1, the robust $M-$estimate, $\hat{\alpha}_{1M}$, is a solution of the following estimating equation:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} x_{i-1,j}\psi(\hat{\varepsilon}_{i,j_M}) = \sum_{i=1}^{m}\sum_{j=1}^{n} \psi_{1M}(\hat{\varepsilon}_{ij}) = 0,$$

where, $\psi_{1M}(\hat{\varepsilon}_{ij}) = x_{i-1,j}\psi(\hat{\varepsilon}_{i,j_M})$. Therefore, we prove that the asymptotic normality of the Huber robust $M-$estimator for α_1 is:

$$\hat{\alpha}_{1M} \sim N(\alpha_1, N^{-1}Var_M(\alpha_1)).$$

where

$$Var_M(\alpha_1) = -\frac{\int_{-c_1}^{c_1} \psi_{1M}^2(\varepsilon)dF(\varepsilon) + c_1^2x^2\int_{-\infty}^{-c_1} dF(\varepsilon) + c_1^2x^2\int_{c_1}^{\infty} dF(\varepsilon)}{[\frac{\partial \int_{-c_1}^{c_1} \psi_{1M}(\varepsilon)dF(\varepsilon)}{\partial\alpha_1}]^2}.$$

Compare Eq.(9) and Eq.(12) and conclude that the robust M-estimator is a special case of the robust RA-estimator with the Mallows function $\phi(u,v) = \psi(u)\psi(v)$, where $\psi(\cdot)$ is a Huber function.

For α_1, the robust $RA-$estimate, $\hat{\alpha}_{1RA}$, is a solution of the following estimating equation:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(x_{i-1,j})\psi(\hat{\varepsilon}_{i,j_{RA}}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\phi_{1RA}(\hat{\varepsilon}_{ij}) = 0,$$

where, $\phi_{1RA}(\hat{\varepsilon}_{ij}) = \psi(x_{i-1,j})\psi(x_{i,j}, \hat{\varepsilon}_{i,j_{RA}})$. There exists six non-differentiable points:

$$-c_1c_2, -c_2\varepsilon, -c_1x, c_1x, c_2\varepsilon, c_1c_2,$$

where $c_1 = c*\xi_1$, c is the tuning constant of a Huber function, $\xi_1 = med(|\varepsilon_{i,j_{RA}}|)/0.6745$, $c_2 = c*\xi_2$, $\xi_2 = med(|x_{i-1,j}|)/0.6745$, $i = 1,\ldots,m, j = 1,\ldots,n$, and assume $c_2\varepsilon \geq c_1x$. The ϕ function is differentiable by separating into seven parts: $(-\infty, -c_1c_2)$, $(-c_1c_2, -c_2\varepsilon), (-c_2\varepsilon, -c_1x), (-c_1x, c_1x), (c_1x, c_2\varepsilon), (c_2\varepsilon, c_1c_2)$, and (c_1c_2, ∞).

The asymptotic normality of the Huber robust $RA-$estimator for α_1 is:

$$\hat{\alpha}_{1RA} \sim N(\alpha_1, N^{-1}Var_{RA}(\alpha_1)).$$

where

$$Var_{RA}(\alpha_1) = -\frac{B_{RA}(\alpha_1)}{[\frac{\partial A_{RA}(\alpha_1)}{\partial\alpha_1}]^2} \tag{13}$$

where

$$B_{RA}(\alpha_1) = \int_{-c_1c_2}^{-c_2\varepsilon}\phi_{1RA}^2(x,\varepsilon)dF(\varepsilon) + \int_{-c_2\varepsilon}^{-c_1x}\phi_{1RA}^2(x,\varepsilon)dF(\varepsilon) + \int_{-c_1x}^{c_1x}\phi_{1RA}^2(x,\varepsilon)dF(\varepsilon) +$$

$$\int_{c_1x}^{c_2\varepsilon} \phi_{1RA}^2(x,\varepsilon)dF(\varepsilon) + \int_{c_2\varepsilon}^{c_1c_2} \phi_{1RA}^2(x,\varepsilon)dF(\varepsilon) + c_1^2c_2^2 \int_{-\infty}^{-c_1c_2} dF(\varepsilon) + c_1^2c_2^2 \int_{c_1c_2}^{\infty} dF(\varepsilon)$$

and

$$A_{RA}(\alpha_1) = \int_{-c_1c_2}^{-c_2\varepsilon} \phi_{1RA}(x,\varepsilon)dF(\varepsilon) + \int_{-c_2\varepsilon}^{-c_1x} \phi_{1RA}(x,\varepsilon)dF(\varepsilon) +$$

$$\int_{-c_1x}^{c_1x} \phi_{1RA}(x,\varepsilon)dF(\varepsilon) + \int_{c_1x}^{c_2\varepsilon} \phi_{1RA}(x,\varepsilon)dF(\varepsilon) + \int_{c_2\varepsilon}^{c_1c_2} \phi_{1RA}(x,\varepsilon)dF(\varepsilon).$$

To extend the univariate robust RA-estimator to the multivariate robust M-estimators, we use the same procedure as in the proof of Theorem 1. However this case will be more complicated because of the technique needed to separate the non-differtiable points.

There are 4 nearest neighbor parameters, $\boldsymbol{\theta}^T = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, in the first-order MRF model. Use Lemma 1 and the multivariate CLT. As m and n go to infinity (N goes to infinity)

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{RA} - \theta) \sim \boldsymbol{N}_4[\mathbf{0}, \Sigma^*_{\boldsymbol{\theta_{RA}}}],$$

where $\Sigma^*_{\boldsymbol{\theta_{RA}}} = \begin{bmatrix} \sigma_{11}^* & \sigma_{12}^* & \sigma_{13}^* & \sigma_{14}^* \\ \sigma_{21}^* & \sigma_{22}^* & \sigma_{23}^* & \sigma_{24}^* \\ \sigma_{31}^* & \sigma_{32}^* & \sigma_{33}^* & \sigma_{34}^* \\ \sigma_{41}^* & \sigma_{42}^* & \sigma_{43}^* & \sigma_{44}^* \end{bmatrix}$

Here we show two examples:

$$\sigma_{11}^* = Var_{RA}(\alpha_1) = -\frac{B_{RA}(\alpha_1)}{[\frac{\partial A_{RA}(\alpha_1)}{\partial \alpha_1}]^2}.$$

where $A_{RA}(\alpha_1)$ and $B_{RA}(\alpha_1)$ are defined in Eq.(13).

$$\sigma_{12}^* = \sigma_{21}^* = COV_{RA}(\alpha_1, \alpha_2) = -\frac{B_{RA}(\alpha_1, \alpha_2)}{\frac{\partial A_{RA}(\alpha_1)}{\partial \alpha_1} \frac{\partial A_{RA}(\alpha_2)}{\partial \alpha_2}},$$

where

$$B_{RA}(\alpha_1,\alpha_2)=\int_{-c_1c_2}^{-c_2\varepsilon}\phi_{1RA}(x,\varepsilon)[\int\phi_{2RA}(x,\varepsilon)dF(\varepsilon)]dF(\varepsilon)+$$

$$\int_{-c_2\varepsilon}^{-c_1x}\phi_{1RA}(x,\varepsilon)[\int\phi_{2RA}(x,\varepsilon)dF(\varepsilon)]dF(\varepsilon)+\int_{-c_1x}^{c_1x}\phi_{1RA}(x,\varepsilon)[\int\phi_{2RA}(x,\varepsilon)dF(\varepsilon)]dF(\varepsilon)+$$

$$\int_{c_1x}^{c_2\varepsilon}\phi_{1RA}(x,\varepsilon)[\int\phi_{2RA}(x,\varepsilon)dF(\varepsilon)]dF(\varepsilon)+\int_{c_2\varepsilon}^{c_1c_2}\phi_{1RA}(x,\varepsilon)[\int\phi_{2RA}(x,\varepsilon)dF(\varepsilon)]dF(\varepsilon)+$$

$$c_1c_2\int_{-\infty}^{-c_1c_2}[\int\phi_{2RA}(x,\varepsilon)dF(\varepsilon)]dF(\varepsilon)+c_1c_2\int_{c_1c_2}^{\infty}[\int\phi_{2RA}(x,\varepsilon)dF(\varepsilon)]dF(\varepsilon)$$

where

$$\int\phi_{2RA}(\varepsilon)dF(\varepsilon)=\int_{-c_3c_4}^{-c_4\varepsilon}\phi_{2RA}(x,\varepsilon)dF(\varepsilon)+\int_{-c_4\varepsilon}^{-c_3x}\phi_{2RA}(x,\varepsilon)dF(\varepsilon)+\int_{-c_3x}^{c_3x}\phi_{2RA}(x,\varepsilon)dF(\varepsilon)+$$

$$\int_{c_3x}^{c_4\varepsilon}\phi_{2RA}(x,\varepsilon)dF(\varepsilon)+\int_{c_4\varepsilon}^{c_3c_4}\phi_{2RA}(x,\varepsilon)dF(\varepsilon)+c_3c_4\int_{-\infty}^{-c_3c_4}dF(\varepsilon)+c_3c_4\int_{c_3c_4}^{\infty}dF(\varepsilon)$$

where $c_3=c*\xi_3$, c is the tuning constant of a Huber function, $\xi_3=med(|\varepsilon_{i,j_{RA}}|)/0.6745$, $c_4=c*\xi_4$, $\xi_4=med(|x_{i+1,j}|)/0.6745$, $i=1,\ldots,m, j=1,\ldots,n$, and assume $c_4\varepsilon\geq c_3x$.

$A_{RA}(\alpha_1)$ is defined in Eq.(13). $A_{RA}(\alpha_2)$ and the other elements of $\Sigma^*_{\boldsymbol{\theta_{RA}}}$ could be found by the same trick. Theorem 2 is proved.

1.4.3 Discussion of the Identical Assumption

Huber (1964) and Serfling (1980) derived asymptotic normality under the *i.i.d.* random variables situation. If the outlier model is from a heavy-tailed t-distribution,

the *i.i.d.* assumption holds. However, if the outlier model is from mixture distributions, the identical distribution assumption in Eq.(2) does not hold. Note that the independent assumption still holds. This section discusses procedures to handle the non-identical distribution problem.

Eicker (1963) introduced the central limit theorems for samples that are not from identical distributions. Let the sample $x_1, \ldots, x_n$ be independent but not identically distributed, for example, with different variance $\sigma_1^2, \ldots, \sigma_n^2$. If the following three conditions are simultaneously satisfied:

(I) $\max_{i=1,\ldots,n} x_i^2 / \sum_{i=1}^n x_i^2 \longrightarrow 0$,

(II) $\sup_{\boldsymbol{x}} \int_{|x_i|>c} x_i^2 dF(x_i) \longrightarrow 0$ as $c \longrightarrow \infty$.

(III) $\inf_{\boldsymbol{x}} \int x_i^2 dF(x_i) > 0$.

Let

$$\zeta = \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 \sigma_i^2}.$$

ζ will converge uniformly in x to a Normal distribution as n goes to infinity.

The first-order MRF additive outlier model in Eq.(3) satisfies the three conditions above. The robust estimators are the solutions of the estimating equations in Eq.(8) and Eq.(11). The robust estimators are in the form (family) of a weighted sum over the entire grid, like ζ above. Therefore, Theorem 1 and Theorem 2 hold even though the identical distribution assumption does not hold. The theorems apply in the case of additive outliers (mixture distributions: main structure and the outliers effect).

1.5 Simulation Studies

In the simulation studies, we constructed the first order MRF models with different structure of outliers from three distributions: a heavy-tailed distribution, a normal distribution, and a skew distribution. We compared the estimation quality of the proposed two robust estimators, robust M-estimator and RA-estimator, against the LSE of neighbor parameters $\alpha_1, \alpha_2, \alpha_3$, and α_4 using the complete edge data information (Section 1.5.2).

Section 1.3.3 compare our robust estimators with two alternative, such as ML estimators, multivariate $N - MLE$ and multivariate $t - MLE$, we estimated the neighbor parameters only in the complete neighborhood data. The edge terms are not used (Section 1.5.3). We compared estimators' performance in the simulation results by using the edge data and without the edge data in Section 1.5.4.

1.5.1 Study Setup for Analysis

Monte Carlo simulation studies with 100 replications in each case were conducted for comparing the estimation quality of the proposed two robust estimators, the robust M-estimator and the robust RA-estimator, against the LSE of neighbor parameters $\alpha_1, \alpha_2, \alpha_3$, and α_4. Zero overall mean is considered to simplify the simulation comparisons.

Simulations include various data sizes in square lattices of 7, 10, 15, 25, and 40 nodes to a side. The neighbor parameters' values were fixed as $\alpha_1 = 0.7, \alpha_2 = 0.7, \alpha_3 = 0.4$, and $\alpha_4 = 0.4$. The procedure to generate a grid of data $x_{i,j}$ follows Eq.(2) from residuals in Eq.(4). The distribution of residuals is not restricted to the normal distribution. Any distributions such as the t-distribution or other heavy tail distributions can be used. Section 1.5.2.1 and 1.5.3.1 present the cases for $\varepsilon_{i,j}$ from, $t_{(1)} \cdot \tau$, $t-$distribution with degree of freedom 1 and τ^2 fixed at 36. A $t-$distribution has

heavier tails than a normal distribution and the outliers occur randomly. Therefore, we don't know how many and where outliers occurred.

In Section 1.5.2.2 and 1.5.3.2, we setup outlier patterns to see the estimators' performance. Our studies consider four scenarios with different outlier patterns: (1.) no outliers present; (2.) a single AO, at a fixed point near the center; (3.) 10% of the observations are AO's in a cluster near the center; and (4.) 10% of the observations are AO's, randomly dispersed throughout the grid. These four scenarios are referred to as Case 1 to 4, respectively. For the no outliers case, $\varepsilon_{i,j}$ are from the standard normal distribution. In the case of randomly dispersed AO outliers, the uncontaminated data x_{ij} are generated from the MRF model in Eq.(2). Then, we generate outlier locations randomly from the uniform(0,1) distribution. Then v_{ij} are generated from $N(0, \tau^2)$ with τ^2 fixed at 36. Finally, if a location does not include an outlier: $z_{ij} = x_{ij}$; otherwise, $z_{ij} = x_{ij} + v_{ij}$ (Eq.(3)). In both robust estimation procedures Huber's robustifying function in Eq.(9) with a tuning constant c =1.5 is used. The Mallows function $\phi(\mu, \nu) = \psi(\mu)\psi(\nu)$ is used for the robust RA-estimates.

Because the $t-$distribution and normal distribution are symmetric, we set up a skew distribution to see the effect of estimators in Section 1.5.2.3 and 1.5.3.3. We only consider Case 4, the worst situation with 10% of the observations are AO's, randomly dispersed throughout the grid. Then v_{ij} are generated from a log-normal distribution, $LN(4, 1)$.

1.5.2 Results from Analysis with Edge Data

LSE, $M-$ and RA-estimates do not require full neighborhood information for neighborhood parameters estimation. Therefore, one of the advantages of those three estimators is using more information from the edge of the grid.

1.5.2.1 Evaluation of Results of a $t-$distribution from Edge Data

Tables 1-3 summarize the simulation results of mean and mean squared error (MSE) of LSE, robust $M-$ and $RA-$estimates from 7×7, 10×10 , and 25×25 grids under the situation that error terms came from a $t-$distribution with one degree of freedom. The tables also present the proportion of bias (Bias). The Bias is defined as, $(\alpha_i - \hat{\alpha_i})/\alpha_i, i = 1, 2, 3, 4$.

In the smaller grid size, 7×7 grid, when the outliers are present randomly from a $t-$distribution with one degree of freedom, robust estimates have smaller bias and MSE than LSE (see Table 1). The proportion of bias of LSE is about 0.0006 to 0.08. The proportion of bias of robust M-estimate and RA-estimates are about 0.001 to 0.02. The bias of the three estimates are small. However, robust estimates have smaller MSE compare to LSE. The estimates of α_3 and α_4 have larger bias and MSE than the estimates of α_1 and α_2 . This could be due to the smaller parameter values of α_3 and α_4 making them more difficult to estimate. Tables 2-3 indicate robust estimators are better than LSE with much smaller MSE and bias. In 10×10 grid, the proportion of bias of LSE is about 0.02 to 0.045. The proportion of bias of robust M-estimate and RA-estimates are about 0.002 to 0.009. The robust estimates have smaller MSE than LSE's. For a 25×25 grid, the robust estimates have almost 0 bias, and the robust estimates have smaller MSE compare to LSE.

1.5.2.2 Evaluation of Results of Normal Distribution from the Edge Data

Tables 4-7 and 8-11 summarize the results from 7×7 and 40×40 grids and four outlier configurations discussed in Section 1.5.1. In the situation with smaller grid size, e.g., 7×7 grid, when there is no outlier present, robust estimates have similar bias and MSE compared with LSE (see Table 4). The three methods are compatible for the other three cases. Tables 5-7 indicate robust estimators are better than LSE with

Table 1: Estimates of $\alpha_1 = 0.7, \alpha_2 = 0.7, \alpha_3 = 0.4$, $\alpha_4 = 0.4$, under a $t-$Distribution (7×7 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6986	0.00200	0.00024	0.6996	0.00057	0.00037
M-est.	0.7008	-0.00114	0.00004	0.6987	0.00186	0.00008
RA-est.	0.7012	-0.00171	0.00006	0.6986	0.00200	0.00009
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.3672	0.08200	0.00237	0.367	0.08250	0.00217
M-est.	0.3925	0.01875	0.00035	0.3926	0.01850	0.00030
RA-est.	0.3927	0.01825	0.00044	0.3929	0.01775	0.00037

Table 2: Parameter Estimates under a $t-$Distribution (10×10 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6856	0.02057	0.00054	0.6852	0.02114	0.00047
M-est.	0.6980	0.00286	0.00003	0.6973	0.00386	0.00004
RA-est.	0.6981	0.00271	0.00003	0.6973	0.00386	0.00004
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.3819	0.04525	0.00148	0.3861	0.03475	0.00133
M-est.	0.3966	0.00850	0.00012	0.3974	0.00650	0.00009
RA-est.	0.3965	0.00875	0.00011	0.3975	0.00625	0.00008

Table 3: Parameter Estimates under a $t-$Distribution (25×25 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6954	0.00657	$5.68e^{-5}$	0.6956	0.00629	$5.42e^{-5}$
M-est.	0.6999	0.00014	$2.46e^{-7}$	0.6999	0.00014	$2.32e^{-7}$
RA-est.	0.6999	0.00014	$2.77e^{-7}$	0.6999	0.00014	$2.42e^{-7}$
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.3924	0.01900	$1.71e^{-4}$	0.3922	0.01950	$1.82e^{-4}$
M-est.	0.3999	0.00025	$2.57e^{-7}$	0.3998	0.00050	$5.27e^{-7}$
RA-est.	0.3999	0.00025	$3.26e^{-7}$	0.3998	0.00050	$5.23e^{-7}$

Table 4: Estimates of $\alpha_1 = 0.7, \alpha_2 = 0.7, \alpha_3 = 0.4$, $\alpha_4 = 0.4$, in Case 1 (7×7 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6994	0.00086	0.00035	0.7002	-0.00029	0.00031
M-est.	0.6995	0.00071	0.00032	0.7005	-0.00071	0.00029
RA-est.	0.6997	0.00043	0.00028	0.7013	-0.00186	0.00030
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.3705	0.07375	0.00208	0.3685	0.07875	0.00229
M-est.	0.3715	0.07125	0.00209	0.3693	0.07675	0.00229
RA-est.	0.3717	0.07075	0.00205	0.3692	0.07700	0.00236

Table 5: Parameter Estimates in Case 2 (7×7 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6792	0.02971	0.00231	0.6812	0.02686	0.00247
M-est.	0.6898	0.01457	0.00070	0.6908	0.01314	0.00091
RA-est.	0.6904	0.01371	0.00066	0.6917	0.01186	0.00077
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.3296	0.17600	0.00957	0.3266	0.18350	0.01001
M-est.	0.3423	0.14425	0.00623	0.3399	0.15025	0.00640
RA-est.	0.3377	0.15575	0.00683	0.3352	0.16200	0.00706

much smaller MSE and bias. When there are more outliers (see Tables 6 and 7), the robust RA-estimate is better than robust M-estimator in both bias and MSE.

When the grid size is as large as 40×40, the bias is close to zero unless there are many outliers uniformly distributed throughout the grid (see Table 11 with 10% outliers). In this case, the robust RA method leads to the smallest bias and MSE.

The next two tables report performance of three estimates for two representative parameters $\alpha_1 = 0.7$ and $\alpha_3 = 0.4$ under various grid sizes under the most difficult case, Case 4 with possible outliers in 10% of all observations. Naturally, with increasing grid sizes, both bias and MSE become smaller. The two robust estimation methods perform better than the LSE, and the robust RA-estimate has a slightly smaller bias and MSE than the robust M-estimate does. This seems to be the general

Table 6: Parameter Estimates in Case 3 (7×7 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6570	0.06143	0.00580	0.6598	0.05743	0.00563
M-est.	0.6743	0.03671	0.00349	0.6793	0.02957	0.00288
RA-est.	0.6817	0.02614	0.00143	0.6855	0.02071	0.00106
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.2723	0.31925	0.03193	0.2693	0.32675	0.03252
M-est.	0.3152	0.21200	0.01369	0.3119	0.22025	0.01385
RA-est.	0.3159	0.21025	0.01288	0.3124	0.21900	0.01313

Table 7: Parameter Estimates in Case 4 (7×7 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6555	0.06357	0.00768	0.6566	0.06200	0.00777
M-est.	0.6887	0.01614	0.00341	0.6713	0.04100	0.00345
RA-est.	0.6731	0.03843	0.00257	0.6754	0.03514	0.00263
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.2670	0.33250	0.02863	0.2636	0.34100	0.02923
M-est.	0.2853	0.28675	0.02129	0.2818	0.29550	0.02210
RA-est.	0.2938	0.26550	0.01892	0.2902	0.27450	0.01979

Table 8: Estimates of $\alpha_1 = 0.7, \alpha_2 = 0.7, \alpha_3 = 0.4,\ \alpha_4 = 0.4$, in Case 1 (40×40 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6991	0.00129	$2.25e^{-6}$	0.6991	0.00129	$3.02e^{-6}$
M-est.	0.6991	0.00129	$2.40e^{-6}$	0.6991	0.00129	$2.98e^{-6}$
RA-est.	0.6991	0.00129	$2.35e^{-6}$	0.6991	0.00129	$3.11e^{-6}$
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.3968	0.00800	$1.60e^{-5}$	0.3968	0.00800	$1.69e^{-5}$
M-est.	0.3968	0.00800	$1.61e^{-5}$	0.3967	0.00825	$1.68e^{-5}$
RA-est.	0.3969	0.00775	$1.51e^{-5}$	0.3968	0.00800	$1.59e^{-5}$

Table 9: Parameter Estimates in Case 2 (40×40 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6991	0.00129	$2.27e^{-6}$	0.6991	0.00129	$3.05e^{-6}$
M-est.	0.6991	0.00129	$2.40e^{-6}$	0.6991	0.00129	$3.00e^{-6}$
RA-est.	0.6991	0.00129	$2.36e^{-6}$	0.6991	0.00129	$3.16e^{-6}$
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.3968	0.00800	$1.61e^{-5}$	0.3968	0.00800	$1.69e^{-5}$
M-est.	0.3968	0.00800	$1.61e^{-5}$	0.3967	0.00825	$1.68e^{-5}$
RA-est.	0.3969	0.00775	$1.52e^{-5}$	0.3968	0.00800	$1.60e^{-5}$

Table 10: Parameter Estimates in Case 3 (40×40 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6991	0.00129	$2.19e^{-6}$	0.6991	0.00129	$2.94e^{-6}$
M-est.	0.6991	0.00129	$2.36e^{-6}$	0.6991	0.00129	$2.90e^{-6}$
RA-est.	0.6991	0.00129	$2.31e^{-6}$	0.6991	0.00129	$3.23e^{-6}$
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.3968	0.00800	$1.61e^{-5}$	0.3968	0.00800	$1.69e^{-5}$
M-est.	0.3968	0.00800	$1.63e^{-5}$	0.3967	0.00825	$1.69e^{-5}$
RA-est.	0.3969	0.00775	$1.54e^{-5}$	0.3968	0.00800	$1.62e^{-5}$

Table 11: Parameter Estimates in Case 4 (40×40 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6931	0.00986	0.00008	0.6930	0.01000	0.00008
M-est.	0.6961	0.00557	0.00003	0.6961	0.00557	0.00003
RA-est.	0.6962	0.00543	0.00003	0.6962	0.00543	0.00003
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.3858	0.03550	0.00029	0.3858	0.03550	0.00029
M-est.	0.3894	0.02650	0.00016	0.3894	0.02650	0.00016
RA-est.	0.3898	0.02550	0.00014	0.3894	0.02650	0.00015

Table 12: Estimates of $\alpha_1 = 0.7$ in Case 4.

Estimate	Grid size					
	10×10			15×15		
	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_1$	Bias	MSE
LSE	0.6099	0.12871	0.01337	0.6723	0.03957	0.00127
M-est.	0.6313	0.09814	0.00847	0.6843	0.02243	0.00049
RA-est.	0.6416	0.08343	0.00595	0.6857	0.02043	0.00041
Estimate	Grid size					
	25×25			40×40		
	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_1$	Bias	MSE
LSE	0.6482	0.07400	0.00315	0.6931	0.00986	0.00008
M-est.	0.6717	0.04043	0.00102	0.6961	0.00557	0.00003
RA-est.	0.6743	0.03671	0.00083	0.6962	0.00543	0.00003

comment for all case studies.

1.5.2.3 Evaluation of Results of Skew Distribution from Edge Data

LSE, robust $M-$ and RA-estimates are free form symmetric assumption. Tables 14 shows the results from 10×10 grids in Case 4 outlier configurations in skew distribution. We saw LSE and robust $M-$estimates are about the same from the bias and MSE. The robust $RA-$estimates has better performance in in both bias and MSE than LSE and robust $M-$estimates.

1.5.2.4 Summary for Simulations with Edge Data

Simulation studies with different distributions, we concluded that in the outliers situation robust estimates are better than LSE in less bias and MSE.

The outliers were randomly generated from a $t-$distribution; we didn't know where and how many outliers occurred in the grid. $t-$distribution is also a heavy-tailed distribution. From the simulations studies, in the smaller grid size, robust estimates have smaller bias and MSE than LSE. In the large grid size, the robust

Table 13: Estimates of $\alpha_3 = 0.4$ in Case 4.

Estimate	Grid size					
	10×10			15×15		
	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_3$	Bias	MSE
LSE	0.2810	0.29750	0.02721	0.3460	0.13500	0.00571
M-est.	0.2993	0.25175	0.01934	0.3562	0.10950	0.00412
RA-est.	0.3099	0.22525	0.01601	0.3586	0.10350	0.00372
Estimate	Grid size					
	25×25			40×40		
	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_3$	Bias	MSE
LSE	0.3679	0.08025	0.00228	0.3858	0.03550	0.00029
M-est.	0.3749	0.06275	0.00143	0.3894	0.02650	0.00016
RA-est.	0.3771	0.05725	0.00121	0.3898	0.02550	0.00014

Table 14: Estimates of $\alpha_1 = 0.7, \alpha_2 = 0.7, \alpha_3 = 0.4$, $\alpha_4 = 0.4$, in Case 4 under a Skew Distribution (10×10 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.1147	0.83614	0.35594	0.1138	0.83743	0.35690
M-est.	0.1121	0.83986	0.36715	0.1192	0.82971	0.35708
RA-est.	0.4547	0.35043	0.07451	0.4585	0.34500	0.07318
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.0257	0.93575	0.14438	0.0255	0.93625	0.14490
M-est.	0.0003	0.99925	0.16083	0.0027	0.99325	0.15921
RA-est.	0.0641	0.83975	0.12138	0.0673	0.83175	0.11978

estimates have almost 0 bias, and the robust estimates have very small MSE compare to LSE.

When there are no outliers present, robust estimates have similar bias and MSE compared with LSE. In additive outliers MRF models, when there are more outliers, the robust RA-estimate is better than the robust M-estimator in both bias and MSE. In the simulation setting for many outliers uniformly distributed throughout the grid has higher bias and MSE compare to other three cases. In this case, the robust RA method leads to the smallest bias and MSE. Naturally, with increasing grid sizes, both bias and MSE become smaller. The two robust estimation methods perform better than the LSE, and the robust RA-estimate has a slightly smaller bias and MSE than the robust M-estimate does. Again, this seems to be the general comment for all case studies.

LSE, robust $M-$ and RA-estimates didn't require symmetric assumption in estimation. Under the most difficult case with possible outliers in 10% of all observations, LSE and robust $M-$estimates are about the same from the bias and MSE. The robust $RA-$estimates have better performance in both bias and MSE than LSE and robust $M-$estimates.

1.5.3 Results from Analysis without Edge Data

Multivariate $t-MLE$ is flexible in the situations with outliers because of its heavy-tailed property. However, ML methods need the full information of the neighborhood data. In order to compare LSE, robust $M-$ and RA-estimates in the same situation, we restrict to use partial data: only use the full neighborhood information to estimate neighbor parameters. The first-order MRF with the full information from neighborhood data is defined as:

$$X_{i,j} = \alpha_1 X_{i-1,j} + \alpha_2 X_{i+1,j} + \alpha_3 X_{i,j-1} + \alpha_4 X_{i,j+1} + \varepsilon_{ij}, i = 2, \ldots, m-1, j = 2, \ldots, n-1,$$

Here we use $t_{(1)}MLE$ with one degree of freedom, and $NMLE$ to estimate neighborhood parameters.

1.5.3.1 Evaluation of Results of $t-$distribution without Edge Data

Tables 15-16 summarize the simulation results of mean and mean square error (MSE) of LSE, robust $M-$, robust $RA-$estimates, $t_{(1)}MLE$, and $NMLE$ only use the full neighborhood information from 7×7, and 25×25 grids under the situation that error terms came from a $t-$distribution setting in Section 1.5.1.

In the smaller grid size, 7×7 grid (Tables 15), when the outliers present randomly from $t-$distribution, robust $RA-$estimates have a little smaller bias and MSE than even the true model $t_{(1)}MLE$. In decreasing order of estimates of bias and MSE, we have: $NMLE$, LSE, $t_{(1)}MLE$, robust $M-$estimates and $RA-$estimates. For large grid size (Tables 16), $t_{(1)}MLE$, robust $M-$ and $RA-$estimates have similar bias and MSE. Three estimates have almost 0 bias, and the robust estimates have very small MSE compare to LSE.

1.5.3.2 Evaluation of Results of Normal Distribution without Edge Data

Tables 17-20 summarize the simulation results of mean and MSE of LSE, robust $M-$, robust $RA-$estimates, $t_{(1)}MLE$, and $NMLE$ only use the full neighborhood information from 7×7, 10×10, 25×25, and 40×40 under the most difficult situation, case 4, with possible outliers in 10% randomly throughout of all observations.

In the smaller grid size, 7×7 grid (Tables 17), robust $M-$ and $RA-$estimates have a little smaller bias and MSE than $t_{(1)}MLE$. In decreasing order of estimates of bias and MSE, we have: $NMLE$, LSE, $t_{(1)}MLE$, robust $M-$estimates

Table 15: Estimates of $\alpha_1 = 0.7, \alpha_2 = 0.7, \alpha_3 = 0.4$, $\alpha_4 = 0.4$, under a $t-$Distribution without Edge Data (10×10 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6852	0.02114	0.00084	0.6852	0.02114	0.00073
M-est.	0.6982	0.00257	0.00004	0.6968	0.00457	0.00006
RA-est.	0.6983	0.00243	0.00003	0.6969	0.00443	0.00006
$NMLE$	0.6935	0.00929	0.00041	0.6892	0.01543	0.00068
$t_{(1)}MLE$	0.6993	0.00100	0.00004	0.6978	0.00314	0.00005
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.3780	0.05500	0.00217	0.3817	0.04575	0.00201
M-est.	0.3965	0.00875	0.00009	0.3975	0.00625	0.00006
RA-est.	0.3965	0.00875	0.00009	0.3975	0.00625	0.00006
$NMLE$	0.3841	0.03975	0.00068	0.3889	0.02775	0.00112
$t_{(1)}MLE$	0.3972	0.00700	0.00005	0.3982	0.00450	0.00010

Table 16: Estimates of $\alpha_1 = 0.7, \alpha_2 = 0.7, \alpha_3 = 0.4$, $\alpha_4 = 0.4$, under a $t-$Distribution without Edge Data (25×25 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6951	0.00700	$7.61e^{-5}$	0.6950	0.00714	$7.09e^{-5}$
M-est.	0.6999	0.00014	$3.20e^{-7}$	0.6998	0.00029	$3.50e^{-7}$
RA-est.	0.6999	0.00014	$3.55e^{-7}$	0.6998	0.00029	$3.52e^{-7}$
$NMLE$	0.6959	0.00586	$1.28e^{-4}$	0.6981	0.00271	$5.94e^{-5}$
$t_{(1)}MLE$	0.7001	-0.00014	$1.57e^{-7}$	0.6999	0.00014	$2.08e^{-7}$
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.3912	0.02200	$2.37e^{-4}$	0.3911	0.02225	$2.48e^{-4}$
M-est.	0.3999	0.00025	$4.71e^{-7}$	0.3997	0.00075	$7.02e^{-7}$
RA-est.	0.3999	0.00025	$5.59e^{-7}$	0.3997	0.00075	$7.50e^{-7}$
$NMLE$	0.3955	0.01125	$1.21e^{-4}$	0.3964	0.00900	$1.08e^{-4}$
$t_{(1)}MLE$	0.3999	0.00025	$1.39e^{-7}$	0.3999	0.00025	$3.20e^{-7}$

Table 17: Estimates of $\alpha_1 = 0.7, \alpha_2 = 0.7, \alpha_3 = 0.4,\ \alpha_4 = 0.4$, in Case 4 without Edge Data (7×7 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6525	0.06786	0.00895	0.6539	0.06586	0.00997
M-est.	0.6620	0.05429	0.00628	0.6657	0.04900	0.00715
RA-est.	0.6656	0.04914	0.00531	0.6683	0.04529	0.00579
$NMLE$	0.6536	0.06629	0.01613	0.6188	0.11600	0.02489
$t_{(1)}MLE$	0.6698	0.04314	0.00744	0.6645	0.05071	0.00783
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.2598	0.35050	0.03756	0.2536	0.36600	0.03958
M-est.	0.2773	0.30675	0.02966	0.2705	0.32375	0.03197
RA-est.	0.2894	0.27650	0.02623	0.2825	0.29375	0.02844
$NMLE$	0.1889	0.52775	0.08137	0.1867	0.53325	0.08684
$t_{(1)}MLE$	0.2368	0.40800	0.04946	0.2380	0.40500	0.05066

and $RA-$estimates. For larger grid size (Tables 18-20), $t_{(1)}MLE$, has similar bias and MSE. In decreasing order of estimates of bias and MSE, we have: $NMLE$, LSE, robust $M-$estimates, robust $RA-$estimates, and $t_{(1)}MLE$. Therefore, robust estimates have an advantage in the small grid sizes compare to $t_{(1)}MLE$. Naturally, with increasing grid sizes, both bias and MSE of all estimators become smaller.

1.5.3.3 Evaluation of Results of Skew Distribution without Edge Data

Tables 21 shows the simulation results from 10×10 grids in case 4 outlier configuration in skew distribution by using the full neighborhood information. $NMLE$, LSE and $M-$estimates have the similar bias and MSE. The robust $RA-$estimates have better performance in both bias and MSE than the other estimates. The decreasing order of estimates of bias and MSE: $NMLE$, robust $M-$estimates, LSE, $t_{(1)}MLE$, and robust $RA-$estimates. Therefore, under a very skew distribution, robust $RA-$estimates have the best performance.

Table 18: Parameter Estimates in Case 4 without Edge Data (10×10 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.5997	0.14329	0.01823	0.5954	0.14943	0.01804
M-est.	0.6222	0.11114	0.01154	0.6226	0.11057	0.00981
RA-est.	0.6345	0.09357	0.00826	0.6304	0.09943	0.00828
$NMLE$	0.5830	0.16714	0.02739	0.5643	0.19386	0.03370
$t_{(1)}MLE$	0.6530	0.06714	0.00630	0.6492	0.07257	0.00546
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.2692	0.32700	0.03219	0.2680	0.33000	0.03401
M-est.	0.2893	0.27675	0.02273	0.2891	0.27725	0.02414
RA-est.	0.2978	0.25550	0.02016	0.2975	0.25625	0.02153
$NMLE$	0.2650	0.33750	0.04297	0.2578	0.35550	0.04594
$t_{(1)}MLE$	0.3222	0.19450	0.01432	0.3203	0.19925	0.01642

Table 19: Parameter Estimates in Case 4 without Edge Data (25×25 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6456	0.07771	0.00351	0.6457	0.07757	0.00349
M-est.	0.6692	0.04400	0.00123	0.6685	0.04500	0.00128
RA-est.	0.6715	0.04071	0.00103	0.6711	0.04129	0.00106
$NMLE$	0.6454	0.07800	0.00357	0.6454	0.07800	0.00356
$t_{(1)}MLE$	0.6880	0.01714	0.00029	0.6880	0.01714	0.00028
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.3643	0.08925	0.00270	0.3639	0.09025	0.00262
M-est.	0.3716	0.07100	0.00176	0.3703	0.07425	0.00171
RA-est.	0.3738	0.06550	0.00152	0.3724	0.06900	0.00149
$NMLE$	0.3645	0.08875	0.00268	0.3643	0.08925	0.00258
$t_{(1)}MLE$	0.3834	0.04150	0.00075	0.3824	0.04400	0.00070

Table 20: Parameter Estimates in Case 4 without Edge Data (40×40 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.6925	0.01071	0.00009	0.6928	0.01029	0.00009
M-est.	0.6958	0.00600	0.00003	0.6959	0.00586	0.00003
RA-est.	0.6958	0.00600	0.00003	0.6960	0.00571	0.00003
$NMLE$	0.6924	0.01086	0.00010	0.6927	0.01043	0.00010
$t_{(1)}MLE$	0.6980	0.00286	0.00001	0.6980	0.00286	0.00001
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.3854	0.03650	0.00031	0.3854	0.03650	0.00031
M-est.	0.3891	0.02725	0.00017	0.3891	0.02725	0.00017
RA-est.	0.3894	0.02650	0.00016	0.3894	0.02650	0.00016
$NMLE$	0.3845	0.03875	0.00036	0.3845	0.03875	0.00036
$t_{(1)}MLE$	0.3928	0.01800	0.00008	0.3928	0.01800	0.00008

Table 21: Estimates of $\alpha_1 = 0.7, \alpha_2 = 0.7, \alpha_3 = 0.4,\ \alpha_4 = 0.4$, in Case 4 under a Skew Distribution without Edge Data (10×10 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
LSE	0.1228	0.82457	0.35941	0.1097	0.84329	0.35889
M-est.	0.1152	0.83543	0.36694	0.1204	0.82800	0.35702
RA-est.	0.4146	0.40771	0.10615	0.4109	0.41300	0.10931
$NMLE$	0.0624	0.91086	0.41254	0.0594	0.91514	0.41832
$t_{(1)}MLE$	0.2830	0.59571	0.26397	0.2684	0.61657	0.27686
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
LSE	0.0273	0.93175	0.14777	0.0241	0.93975	0.15168
M-est.	-0.0036	1.00900	0.16629	0.0038	0.99050	0.15976
RA-est.	0.0433	0.89175	0.13843	0.0496	0.87600	0.13427
$NMLE$	-0.0083	1.02075	0.17831	-0.0017	1.00425	0.17199
$t_{(1)}MLE$	0.0255	0.93625	0.14989	0.0374	0.90650	0.14203

1.5.3.4 Summary for Simulations without Edge Data

Multivariate $t-MLE$ is flexible with the outliers situations because of its heavy-tailed property. In order to compare robust estimators with $t-MLE$, the estimation procedure used the full information of the neighborhood data. Note that robust estimators and LSE did not have this restriction. Simulation studies with different distributions, we conclude that robust estimates are better than LSE, $NMLE$ in less bias and MSE. In the small sample grid size, robust $RA-$estimators are better than $t-MLE$ in less bias and MSE. In the large sample grid size, the relationship between robust estimators and $t-MLE$ varies.

When simulate the outliers present randomly from a $t-$distribution, in the smaller grid size, robust $RA-$estimates have a little smaller bias and MSE than even $t_{(1)}MLE$. The decreasing order of estimates of bias and MSE: $NMLE$, LSE, $t_{(1)}MLE$, robust $M-$estimates and $RA-$estimates. In the large grid size, $t_{(1)}MLE$, robust $M-$ and $RA-$estimates have similar bias and MSE. Three estimates have almost 0 bias, and the robust estimates have very small MSE compare to LSE.

In additive outliers MRF models, under the most difficult situation with possible outliers in 10% of all observations, robust $M-$ and $RA-$estimates have a little smaller bias and MSE than $t_{(1)}MLE$. In decreasing order of estimates of bias and MSE, we have: $NMLE$, LSE, $t_{(1)}MLE$, robust $M-$estimates and $RA-$estimates. For large grid size, $t_{(1)}MLE$, has similar bias and MSE. The decreasing order of estimates of bias and MSE: $NMLE$, LSE, robust $M-$estimates, robust $RA-$estimates, and $t_{(1)}MLE$. Therefore, robust estimates have an advantage in the small grid sizes compared to $t_{(1)}MLE$. Naturally, with increasing grid sizes, both the bias and MSE of all estimators become smaller.

LSE, robust $M-$ and RA-estimates did not require symmetric assumption in

estimation. Under the most difficult case with possible outliers in 10% of all observations, $NMLE$, LSE and robust $M-$estimates have similar bias and MSE. The robust $RA-$estimates have better performance in both bias and MSE than the other estimates. In decreasing order of estimates of bias and MSE, we have: $NMLE$, robust $M-$estimates, LSE, $t_{(1)}MLE$, and robust $RA-$estimates. Under a very skew distribution, robust $RA-$estimates have the best performance.

1.5.4 Comparison of Simulation Results from 1.5.2 and 1.5.3

Here we study the edge effect in LSE, robust $M-$ and RA-estimates. In the small grid size will loss more information than larger grid size. The robust $RA-$ estimates suffer more from using only neighborhood data compared to LSE and robust $M-$ estimates in Section 1.5.4.1.

The ML methods need the full information of the neighborhood data. In the other hand, LSE, and robust estimates do not have the restriction for the full information of the neighborhood data. We show the advantage of the estimation bias and MSE are smaller than the ML methods in Section 1.5.4.2.

1.5.4.1 Comparison of Simulation Results for LSE, and robust estimates under using partial data and full data

Table 22 shows the simulation results from a 25×25 grid in Case 4 outlier configuration in a heavy-tailed distribution to compare the edge effect. Proportion of Bias increased (BiasIn) is defined as estimation proportion of bias by using full neighborhood data minus estimation proportion of bias by using edge data. Ratio of MSEs (RMSE) is estimator of MSE of using full neighborhood data divided by estimator of MSE of using edge data. Proportion of Bias increased are small. LSE's proportion bias increased about 0.003 to 0.00275. Robust $M-$ and $RA-$ estimates' proportion bias

Table 22: Edge Effect Study of Estimates of $\alpha_1 = 0.7, \alpha_2 = 0.7, \alpha_3 = 0.4,\ \alpha_4 = 0.4$, under a $t-$Distribution (25×25 Grid).

	$\hat{\alpha}_1$		$\hat{\alpha}_2$		$\hat{\alpha}_3$		$\hat{\alpha}_4$	
Estimate	BiasIn	RMSE	BiasIn	RMSE	BiasIn	RMSE	BiasIn	RMSE
LSE	0.00043	1.34	0.00086	1.31	0.00030	1.39	0.00275	1.36
M-est.	0	1.30	0.00014	1.51	0	1.83	0.00025	1.33
RA-est.	0	1.28	0.00014	1.45	0	1.71	0.00025	1.43

increased about 0 to 0.00025. MSE of LSE by using full neighborhood data increase about 30 % to 40 % compare to MSE of LSE by using edge data. MSE of robust estimates by using full neighborhood data increase about 30 % to 80 % compared to MSE of robust estimates by using edge data.

In additive outliers MRF models, under the most difficult situation with possible outliers in 10% of all observations, robust $M-$ and $RA-$estimates have loss more information than LSE. In the smaller grid size (Table 23), proportions of bias increased are small. LSE's proportions of bias increased range from 0.004 to 0.025. The robust $M-$ estimates' proportions of bias increased range from 0.08 to 0.038. Robust $RA-$ estimates' proportions of bias increased range from 0.01 to 0.02. The MSE of LSE by using full neighborhood data increased about 17 % to 35 % compared to the MSE of LSE by using edge data.The MSE of robust estimates by using full neighborhood data increased about 40 % to 120 % compared to MSE of robust estimates by using edge data. For large grid size (Table 24), LSE's proportions of bias increased range from 0.00029 to 0.001. The robust $M-$ and $RA-$ estimates' proportions of bias increased range from 0 to 0.001. The MSE of LSE by using full neighborhood data increased range from 7 % to 13 % compared to the MSE of LSE by using edge data. The MSEs of robust estimates were not affected by using the edge data or full neighborhood data.

Table 23: Edge Effect Study of Estimates of $\alpha_1 = 0.7, \alpha_2 = 0.7, \alpha_3 = 0.4, \alpha_4 = 0.4$, under Case 4 (7×7 Grid).

	$\hat{\alpha}_1$		$\hat{\alpha}_2$		$\hat{\alpha}_3$		$\hat{\alpha}_4$	
Estimate	BiasIn	RMSE	BiasIn	RMSE	BiasIn	RMSE	BiasIn	RMSE
LSE	0.00429	1.17	0.00386	1.28	0.01800	1.31	0.02500	1.35
M-est.	0.03814	1.84	0.00800	2.07	0.02000	1.39	0.02825	1.45
RA-est.	0.01071	2.07	0.01014	2.20	0.01100	1.39	0.01925	1.44

Table 24: Edge Effect Study of Parameter Estimates under Case 4 (40×40 Grid).

	$\hat{\alpha}_1$		$\hat{\alpha}_2$		$\hat{\alpha}_3$		$\hat{\alpha}_4$	
Estimate	BiasIn	RMSE	BiasIn	RMSE	BiasIn	RMSE	BiasIn	RMSE
LSE	0.00086	1.13	0.00029	1.13	0.00100	1.07	0.00100	1.07
M-est.	0.00043	1.00	0.00029	1.00	0.00075	1.06	0.00075	1.06
RA-est.	0.00057	1.00	0.00029	1.00	0.00100	1.14	0	1.07

Tables 25 shows the simulation results from 10×10 grids in case 4 outlier configurations in skew distribution to compare the edge effect. LSE's proportion bias increased about -0.01 to 0.006. The robust $M-$ estimates' proportions of bias increased range from -0.004 to 0.01. The MSEs of LSE and robust $M-$ estimates were not affected by using the edge data or only full neighborhood data. The MSE of robust $RA-$ estimates by using full neighborhood data increased about 10 % to 50 % compared to the MSE of robust estimates by using edge data.

1.5.4.2 Comparison of Simulation Results for the robust RA−estimates using full data versus t − MLE using partial data

From the simulation studies, the robust $RA-$estimator and $t-MLE$ both handle the outliers situation very well. $t-MLE$ worked better when the underlying distribution was a $t-$distribution and also when the grid size was large. In the other cases, the robust $RA-$estimator was superior.

Table 25: Edge Effect Study of Estimates of $\alpha_1 = 0.7, \alpha_2 = 0.7, \alpha_3 = 0.4, \alpha_4 = 0.4$, under a Skew Distribution (10×10 Grid).

	$\hat{\alpha}_1$		$\hat{\alpha}_2$		$\hat{\alpha}_3$		$\hat{\alpha}_4$	
Estimate	BiasIn	RMSE	BiasIn	RMSE	BiasIn	RMSE	BiasIn	RMSE
LSE	-0.01157	1.01	0.00586	1.01	-0.00400	1.02	0.00350	1.05
M-est.	-0.00443	1.00	-0.00171	1.00	0.00975	1.03	-0.00275	1.00
RA-est.	0.05729	1.42	0.06800	1.49	0.05200	1.14	0.04425	1.12

Table 26: Comparison of the Robust $RA-$Estimates under Full Data and $t-MLE$ under Partial Data for a $t-$Distribution (10×10 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
RA-est.$_f$	0.6981	0.00271	0.00003	0.6973	0.00386	0.00004
$t_{(1)}MLE_p$	0.6993	0.00100	0.00004	0.6978	0.00314	0.00005
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
RA-est.$_f$	0.3965	0.00875	0.00011	0.3975	0.00625	0.00008
$t_{(1)}MLE_p$	0.3972	0.00700	0.00005	0.3982	0.00450	0.00010

Therefore, we compare the robust $RA-$estimates under full data versus $t-MLE$ under partial data in different situations. Notation for Table 26–30 is as follows: subscript "f" means using full data and subscript "p" means using partial data (only the full neighborhood data).

In Table 26 and 27 are 10×10 and 25×25 grids under the situation that error terms came from a $t-$distribution with one degree of freedom. In this situation, we do not know where the outliers are. The MSE of the robust $RA-$estimates are a little smaller than $t_{(1)}MLE$ for small grid size 10×10.

For a larger grid size, $t_{(1)}MLE$ is slightly better than robust $RA-$estimates. When the grid size is larger, the loss of edge data is less significant for $t_{(1)}MLE$.

In Table 28 and 29 are 7×7 and 40×40 grids under the situation that error terms came from a Normal distribution and Case 4 with possible outliers in 10% of all observations. The bias and MSE of the robust $M-$ and $RA-$estimates are smaller

Table 27: Comparison of the Robust $RA-$Estimates under Full Data and $t-MLE$ under Partial Data for a $t-$Distribution (25×25 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
RA-est.$_f$	0.6999	0.00014	$2.77e^{-7}$	0.6999	0.00014	$2.42e^{-7}$
$t_{(1)}MLE_p$	0.7001	-0.00014	$1.57e^{-7}$	0.6999	0.00014	$2.08e^{-7}$
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
RA-est.$_f$	0.3999	0.00025	$3.26e^{-7}$	0.3998	0.00050	$5.23e^{-7}$
$t_{(1)}MLE_p$	0.3999	0.00025	$1.39e^{-7}$	0.3999	0.00025	$3.20e^{-7}$

Table 28: Comparison of the Robust $RA-$Estimates under Full Data and $t-MLE$ under Partial Data for Case 4 (7×7 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
RA-est.$_f$	0.6731	0.03843	0.00257	0.6754	0.03514	0.00263
$t_{(1)}MLE_p$	0.6698	0.04314	0.00744	0.6645	0.05071	0.00783
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
RA-est.$_f$	0.2938	0.26550	0.01892	0.2902	0.27450	0.01979
$t_{(1)}MLE_p$	0.2368	0.40800	0.04946	0.2380	0.40500	0.05066

than $t_{(1)}MLE$ for small grid size 7×7.

For larger grid size, $t_{(1)}MLE$ is slightly better than robust $RA-$estimates. When the grid size is larger, the loss information of the edge-data is no longer effective.

Table 30 is 10×10 grid under the situation that mixture error terms came from a skew distributions for Case 4 with possible outliers in 10% of all observations. The bias and MSE of the robust $RA-$estimates are smaller than $t_{(1)}MLE$.

We summarized that when the grid size is small, the loss information of using only partial data is a lot. The robust $RA-$ estimators are better than $t_{(1)}MLE$ when grid size is small or under very un-symmetric condition. Otherwise, when the grid size is large, the $t_{(1)}MLE$ is a flexible estimator under a true $t-$distribution and mixture normal distributions.

Table 29: Comparison of the Robust $RA-$Estimates under Full Data and $t-MLE$ under Partial Data for Case 4 (40×40 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
RA-est.$_f$	0.6962	0.00543	0.00003	0.6962	0.00543	0.00003
$t_{(1)}MLE_p$	0.6980	0.00286	0.00001	0.6980	0.00286	0.00001
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
RA-est.$_f$	0.3898	0.02550	0.00014	0.3894	0.02650	0.00015
$t_{(1)}MLE_p$	0.3928	0.01800	0.00008	0.3928	0.01800	0.00008

Table 30: Comparison of the Robust $RA-$Estimates under Full Data and $t-MLE$ under Partial Data for Case 4 under a Skew Distribution (10×10 Grid).

Estimate	$\hat{\alpha}_1$	Bias	MSE	$\hat{\alpha}_2$	Bias	MSE
RA-est.$_f$	0.4547	0.35043	0.07451	0.4585	0.34500	0.07318
$t_{(1)}MLE_p$	0.2830	0.59571	0.26397	0.2684	0.61657	0.27686
Estimate	$\hat{\alpha}_3$	Bias	MSE	$\hat{\alpha}_4$	Bias	MSE
RA-est.$_f$	0.0641	0.83975	0.12138	0.0673	0.83175	0.11978
$t_{(1)}MLE_p$	0.0255	0.93625	0.14989	0.0374	0.90650	0.14203

1.5.5 Summary for Simulation Studies

We summarize the simulation studies with different distributions:

(1.) Results of analysis with edge data:

(1.1) When residuals present randomly from $t-$distribution, we didn't know where and how much outliers occurred in the grid. $t-$distribution is a heavy-tailed distribution.

(1.1.1) In the smaller grid size, robust estimates have smaller bias and MSE than LSE.

(1.1.2) In the large grid size, the robust estimates have almost 0 bias, and the robust estimates have very small MSE compared to LSE.

(1.2) In additive outliers MRF models, we presented the cases for $\varepsilon_{i,j}$ and v_{ij}

from the standard normal distribution and normal distribution, respectively.

(1.2.1) When there is no outlier present, robust estimates have similar bias and MSE compared with LSE.

(1.2.2) When there are more outliers, the robust RA-estimate is better than robust M-estimator in both bias and MSE.

(1.2.3) In the simulation setting for many outliers uniformly distributed throughout the grid has higher bias and MSE compare to other three cases. In this case, the robust RA method leads to the smallest bias and MSE.

(1.2.4) Naturally, with increasing grid sizes, both bias and MSE become smaller.

(1.3) LSE, robust $M-$ and RA-estimates did not require symmetric assumption in estimation. Therefore, we try to see the effect of estimations under skew distribution outliers setting under the most difficult case with possible outliers in 10% of all observations. For additive outliers MRF models, we presented the cases for $\varepsilon_{i,j}$ from the standard normal distribution and v_{ij} from a log-normal distribution.

(1.3.1) LSE and $M-$estimates are about the same from the bias and MSE.

(1.3.2) $RA-$estimates has better performance in in both bias and MSE than LSE and $M-$estimates.

(2.) Multivariate $t - MLE$ is flexible in the situations with outliers because of its heavy-tailed property. In order to compare robust estimators with $t-MLE$, the estimation procedure were used the full information of the neighborhood data.

Note that robust estimators and LSE didn't need this restriction. Results of analysis without edge data:

(2.1) When we simulated the outliers present randomly from a $t-$distribution:

(2.1.1) In the smaller grid size, robust $RA-$estimates have a little smaller bias and MSE than even true $t_{(1)}MLE$. The decreasing order of estimates of bias and MSE: $NMLE$, LSE, $t_{(1)}MLE$, robust $M-$estimates and $RA-$estimates.

(2.1.2) In the large grid size, $t_{(1)}MLE$, robust $M-$ and $RA-$estimates have similar bias and MSE. Three estimates have almost 0 bias, and the robust estimates have very small MSE compare to LSE.

(2.2) In additive outliers MRF models, we presented the cases for $\varepsilon_{i,j}$ and v_{ij} from the standard normal distribution and a normal distribution, under the most difficult situation with possible outliers in 10% of all observations:

(2.2.1) In the smaller grid size, the robust $M-$ and $RA-$estimates have a little smaller bias and MSE than $t_{(1)}MLE$. In decreasing order of estimates of bias and MSE, we have: $NMLE$, LSE, $t_{(1)}MLE$, robust $M-$estimates and $RA-$estimates.

(2.2.2) For large grid size, $t_{(1)}MLE$ has similar bias and MSE. In decreasing order of estimates of bias and MSE, we have: $NMLE$, LSE, robust $M-$estimates, robust $RA-$estimates, and $t_{(1)}MLE$. Therefore, robust estimates have an advantage in the small grid sizes compared to $t_{(1)}MLE$.

(2.2.3) Naturally, with increasing grid sizes, both bias and MSE of all estimates become smaller.

(2.3) For additive outliers MRF models, we presented the cases for $\varepsilon_{i,j}$ from

the standard normal distribution and v_{ij} from a skew distribution, a log-normal distribution, and the most difficult case with possible outliers in 10% of all observations:

(2.3.1) $NMLE$, LSE, and robust $M-$estimates have similar bias and MSE. The robust $RA-$estimates have better performance in both bias and MSE than the other estimates. In decreasing order of estimates of bias and MSE: $NMLE$, robust $M-$estimates, LSE, $t_{(1)}MLE$, and robust $RA-$estimates.

(2.3.2) Under a very skew distribution, the robust $RA-$estimates have the best performance.

(3.) Here we study the edge effect in LSE, robust $M-$ and RA-estimates.

(3.1) Estimators will lose more information in a small grid size than in a larger grid size.

(3.2) The robust $RA-$ estimates suffer more from using only neighborhood data instead compared to LSE and robust $M-$ estimates.

(4.) Here we compare robust RA-estimates using full data and $t-MLE$ using partial data.

(4.1) The robust $RA-$estimators had an advantage over $t-MLE$ (1) when the grid size was small and data came from a mixture of normal distributions, (2) when the grid size was small and data came from a mixture of normal and skew distributions, and (3) when the grid size was large and data came from a mixture of normal and skew distributions. The robust $RA-$estimators performed comparably to $t-MLE$ when the grid size was large and data came from a mixture of normal distributions.

(4.2) $t - MLE$ performed better than the robust $RA-$estimator when the true error term of the first-order MRF model came from a t-distribution. $t - MLE$ had smaller bias and MSE than the robust $RA-$estimator when the grid size was large; the effect of the loss of information due to using partial data decreased when the grid size became large.

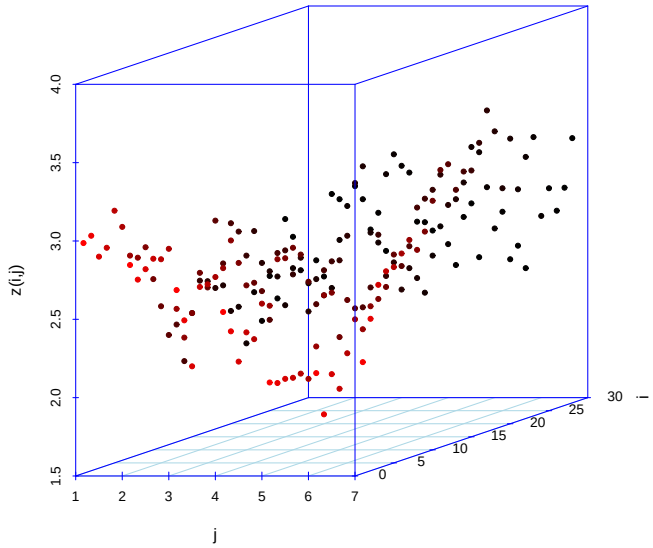


Figure 4: Plot of Kempton and Hower Data (1981).

1.6 Examples

Three data sets taken from the literature, including the Lee-Rawlings data were presented in the introduction. These three data sets are not large sample sizes. From the simulation experience, we know the performance of robust estimation are better than $NMLE$, LSE, and $t-MLE$. In order to avoid losing information from the grid data, here we applied the proposed robust estimators in the three data sets and compared the results to LSE. The three estimators eliminate large scale variation (the mean function of the first order MRF model).

1.6.1 Data from Kempton and Howes (1981)

This data set consists of uniformity trials on a 28×7 grid of spring barley plot yields (Kempton and Howes 1981). The plot in showed in Figure 4. No obvious outliers (or trends) are visually apparent in a plot of the data in Figure 4.

Table 31: Parameter Estimates for Kempton and Howes Data.

Estimate	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$
LSE	0.4470	0.4818	0.0840	0.0668
M-est.	0.4291	0.4886	0.1015	0.0620
RA-est.	0.4420	0.4769	0.1298	0.0649

Because there are no obvious outliers, the robust estimator's tuning constant is set at a small value $c = 1$. Table 31 shows that the three estimates lead to similar results. The left- and right-dependence parameters α_1 and α_2 are larger than the other two dependence parameters.

1.6.2 Data from Cullis, Lill, Fisher, Read, and Gleeson (1989)

Cullis et al. (1989) analyzed data from an experiment assessing the yield potential of test lines from the southern New South Wales wheat breeding programs on a 67×10 grid. Originally, there are two missing values at (7, 1) and (13, 9). This section presents results from two alternative ways of handling the missing data. First, missing values were interpolated using the means of nearest neighbor data shown in Figure 5. Second, two missing values were artificially interpolated as outliers using value "−5" in Figure 6. For both experiments, robust estimator's tuning constant is set at $c = 1.5$.

Table 32 shows that the three methods provide similar estimates in the first experiment. Note that the parameter values in this example range between 0.20 and 0.25. They are smaller than the values of the first two parameters in Section 5.1, but larger than the other two parameter values. The corresponding results for the second experiments are given in Figure 6 and Table 33.

Since there are only two potential outliers in a data set with 670 observations, the impact of these outliers to the estimation results might be limited. Comparison of

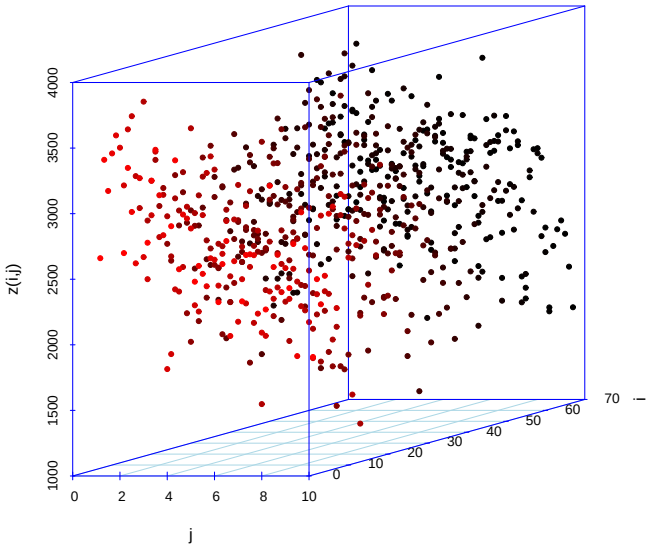


Figure 5: Plot of Cullis et al. Data (1989): neighbor mean for missing values.

Table 32: Parameter Estimates for Cullis et al. Data, Neighbor Mean for Missing Values.

Estimate	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$
LSE	0.2446	0.2375	0.2195	0.2113
M-est.	0.2473	0.2273	0.2220	0.2283
RA-est.	0.2351	0.2334	0.2468	0.2308

Tables 32 and 33 by the percentage of change is shown in Table 34 (calculated as (Table 32 estimate - Table 33 estimate)/ (Table 32 estimate) $\times 100\%$). Note that there is some difference of estimates even in the robust methods. However, the robust estimates are less sensitive to the outliers compared to LSE. The robust RA-estimates are less sensitive to the outliers than the robust M-estimates.

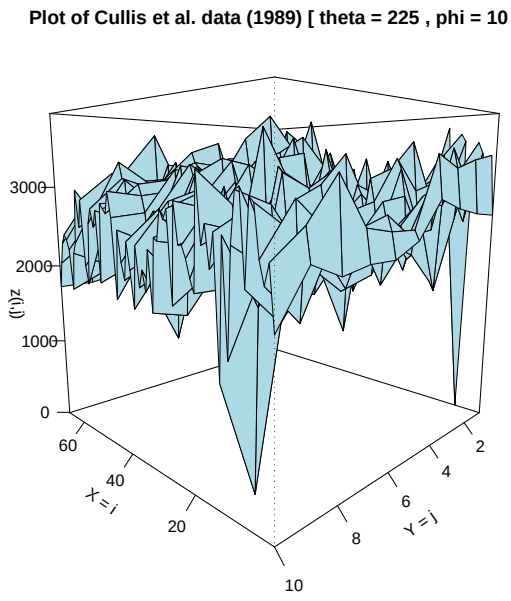


Figure 6: Plot of Cullis et al. Data (1989): "−5" for missing values.

Table 33: Parameter Estimates for Cullis et al. Data, "−5" for Missing Values.

Estimate	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$
LSE	0.2057	0.2008	0.2310	0.2397
M-est.	0.2225	0.2104	0.2254	0.2293
RA-est.	0.2280	0.2208	0.2450	0.2316

Table 34: Parameter Estimates Percentage of Change Comparison in Tables 32 and 33.

Difference(%)	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$
LSE	15.9	15.5	-5.2	13.4
M-est.	10.0	7.4	-1.5	-0.4
RA-est.	3.0	5.4	0.7	-0.3

Table 35: Parameter Estimates for Lee and Rawlings Data.

Estimate	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$
LSE	0.2353	0.1510	0.4519	0.2872
M-est.	0.2322	0.1079	0.4316	0.3188
RA-est.	0.1919	0.1181	0.4613	0.3594

1.6.3 Data from Lee and Rawlings (1982)

Lee and Rawlings (1982) data is from a 6×10 soybean-plant growth chamber, there are two potential outliers in grid (3, 10) and (6, 3) (see Figure 2). Use $c = 1.5$, as the tuning constant. Table 35 shows that the estimates obtained from the two robust methods are much different to the LSE in this example. The third dependence parameter has the largest estimate.

1.7 Concluding Remarks and Future Work

We showed the general asymptotic properties of the robust estimators under different robustifying functions for the first-order MRF model. We also relaxed the assumption of the identical distribution of the error term.

The simulation experiments and real-life examples show that when there are no outliers, the robust estimation methods perform comparably against the least squares method in estimating dependence parameters in the first-order MRF model. When there are additive outliers from the normal distribution, studies in three scenarios for single and multiple outliers located in various patterns indicate that the two robust methods lead to smaller bias and variance. In particular, the robust RA method gives consistently slightly better performance than the robust M method. Naturally, when the grid size is smaller, the improvement from using the robust methods over the LS method is larger.

The distribution of residuals is not restricted to the normal distribution. Any distributions such as t-distribution or other heavy tail distributions can be used. The robust estimates have smaller bias and MSE than LSE. In the large grid size, the robust estimates have almost 0 bias, and the robust estimates have very small MSE compared to LSE.

The ML method is commonly used for parameter estimation in many statistical applications. It is the best estimator when we know the true model distribution. However, it is also well known that multivariate normal MLE is not resistant to outliers, and in this respect it is even worse than LSE.

LSE, robust $M-$ and RA-estimates do not require a symmetric distribution assumption for estimation. Therefore, we tried to see the effect on estimations of a skew distribution of outliers under Case 4. The robust $RA-$estimates had better performance in both bias and MSE than LSE, the robust $M-$estimates, normal MLE,

and $t - MLE$.

Multivariate $t - MLE$ is flexible in situations with outliers because of its heavy-tailed property. The performance of $t_{(1)}MLE$ is better than robust $RA-$estimators hen the first-order MRF model with true error term comes from t-distribution.

Multivariate normal MLE and $t - MLE$ require full neighborhood information. However, LSE, robust $M-$, and RA-estimates did not require this. Edge data is not fully exploited by multivariate normal MLE and $t - MLE$. $t - MLE$ performed better than the robust $RA-$estimator when the true error term of the first-order MRF model came from a t-distribution. $t-MLE$ had smaller bias and MSE than the robust $RA-$estimator when the grid size was large; the effect of the loss of information due to using partial data decreased when the grid size became large.

The robust $RA-$estimators had an advantage over $t - MLE$ (1) when the grid size was small and data came from a mixture of normal distributions, (2) when the grid size was small and data came from a mixture of normal and skew distributions, and (3) when the grid size was large and data came from a mixture of normal and skew distributions. The robust $RA-$estimators performed comparably to $t - MLE$ when the grid size was large and data came from a mixture of normal distributions.

Future work for the robust estimations for the spatial MRF models:

(1.) If removing outlier(s) or understanding the causes of outliers is important, future work will develop an outlier detection procedure to locate potential outlier(s) (Neter et al. 1996 and Rousseeuw and Leroy 1987).

(2.) For simplicity, we naturally assume the overall mean of the first order MRF model is zero. Intuitively, using the overall mean will make the model more stable. Therefore, we could use an iteration algorithm to find the estimate of the overall mean and neighbor parameters.

(3.) For the edge data, we can develop an iteration algorithm to find estimates of the neighborhood parameters. Then these could be used to predict the missing neighborhood data for edge values. Therefore, we have more information about the edge data and complete neighborhood information for all grid points. We could iterate the estimation procedure again to get better neighborhood parameter estimates.

(4.) Extension to the second-order MRF model.

CHAPTER II

EXTENDING THE SKILL PLOT FOR DISEASE DIAGNOSIS

2.1 Introduction

Classification and prediction are both important in many areas: meteorology, economics, computer science, etc. They have been part of discovery by statistical research for a long time. Because of biotechnological innovation, there is even more attention in the relative fields. Sing et al. (2005) pointed out that pattern classification, scoring or ranking predictors are vital in a wide range of biological problems. Examples include predicting phenotypic properties of HIV-1 from genotypic information, microarray analysis for prediction of tissue condition based on gene expression, predicting bio-availability or toxicity of drug compounds, and gauging treatment effect in clinical trials (Brumback et al. 2006). In many cases, robustness and efficiency of the markers of a diagnostic test are critical and the cost of misclassification is a primary factor for classification and prediction.

Diagnostic tests that use markers to determine whether a patient is diseased or healthy are standard tools in medical screening. Early detection is considered essential to effective treatment. Finding new markers that are less invasive, less expensive, and more accurate than existing measures are important in disease prevention (Dodd and Pepe 2003). For the diagnosis of many modern diseases, the difference in marker measurements used to screen healthy patients from diseased patients can be subtle, and statistical researchers work to develop the most effective tool to discern this difference. Misclassification costs are often asymmetric; that is, the cost of misclassifying a healthy patient into the diseased group (a *false positive* result) is often less than

the cost of misclassifying a diseased patient into the healthy group. One tool that has been especially useful in recent decades is the receiver operating characteristic (ROC) curve.

2.1.1 ROC Curve

The receiver operating characteristic (ROC) curve (see Eq.(24) for details) was originally developed for signal detection theory and then was successfully adapted to radiology (Green and Swets 1966, Swets and Pickett 1982, Hanley et al. 1982). In modern medical research, the ROC curve is a widely used technique for quantifying the discriminative ability of screening tests (Pepe 2000, and Chambless and Diao 2006). Those screening tests are applied in detecting a variety of diseases, including cancers (Zou et al. 1997, McIntosh and Pepe 2002, Dodd and Pepe 2003, Briggs and Ruppert 2005), dental disease (Pretty and Maupome 2004), hearing impairment (Pepe and Longton 2005), cardiovascular disease (Cai et al. 2006), and appendicitis (Briggs and Zaretzki 2008).

To describe the ROC, for any case that is subjected to a dichotomous test, there are four possible outcomes. Classically, a 2×2 table illustrates these categories (Table 31). Let Y equal one if the patient actually has the disease, and Y is zero otherwise. Let $p = p_{1+} = P(Y = 1)$ represent the proportion of diseased patients in the total population. Suppose each patient has diagnostic marker response X. If the patient is a member of the healthy population then X has cumulative distribution function (CDF) F, and if the patient is from the diseased group then X has CDF G. Without loss of generality, we will assume the patient is diagnosed as diseased if $X \geq c$ for some fixed threshold c. In many cases, it might be assumed that $F(x) \geq G(x)$ for all values of x.

The sensitivity of a test refers to its ability to correctly identify cases of disease. Note that it depends on the choice of c.

Table 36: Sample of a 2×2 Table.

	Disease Present ($Y = 1$)	Disease absent ($Y = 0$)
Positive test ($X \geq c$)	True-positive	False-positive
Negative test ($X < c$)	False-negative	True-negative

Sensitivity = true-positive rate= $\frac{\text{True}-\text{positive}}{\text{True}-\text{positive}+\text{False}-\text{negative}}$,

and the specificity of a test refers to its ability to correctly identify cases of non-disease. It also depends on the choice of c.

Specificity = 1 - false-positive rate = $\frac{\text{True}-\text{negative}}{\text{False}-\text{positive}+\text{True}-\text{negative}}$.

The ROC curve is a plot of the true-positive rate (TPR = sensitivity) versus the false-positive rate (FPR = 1 − specificity), for a classification rule based on a continuously increasing sequence of threshold values. The graph of TPR ($P(X \geq c|Y = 1)$) vs. FPR ($P(X \geq c|Y = 0)$) defines the ROC curve:

$$\mathrm{R}(t) = 1 - G\left(F^{-1}(1-t)\right) \tag{14}$$

where $0 < t < 1$. The curve shows the inherent trade-off between FPR and TPR. A test can be judged according to how its corresponding ROC curve arches over the 45° line—the more concave, the better.

A typical diagnostic test classifies patients according to a single marker, and the misclassification rates depend on the threshold value that distinguishes the two screening outcomes. However, it is not always certain how to determine the optimal cutoff point (Jager 2001). Pepe (2003, Chapter 4) suggests that factors such as health care

resources, invasive examination etc., can influence the choice of threshold. In general, the choice of the cutoff point depends on:

(1.) The fixed cost k_{10} for classifying a diseased person as a healthy one.

(2.) The fixed cost k_{01} for classifying a healthy person as a diseased patient.

(3.) The overall misclassification probability, $p(1 - \text{TPR}) + (1 - p)\text{FPR}$. By minimizing the overall misclassification probability, the slope of the ROC curve at the optimal cutoff point is $(1 - p)/p$.

(4.) The expected cost of misclassification, $k_{10}p(1 - \text{TPR}) + k_{01}(1 - p)\text{FPR}$. By minimizing the expected cost of misclassification, the slope of the ROC curve at the optimal cutoff point is $(k_{01}(1 - p)) \,/\, (k_{10}p)$.

Theoretical results for ROC curves are well-established. Pepe (2000) developed a semiparametric estimator for ROC curves within the generalized linear model framework for binary regression. Hsieh and Turnbull (1996) considered nonparametric estimators based on empirical distribution functions and derived asymptotic properties. Lloyd and Yong (1999) showed smooth kernel-based estimators outperform this strictly empirical estimator. Claeskens et al. (2003), Dodd and Pepe (2003) and Hall et al. (2004) studied nonparametric methods, e.g. empirical likelihood method or bootstrap, for constructing confidence intervals and confidence bands for estimators of ROC curves.

Briggs and Zaretzki (2006) pointed out advantages of the ROC curve:

(1.) From Eq.(24), an invariance to monotonic transformations of X.

(2.) ROC curves do not depend on the scale of the original test, making it possible to compare multiple curves on the same axis.

(3.) The ability to consider the diagnostic strength of the variable without regard to a cutoff point which must be estimated or chosen arbitrarily.

However, the ROC curve does have a disadvantage: The optimal threshold cutoff point is not easily apparent directly from the plot (Jager 2001) or there might be more than one optimal threshold cutoff point (see our motivating example).

The major statistical challenges for evaluating diagnostic tests in general and for applying ROC methodology are (from Briggs and Zaretzki, 2008):

(1.) If a definitive gold standard assessment of disease status is not available, how can inference for an ROC curve be accomplished? There must be some simple, basic (or relatively not that good) markers from expertise experience that we could use. On the other hand, there may be more than one gold standard marker; Pepe and Longton (2005) try to utilize this information for better performance.

(2.) The test results may be much more complicated, involving several components. Do ROC curves have a role to play in determining how to combine different sources of information to optimize diagnostic accuracy? Many papers used (logistic) regression for combining predictors of classification (Dodd and Pepe 2003, Pepe et al. 2006). When there were more than two diagnostic alternatives, Mossman (1999) and Heckerling (2001) handled this situation by extension to three-way ROCs (three-dimensional view of ROC surface).

(3.) Disease status is often not fixed, but rather can evolve over time. How can the time aspect be incorporated sensibly into ROC analysis? Cai et al. (2006) take the time lag into marker that is measured closer to the time of disease occurrence. Then consider inference for sensitivity and specificity functions that are defined as functions of time. Muijtjens et al. (2006) consider the ROC and loss function analysis in sequential testing.

(4.) Are there alternatives to the ROC curve for describing test accuracy?

2.1.2 ROC for Diagnostic Tests

There are numerous ways of summarizing the ROC curve into an objective test statistic. The area under the ROC curve (AUC), defined as $\int_0^1 \mathrm{R}(t)dt$ was one of the first commonly used measures of test quality. The concept of AUC came from the calculus integral. A no-information diagnostic test can be considered to have 0.5 as its AUC. It can be easily shown that if $Z_0 \sim F$ and $Z_1 \sim G$ are independent, then

$$\int_0^1 \mathrm{R}(t)dt = P(Z_0 \leq Z_1). \tag{15}$$

For continuous data, AUC is equivalent to the probability that a random observation coming from the diseased population (Z_1) is larger than that from the non-diseased population (Z_0).

The AUC is the most commonly used method of summarizing a diagnostic test's overall accuracy (Pretty and Maupome 2004). However, the AUC summarizes test performance over regions of the ROC space that are of no practical interest. The partial area under the curve (PAUC) restricts the AUC-integration to an area of interest, based on FPRs that are considered clinically relevant:

$$\mathrm{A}(t_0, t_1) = \int_{t_0}^{t_1} \mathrm{ROC}(t)dt, \tag{16}$$

where the interval (t_0, t_1) denotes the false-positive rates of interest. If a diagnostics test has an $\mathrm{A}(t_0, t_1)$ which equals $(t_1^2 - t_0^2)/2$, it has no information/no ability to classify individuals correctly. Dodd and Pepe (2003) emphasized that although the partial AUC estimator is a more clinically relevant summary measure of accuracy, the choice of the appropriate restricted region may be controversial. In their simulation studies, they use $(t_0, t_1) = \{(0, 0.1), (0, 0.2), (0.1, 0.2), (0.1, 0.3)\}$ as illustrations. However,

reasonable choices depend on information about the cost associated with true- and false-positive diagnoses. For example, if a diagnostic test is not particularly efficient at screening a disease that has affected most of the at-risk set, the naive guess that every patient has the disease might actually be cost effective.

Dodd and Pepe (2003) describe the significance of the PAUC through the odds,

$$\Lambda(t_0, t_1) = \frac{\mathrm{A}(t_0, t_1)}{(t_1 - t_0) - \mathrm{A}(t_0, t_1)}. \tag{17}$$

This is the odds of the probability of a correct classification to the probability of an mistaken classification, given the test result is from the healthy population in the region (t_0, t_1). Note if the test has an odds of $(t_0 + t_1)/(2 - (t_0 + t_1))$, then the test conveys no information. If the test is perfect, the odds will increase to infinity.

The Skill score and Climate Skill test (Skill test) were recently studied by Mozer and Briggs (2003), Briggs and Ruppert (2005), Briggs and Zaretzki (2008) as a method to evaluate simple yes/no or probabilistic forecasts of binary events by their accuracy. The Skill score and test are unique in that they can take into account both cost of the forecast and the loss of making incorrect forecasts. The Skill plot is a novel alternative to the ROC curve for describing test accuracy used in the above literature. We investigate the skill statistic as an alternative to the AUC, and consider tests at FPRs that produce skillful tests. The Skill score, test and plot are discussed in Section 2.2. Cutoff points in a ROC curve that pass the Skill test are said to have skill. It's very intuitive that only skillful cutoff points should be used. Therefore, with the idea of Skill score, we propose examination of the choice of setting the interval (t_0, t_1) in Eq.(16). We establish properties of the Skill statistics in Section 2.3. Section 2.4 studies the relationship between healthy group and diseased group. Section 2.5 is the simulation studies of PAUC methods. Section 2.6 investigates a motivating example for this research, featuring data for $5,662$ women being diagnosed for osteoporosis. We apply the ideas of the Skill plot, PAUC, odds of PAUC, and the relationship

between non-diseased and diseased groups. We compare two biomarkers to see their abilities for disease diagnosis. Section 2.7 presents concluding remarks and topics for future research.

2.2 The Skill Score

Mozer and Briggs (2003) developed a *skill score* as a method to evaluate probabilistic forecasts of binary events as "skilled" or "not skilled" by integrating the loss from misclassification. They define a diagnostic test as *skillful* if it is more effective in screening disease than the optimal naive guess. It makes sense to use only threshold values that correspond to skillful tests, and the skill score is useful because it considers both the cost of the forecast and the loss of making incorrect forecasts. The *skill plot* (Briggs and Zaretzki, 2008) summarized the diagnostic skill over a range of threshold values and offers a novel alternative to the ROC curve for describing disease diagnosis.

The skill score is based on a simple loss function. If we define

$$\theta = \frac{k_{01}}{k_{01} + k_{10}},$$

then θ is the (relative) loss when $Y = 0$ and $X \geq c$, and $1 - \theta$ is the loss when $Y = 1$ and $X < c$. Without loss of generality, we will assume the misclassification cost of k_{01} is less than k_{10}, so that $\theta \leq 1/2$.

Recall $p = p_{1+} = P(Y = 1)$ is the proportion of diseased patients in the total population. Let $p_{+1} = P(X \geq c) = p\bar{G}(c) + (1 - p)\bar{F}(c)$ = proportion of people classified as diseased, where $\bar{G}(c)$ = P(correctly classify given a diseased person), and $\bar{F}(c)$ = P(classify as diseased person given a healthy person). Probabilities of individual outcomes are summarized in Table 37.

Table 37: Contingency Table.

	$Y = 1$	$Y = 0$	
$X \geq c$	p_{11}	p_{01}	p_{+1}
$X < c$	p_{10}	p_{00}	p_{+0}
	$p = p_{1+}$	$1 - p = p_{0+}$	

Without information from X, the optimal naive classification rule is based solely

on comparing p and θ. There are two possible actions: classify all subjects as healthy people if $p < \theta$ or classify all subjects as diseased patients if $p \geq \theta$. The expected loss for this rule is $E_N = p(1-\theta)I(p < \theta) + (1-p)\theta I(p \geq \theta)$. With the information from X, the expert classification rule is based on a critical cutoff point c. Subjects with $X \geq c$ are classified as diseased, and the others as healthy. A "skill score" can be constructed based upon the relative difference in expected loss between the optimal naive and the expert classification:

$$K_{p,\theta}(c) = \frac{E_N - E_E(c)}{E_N - E_P}, \tag{18}$$

where $E_E(c)$ is the expected loss from the expert guess based on a cutoff point c, and E_P is the expected loss from a perfect classification (we will assume $E_P = 0$). If E_E is based on a diagnostic classification with threshold c, then $E_E(c) = p_{01}\theta + p_{10}(1-\theta)$, and the skill score simplifies to

$$K_{p,\theta}(c) = \frac{p_{11} - p_{+1}\theta}{p(1-\theta)} I(p < \theta) + \frac{p_{+0}\theta - p_{10}}{(1-p)\theta} I(p \geq \theta). \tag{19}$$

The skill plot simply plots $K_{p,\theta}(c)$ versus threshold value of c.

The skill score helps point to an optimal threshold value by finding the value of c that maximizes $\hat{K}(c)$. In a similar vein, Baker (2000) considered a simple linear utility function of FPR and TPR in order to create an optimal test.

2.2.1 Skillful Diagnostic Tests

Because the skill score provides an effective loss-based metric for diagnostic test performance, it seems intuitive that the PAUC should be based only on skillful tests. Instead of integrating the ROC over an arbitrarily chosen range of threshold values, we use only the set of values c for which $K_{p,\theta}(c) \geq 0$. Remark that in general, the definition of the Skill score is greater than zero. We can define a fix value, ω, to minor

adjust the skillful. In the rest of this chapter, we use the original idea of the Skill score is greater than zero. That is $K_{p,\theta}(c) \geq \omega$, where $\omega \in (-Inf, 1)$.

Note that, if $p < \theta$, a positive skill score occurs if TPR $\geq \theta p_{+1}/p_{1+}$, or FPR $\leq (1-\theta)p_{+1}/p_{0+}$. In terms of Y, $K_{p,\theta}(c) \geq 0$ if

$$P(Y = 1|X \geq c) = \frac{p_{11}}{p_{+1}} \geq \theta. \tag{20}$$

For the case $p < \theta$, the PAUC in (16) becomes

$$\int_{K_{p,\theta}(t)\geq 0} \mathrm{R}(t)dt.$$

In the less common scenario when $p \geq \theta$, a positive skill score occurs if TPR $\geq 1 - \theta p_{+0}/p_{1+}$, or equivalently, if $P(Y = 1|X < c) = p_{10}/p_{+0} < \theta$. Figure 7 shows a skill score based on $F(t) = \Phi(t)$, $G(t) = \Phi((t - 1.5)/1.2)$, where Φ represents the standard normal CDF. The relative loss associated with the costs of misclassification determine the range of FPR values (t_0, t_1), which in turn determine the PAUC statistic.

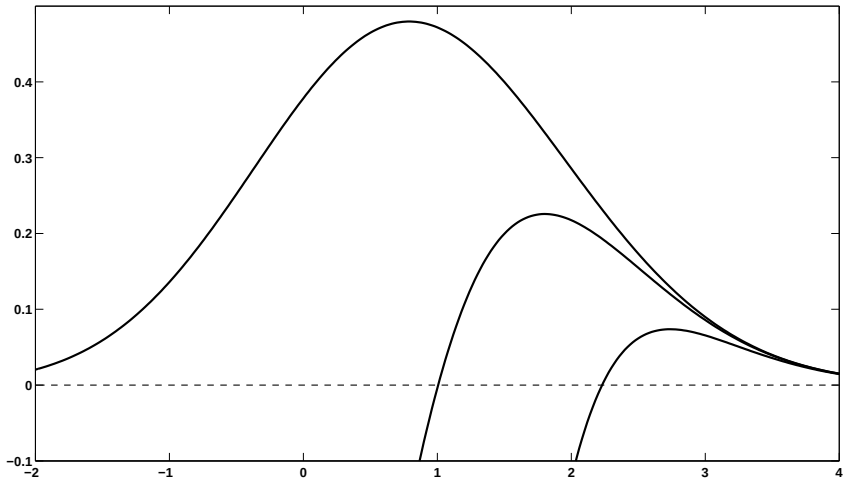


Figure 7: Skill Scores Based on $F(t) = \Phi(t)$, $G(t) = \Phi((t - 1.5)/1.2)$, and $\theta = 0.5$. From left to right, $K(c)$ corresponds to $p = 0.5, 0.2, 0.05$.

2.3 Diagnostic Statistics

Suppose we observe $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ as a training sample consisting of n paired observations where Y_i equals one if the i^{th} person has the disease and equals zero otherwise. The empirical distribution function (EDF) for G, denoted by $G_{n_1}(t)$, is based on $n_1 = \sum_{i=1}^{n} Y_i < n$, and the EDF for F, denoted by $F_{n_0}(t)$, is based on a sample of $n_0 = \sum_{i=1}^{n}(1 - Y_i)$. Note that $\hat{p}_{+1} = n^{-1} \sum_{i=1}^{n} I(X_i \geq c)$. In this paper, we will assume that the sample sizes are such that $n_1/n \to p > 0$ as $n \to \infty$.

The plug-in estimator for the ROC, $\hat{R}(t) = 1 - G_{n_1}(F_{n_0}^{-1}(1 - t))$ simplifies to the proportion of the sample designated as diseased that have marker scores larger than $n_0 p$ out of n_0 sample observations that were classified as healthy (i.e., the p^{th} percentile of F_{n_0}). The nonparametric plug-in estimator $R(t; F_{n_0}, G_{n_1})$ creates a jagged ROC curve (see Figure 15) due to the discrete jumps of F_{n_0} and G_{n_1} at the observations. From Eq.(15), the AUC's empirical analog is the Mann-Whitney U-statistic (Bamber 1975, Hanley and McNeil, 1982, Hsieh and Turnbull 1996). We construct the empirical analog to the skill score, K, in a similar manner. In terms of F and G, we can write

Eq.(19) as

$$K_{p,\theta}(c) = \left(\bar{G}(c) - \frac{(1-p)\theta}{p(1-\theta)}\bar{F}(c)\right) I(p < \theta) + \left(F(c) - \frac{p(1-\theta)}{(1-p)\theta}G(c)\right) I(p \geq \theta). \quad (21)$$

Because we are assuming the healthy population has lower marker scores, if $F(c) \geq G(c)$ and $\theta = p$, then $K_{p,\theta}(c) = F(c) - G(c) \geq 0$. Let H_n be the EDF of the full sample (ignoring group membership) so that

$$\bar{H}_n(c) = \hat{p}_{+1} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \geq c) = \frac{1}{n}\left(n_1\bar{G}_{n_1}(c) + n_0\bar{F}_{n_0}(c)\right).$$

Then the plug-in estimator to Eq.(21) is simply

$$\begin{aligned}\hat{K}_{\hat{p},\theta}(c) &= \left(\bar{G}_{n_1}(c) - \frac{(1-\hat{p})\theta}{\hat{p}(1-\theta)}\bar{F}_{n_0}(c)\right) I(\hat{p} < \theta) + \left(F_{n_0}(c) - \frac{\hat{p}(1-\theta)}{(1-\hat{p})\theta}G_{n_1}(c)\right) I(\hat{p} \geq \theta), \\ &= \frac{\bar{G}_{n_1}(c)\hat{p} - \bar{H}_n(c)\theta}{\hat{p}(1-\theta)} I(\hat{p} < \theta) + \frac{H_n(c)\theta - G_{n_1}(c)\hat{p}}{(1-\hat{p})\theta} I(\hat{p} \geq \theta), \quad (22)\end{aligned}$$

where $\hat{p} = n_1/n$.

2.3.1 Properties of the Skill Statistic

Confidence intervals for the skill score can be constructed using normal approximations, and examples in the next section show these intervals to be effective with sufficiently large samples. In the following theorems, we describe the asymptotic properties of $\hat{K}_{\hat{p},\theta}(c)$ along with estimators for $P(Y = 1|X \geq c)$ and $P(Y = 0|X < c)$.

Theorem 3 Assume F and G are continuous distributions and twice differentiable, have finite mean and variance, and for some $\epsilon > 0$, $\epsilon \leq \theta < 1/2$. Then $\sqrt{n}(\hat{K}_{\hat{p},\theta}(c) - E[\hat{K}_{p,\theta}(c)]) \to N(0, \sigma^2)$, where

$$E[\hat{K}_{\hat{p},\theta}(c)] \approx \frac{\bar{G}(c)p - \bar{H}(c)\theta}{p(1-\theta)} I(p < \theta) + \frac{H(c)\theta - G(c)p}{(1-p)\theta} I(p \geq \theta),$$

and

$$\begin{aligned}\sigma^2 &\approx \frac{1}{p(1-\theta)^2}\left((1-2\theta)\bar{G}(c)G(c)+\frac{\theta^2}{p}\bar{H}(c)H(c)+\frac{1-p}{p^2}\bar{H}^2(c)\theta^2\right)I(p<\theta)\\ &+ \frac{1}{(1-p)^2}\left(\frac{p-2p\theta}{\theta^2}\bar{G}(c)G(c)+\bar{H}(c)H(c)+\frac{p}{(1-p)\theta^2}(H(c)\theta-G(c))^2\right)I(p\geq\theta).\end{aligned}$$

Proof of Theorem 3: By Central Limit Theorem,

$$\sqrt{n}[F_n(x)-F(x)]\rightarrow N[0,F(x)(1-F(x))],$$

Therefore, $E(\bar{H}_n(c))=\bar{H}(c), Var(\bar{H}_n(c))=\frac{1}{n}\bar{H}(c)(1-\bar{H}(c))$.

Same applied on $E(\bar{G}_m(c))=\bar{G}(c), Var(\bar{G}_m(c))=\frac{1}{m}\bar{G}(c)(1-\bar{G}(c))$.

$$\begin{aligned}Cov(\bar{G}_m(c),\bar{H}_n(c)) &= Cov\{\bar{G}_m(c),[\frac{1}{n}[m\bar{G}_m(c)+(n-m)\bar{F}_{n-m}(c)]]\}(independent)\\ &= \frac{m}{n}Var(\bar{G}_m(c))=\frac{1}{n}\bar{G}(c)(1-\bar{G}(c)).\end{aligned}$$

Let $\hat{p}=m/n\sim$binomial$(p), E(\hat{p})=p, Var(\hat{p})=\frac{p(1-p)}{n}$,

$Cov(\hat{p},\bar{H}_n(c))=0$, and $Cov(\hat{p},\bar{G}_m(c))=0$.

Let $W_n^{*T}=[\bar{H}_n,\bar{G}_m,\hat{p}(1-\theta)]$, therefore,

$$\sqrt{n}(W_n^{*T}-E[W_n^{*T}])\rightarrow N[0,\Sigma],$$

where $E[W_n^{*T}]=[\bar{H}(c),\bar{G}(c),p(1-\theta)]$, and,

$$\Sigma=\begin{pmatrix}\sigma_{11}&\sigma_{12}&\sigma_{13}\\ \sigma_{21}&\sigma_{22}&\sigma_{23}\\ \sigma_{31}&\sigma_{32}&\sigma_{33}\end{pmatrix},$$

$\sigma_{11}=\bar{H}(c)(1-\bar{H}(c)), \sigma_{22}=\iota\bar{G}(c)(1-\bar{G}(c))$, where ι is a constant.

$\sigma_{12} = \bar{G}(c)(1-\bar{G}(c)), \sigma_{33} = p(1-p)(1-\theta)^2, \sigma_{13} = 0$, and $\sigma_{23} = 0$.

Use a Taylor expansion: $E[f(X,Y,Z)] \approx f(X_0, Y_0, Z_0)$

$$Var[f(X,Y,Z)] \approx \begin{pmatrix} \frac{\partial f}{\partial X} & \frac{\partial f}{\partial Y} & \frac{\partial f}{\partial Z} \end{pmatrix} \Sigma \begin{pmatrix} \frac{\partial f}{\partial X} \\ \frac{\partial f}{\partial Y} \\ \frac{\partial f}{\partial Z} \end{pmatrix},$$

given $X = X_0, Y = Y_0, Z = Z_0$.

Thus,

$$\sqrt{n}(\hat{K}_\theta(c) - E[\hat{K}_\theta(c)]) \to N(0, \sigma^2),$$

where,

$$E[\hat{K}_\theta(c)] \approx \frac{\bar{H}(c)\left(\bar{G}(c) - \theta\right)}{p(1-\theta)},$$

and

$$\sigma^2 \approx \frac{1}{p^2(1-\theta)^2}\{\frac{1-p}{p}\bar{H}^2(c)[\bar{G}(c)-\theta]^2 + \bar{H}(c)(1-\bar{H}(c))(\bar{G}(c)-\theta)^2$$

$$+2\bar{H}(c)\bar{G}(c)(1-\bar{G}(c))(\bar{G}(c)-\theta) + \iota\bar{H}(c)^2\bar{G}(c)(1-\bar{G}(c))\}.\#$$

Note that appropriate sample sizes for determining the significance of a marker could be defined in Theorem 3, given θ, the diseased rate, the optimal cutoff point of the marker, the proportion of people classified as diseased corresponding to the optimal cutoff point of the marker, and the proportion of diseased persons correctly classified corresponding to the optimal cutoff point of the marker.

The conditional probability from Eq.(20) provides an alternative way to characterize a skillful diagnostic test. The probability of observing a diseased patient conditional on positive diagnosis results $\eta_1(c) = P(Y = 1|X \geq c)$ is expressed in terms of F and G as

$$\eta_1(c) = \frac{\bar{G}(c)p}{\bar{G}(c)p + \bar{F}(c)(1-p)}. \tag{23}$$

Theorem 4 Under the regularity conditions of Theorem 3, for the plug-in estimator of $\eta_1(c)$ in Eq.(23)

$$\hat{\eta}_1(c) = \frac{\bar{G}_{n_1}(c)\hat{p}}{\bar{G}_{n_1}(c)\hat{p} + \bar{F}_{n_0}(c)(1-\hat{p})}, \tag{24}$$

we have $\sqrt{n}\left(\hat{\eta}_1(c) - E(\hat{\eta}_1(c))\right) \to N(0, \sigma^2_{\eta_1}(c))$, where $E(\hat{\eta}_1(c)) \approx \eta_1(c)$ and

$$\sigma^2_{\hat{\eta}_1(c)} \approx \frac{\bar{G}(c)\bar{F}(c)p(1-p)}{(\bar{G}(c)p + \bar{F}(c)(1-p))^4}\left(G(c)\bar{F}(c)(1-p) + \bar{G}(c)F(c)p + \bar{G}(c)\bar{F}(c)\right).$$

When $p < \theta$, recall that $\hat{K} \geq 0$ occurs if $\hat{\eta}_1(c) \geq \theta$. Similarly, when $p \geq \theta$, the probability of observing a healthy subject conditional on negative diagnosis results is expressed in terms of F and G as

$$\eta_0(c) = P(Y = 0|X < c) = \frac{F(c)(1-p)}{G(c)p + F(c)(1-p)}. \tag{25}$$

Theorem 5 Under the regularity conditions of Theorem 3, for the plug-in estimator of $\eta_0(c)$ in Eq.(25)

$$\hat{\eta}_0(c) = \frac{F_{n_0}(c)(1-\hat{p})}{G_{n_1}(c)\hat{p} + F_{n_0}(c)(1-\hat{p})}, \tag{26}$$

we have $\sqrt{n}(\hat{\eta}_0(c) - E(\hat{\eta}_0(c))) \to N(0, \sigma^2_{\eta_0}(c))$, where $E(\hat{\eta}_0(c)) \approx \eta_0(c)$ and

$$\sigma^2_{\hat{\eta}_0(c)} \approx \frac{G(c)F(c)p(1-p)}{(G(c)p + F(c)(1-p))^4}\left(\bar{G}(c)F(c)(1-p) + G(c)\bar{F}(c)p + G(c)F(c)\right).$$

In this case, we assess the skill of a diagnostic test based on whether $\hat{\eta}_0(c) > 1 - \theta$.

The Skill score from Eq.(18) has a more general form called the *Total Expected Misclassification Cost* (TEMC). For example, T_E, the TEMC from the expert forecast, is $np_{01}k_{01} + np_{10}k_{10}$. Under $p < \theta$, the naive guess is that all subjects are healthy, and TEMC based on the naive guess, $T_N = npk_{10}$. The skill score is alternatively expressed as

$$\begin{aligned}\hat{K}_{T,\hat{p},\theta}(c) &= \frac{np_{11}k_{10} - np_{01}k_{01}}{npk_{10}} = \frac{p_{11}k_{10} - p_{01}k_{01}}{pk_{10}} \\ &= \bar{G}_{n_1}(c) - \frac{(1-\hat{p})k_{01}}{pk_{10}}\bar{F}_{n_0}(c).\end{aligned} \tag{27}$$

Note that Eq.(27) is the skill score defined by Expected Misclassification Cost Per Person (EMCPP), which does not depend on the total sample size. These skill score outcomes are summarized in Table 38. The estimates of $\sigma^2_{T_0}$ and $\sigma^2_{T_1}$ in Table 38 are

$$\hat{\sigma}^2_{T_0} \approx \frac{1}{np}\left(\bar{G}(c)G(c) + \frac{1-p}{p}\frac{k_{01}^2}{k_{10}^2}\bar{F}(c)F(c) + \frac{1-p}{p^2}\frac{k_{01}^2}{k_{10}^2}\bar{F}^2(c)\right),$$

and

$$\hat{\sigma}^2_{T_1} \approx \frac{1}{n(1-p)}\left(\frac{1-p}{p}\frac{k_{01}^2}{k_{10}^2}\bar{G}(c)G(c) + \bar{F}(c)F(c) + p(1-2p)^2\frac{k_{01}^2}{k_{10}^2}\bar{G}^2(c)\right).$$

Table 38: The Skill Score Defined by TEMC and EMCPP.

Expert Forecast		
Expected loss	$p_{10}(1-\theta) + p_{01}\theta$	
TEMC	$np_{01}k_{01} + np_{10}k_{10}$	
EMCPP	$p_{01}k_{01} + p_{10}k_{10}$	
Situation	$p < \theta$	$p \geq \theta$
Naive guess	all healthy	all diseased
Expected loss	$p(1-\theta)$	$(1-p)\theta$
TEMC	npk_{10}	$n(1-p)k_{01}$
EMCPP	pk_{10}	$(1-p)k_{01}$
Skill score by TEMC	$\frac{p_{11}k_{10}-p_{01}k_{01}}{pk_{10}}$	$\frac{p_{00}k_{01}-p_{10}k_{10}}{(1-p)k_{01}}$
Estimator	$\bar{G}_m(c) - \bar{F}_{n_0}(c)\frac{1-p}{p}\frac{k_{01}}{k_{10}}$	$F_{n_0}(c) - G_{n_1}(c)\frac{p}{1-p}\frac{k_{10}}{k_{01}}$
Variance	$\hat{\sigma}^2_{T_0}$	$\hat{\sigma}^2_{T_1}$
Skillful	$\frac{p_{11}}{p_{01}} \geq \frac{k_{01}}{k_{10}}$	$\frac{p_{00}}{p_{10}} \geq \frac{k_{10}}{k_{01}}$

2.3.2 Estimating PAUC

We know from Eq.(16) that PAUC is estimated based on a subjectively chosen set of FPR. Dodd and Pepe (2003) admit that controversy is unavoidable with such a subjective choice; FPRs could be unscrupulously chosen in such a way as to maximize the significance of the PAUC statistic, for example. Choices for acceptable FPRs are implicitly a function of the relative loss associated with the type I and type II errors.

With this in mind, the skill score $K(c)$ offers a more objective and coherent way of selecting the subset of FPR according to these fixed loss functions. To eliminate the inherent subjectivity in the PAUC, we consider the set of FPR such that $K(c) \geq 0$, which corresponds to $FPR \leq (1-\theta)p_{+1}/p_{0+}$. The corresponding empirical estimator of PAUC,

$$\hat{A}_K = \int_{\hat{K}_{p,\theta}(t)\geq 0} R(t)dt, \tag{28}$$

is based on the estimated skill score in Eq.(16).

To show how the estimator is constructed, we introduce a practical application in the following section. The results show that the skill score can vary greatly depending on the loss function, so costs for misclassification should not be chosen arbitrarily by the practitioner.

With this infusion of extra empirical information on the skill score comes added uncertainty. In turn, this new estimator would be an inferior choice to the regular PAUC estimator based on expert opinion, as long as the expert opinion is accurate. While this is not the norm, Pepe (2003) includes actual case studies in which past data can aid in deciding valid FPR values.

2.4 Relationship between Diseased and Healthy Groups

In our motivating example, the graphs in Figure 14 indicate, there is a noticeable difference between markers for the patients who experienced hip fractures (G) and those who do not (F). What is more interesting is that the distributions do have a related shape that can be characterized with related relationships through the distributions F and G. For example, F and G might differ only by a location shift, a scale change, or some other simple transformation. Here we try to study the relationship between diseased and healthy groups and the estimators for the disease diagnosis under this relationship.

2.4.1 KG Model

One particular model, introduced by Koziol and Green (1976) has garnered particular interest from researchers in life testing. The Koziol-Green (KG) model stipulates that $G(t) = F(t)^\beta$ for some $\beta > 0$.

This model naturally induces an ordering between F and G, depending on the value of β; one can show that $\beta > 1$ if and only if G is smaller than F in likelihood ratio ordering. That is, if $Z_0 \sim F$ and $Z_1 \sim G$, Z_0 is less than Z_1 in likelihood ratio if and only if $G(F^{-1})$ is convex. The order between F and G is simply reversed in the case $\beta \leq 1$. The refined KG model offers great potential efficiency in survival and censoring data but the model has been shown to be inconveniently restrictive. Csörgő (1981) showed that this assumption is insupportable in typical sets of lifetime data where the relationship between F and G cannot be characterized this simply. However, it will have particular relevance in the motivating data set considered in Section 2.6.2.1.

Originally, the KG model was used to describe the relationship between the lifetime and censoring distribution of the survival function. If it holds that $G(t) = F(t)^\beta$,

then $ROC(t) = 1 - (1-t)^{\beta}$, and $AUC = 1 - 1/(1+\beta)$. In this case, the skill score simplifies to

$$K_{p,\theta}(c) = \left(1 - F^{\beta}(c) - \tfrac{(1-p)\theta}{p(1-\theta)}\bar{F}(c)\right) I(p < \theta) + \left(F(c) - \frac{p(1-\theta)}{(1-p)\theta}F(c)^{\beta})\right) I(p \geq \theta).$$

And

$$\text{PAUC}(t_0, t_1) = t_1 - t_0 - \frac{1}{1+\beta}\left((1-t_0)^{1+\beta} - (1-t_1)^{1+\beta}\right).$$

The odds of $\text{PAUC}(t_0, t_1)$,

$$\Lambda(t_0, t_1) = \frac{(t_1 - t_0)(1+\beta) - (1-t_0)^{1+\beta} + (1-t_1)^{1+\beta}}{(1-t_0)^{1+\beta} - (1-t_1)^{1+\beta}}.$$

2.4.2 Mean-Shift Model

The case where F and G differ by a location shift does not fit into the KG model framework, but it has been considered as a special case of interest. If $G(c) = F(c-\delta)$, with $\delta > 0$, we have $ROC(t) = 1 - F(F^{-1}(1-t) - \delta)$ and $AUC = \int_0^1 1 - F(F^{-1}(1-t) - \delta)dt$. The skill score is

$$K_{p,\theta}(c) = \left[\bar{F}(c-\delta) - \frac{(1-p)\theta}{p(1-\theta)}\bar{F}(c)\right] I(p < \theta) + \left[F(c) - \frac{p(1-\theta)}{(1-p)\theta}F(c-\delta)\right] I(p \geq \theta). \quad (29)$$

And

$$\text{PAUC}(t_0, t_1) = \int_{t_0}^{t_1} 1 - F(F^{-1}(1-t) - \delta)dt.$$

The PAUC odds,

$$\Lambda(t_0, t_1) = \frac{\int_{t_0}^{t_1} 1 - F(F^{-1}(1-t) - \delta)dt}{t_1 - t_0 - \int_{t_0}^{t_1} 1 - F(F^{-1}(1-t) - \delta)dt}.$$

2.4.3 Equality Relationship

In the case of equality $G(c) = F(c)$, we have $ROC(t) = t$, $AUC = 1/2$ and the skill score further simplifies to

$$K_{p,\theta}(c) = \left[\left(1 - \frac{(1-p)\theta}{p(1-\theta)}\right)\bar{F}(c)\right] I(p < \theta) + \left[\left(1 - \frac{p(1-\theta)}{(1-p)\theta}\right)F(c))\right] I(p \geq \theta).$$

We know $p \leq \theta$ implies $(1-p)\theta \geq p(1-\theta)$, and vice versa. In this case, it is easy to show that $K_{p,\theta}(c) \leq 0$ for all values of (p, θ), so the test can be nowhere skillful. Because of this, the skillful range is 0. Thus, the PAUC and its odds are 0.

2.5 Simulation Studies of PAUC

In Section 2.3, we examined properties of the skill score, which depend strongly on the underlying provided loss function as well as the true underlying FPR. In this section, we provide a simulation study that will more practically illustrate the viability of using the Skill score in disease diagnosis.

2.5.1 Study Setup

For comparative purposes, we use the same simulation settings as Dodd and Pepe (2003). A marker score of a patient in the diseased group is generated from a normal distribution with mean 1.5 and variance 1.44. A marker score of a subject from the non-diseased group is distributed as the standard normal distribution. We set up simulations with different sample sizes for four disease rates (p): 0.091, 0.167, 0.333, and 0.5. In order to study the effect of sample size, we set up a doubled sample size and a tenfold sample size in each of the four disease rates. When $p = 0.091$, $n_1 = 10$ and $n_0 = 100$; the doubled sample size has $n_1 = 20$ and $n_0 = 200$; and the tenfold sample size has $n_1 = 100$ and $n_0 = 1000$. For $p = 0.167$, $n_1 = 10$ and $n_0 = 50$; the doubled sample size has $n_1 = 20$ and $n_0 = 100$; and the tenfold sample size has $n_1 = 100$ and $n_0 = 500$. For $p = 0.333$, $n_1 = 50$ and $n_0 = 100$; the doubled sample size has $n_1 = 100$ and $n_0 = 200$; and the tenfold sample size has $n_1 = 500$ and $n_0 = 1000$. For $p = 0.5$, $n_1 = 50$ and $n_0 = 50$; the doubled sample size has $n_1 = 100$ and $n_0 = 100$; and the tenfold sample size has $n_1 = 500$ and $n_0 = 500$.

We take the theoretical quantiles under a uniform partition of the sample size for diseased group from the $N(1.5, 1.44)$ and non-disease group from the standard normal distribution. By transforming back to theoretical values of markers, we calculate the the Skill score by different loss function values (θ) ranging from 0.01 to 0.6. The skillful region of FPR is defined under the Skill score. 1000 simulation results for

mean-result of skillful region of FPR. Note that mean of t_0 and t_1 in simulations did not include the nonskilled part. Based on a skillful region, (t_0 and t_1), we calculate the theoretical PAUC (T. PAUC) and the plug-in PAUC estimate under various θ. The results of the comparison is in Section 2.5.2.

2.5.2 Evaluation of Results of the Skillful Regions and PAUCs

Under $p = 0.091$, Tables 39–41 compare the theoretical skillful region of FPR and the mean skillful region of FPR from 1000 simulations. In Table 39, when θ is higher (0.6) the nonskilled frequency increased to about 30%. The reason is the skillful region of FPR is low (0, 0.01) for $\theta = 0.6$. Therefore, in some cases, although the theoretical skillful region is not zero, the results from simulation might not be skillful at all. When we double the sample size (Table 40), the nonskilled frequency decreased to about 11%. With a tenfold increase in the sample size (Table 41), the nonskilled frequency decreased to about zero. The nonskilled rate decreased when we increased the sample size in our studies.

The theoretical skillful region of FPR and simulations mean skillful region of FPR are about the same. This result is true for all case studies. Later in the simulation studies, we use the simulation skillful region as the comparison of theoretical PAUC and plug-in PAUC estimates. The mean squared errors (MSEs) are provided. In Table 41, the MSEs of PAUC estimates are very small. In the left top of Figure 8, the MSEs decreased to almost zero when we increased the sample sizes. The largest MSE happened for $p = \theta$.

Increasing the disease rate to 0.167, Tables 42–44 compare the theoretical skillful region of FPR and the mean skillful region of FPR from 1000 simulations. When θ is higher (0.5 or 0.6) the nonskilled frequency increased to more than 10%. However, the nonskilled rate decreased when we increased the sample size in our studies. In

Table 39: Comparison of the Theoretical Skillful Region and 1000 Simulation Studies, n_1=10, n_0=100, and p=0.091.

	Theo.		1000 simulations				
θ	t_0	t_1	t_0	t_1	Nonskilled	T. PAUC	Esti. PAUC (MSE)
0.01	0.68	0.99	0.605	0.990	2.3%	0.3571	0.3758 (0.00087)
0.05	0.09	0.99	0.117	0.985	0.1%	0.7513	0.7674 (0.00330)
0.091	0	0.99	0.002	0.987	0.0%	0.8183	0.8162 (0.00547)
0.1	0	0.89	0.002	0.874	0.0%	0.7061	0.7057 (0.00502)
0.2	0	0.31	0.002	0.306	0.3%	0.1880	0.1971 (0.00139)
0.3	0	0.16	0.002	0.144	1.3%	0.0698	0.0786 (0.00050)
0.5	0	0.04	0.001	0.039	7.0%	0.0114	0.0154 (0.00007)
0.6	0	0.01	0.002	0.022	27.6%	0.0391	0.0067 (0.00003)

Table 40: Comparison of the Theoretical Skillful Region and 1000 Simulation Studies, n_1=20, n_0=200, and p=0.091.

	Theo.		1000 simulations				
θ	t_0	t_1	t_0	t_1	Nonskilled	T. PAUC	Esti. PAUC (MSE)
0.01	0.805	0.995	0.659	0.992	1.0%	0.3149	0.3259 (0.00033)
0.05	0.075	0.995	0.096	0.993	0.0%	0.7785	0.7848 (0.00172)
0.091	0	0.995	0	0.994	0.0%	0.8256	0.8230 (0.00263)
0.1	0	0.895	0	0.883	0.0%	0.7152	0.7133 (0.00251)
0.2	0	0.315	0.001	0.314	0.0%	0.1940	0.1974 (0.00074)
0.3	0	0.150	0.001	0.147	0.2%	0.0716	0.0757 (0.00022)
0.5	0	0.040	0	0.039	1.8%	0.0118	0.0141 (0.00003)
0.6	0	0.015	0.001	0.021	11.1%	0.0045	0.0062 (0.00001)

Table 43, the nonskilled rate decreased to under 5%. In Table 44, the MSEs of PAUC estimates are very small. In the right top of Figure 8, the MSEs decreased to almost zero when we increased the sample sizes. The largest MSE happened for $p = \theta$.

When the disease rate was increased to 0.333, Tables 45–47 show that lower θ (0.01) had nonskilled frequency of about 8%. However, the nonskilled rate decreased only 1% for larger sample sizes (Tables 46 and 47). The MSEs of PAUC estimates are very small. In fact, the MSEs of PAUC estimates decreased to almost zero when we increased the sample sizes, and the largest MSE of PAUC estimates happened when

Table 41: Comparison of the Theoretical Skillful Region and 1000 Simulation Studies, n_1=100, n_0=1000, and p=0.091.

	Theo.		1000 simulations				
θ	t_0	t_1	t_0	t_1	Nonskilled	T. PAUC	Esti. PAUC (MSE)
0.01	0.865	0.999	0.768	0.998	1.5%	0.2215	0.2241 (0.00002)
0.05	0.073	0.999	0.075	0.998	0.0%	0.8002	0.8023 (0.00042)
0.091	0	0.998	0	0.998	0.0%	0.8292	0.8300 (0.00052)
0.1	0.058	0.890	0	0.888	0.0%	0.7205	0.7213 (0.00051)
0.2	0	0.323	0	0.322	0.0%	0.2002	0.2013 (0.00017)
0.3	0	0.151	0	0.152	0.0%	0.0745	0.0754 (0.00005)
0.5	0	0.041	0	0.042	0.0%	0.0129	0.0134 (0.00000)
0.6	0	0.021	0	0.022	0.0%	0.0053	0.0057 (0.00000)

Table 42: Comparison of the Theoretical Skillful Region and 1000 Simulation Studies, n_1=10, n_0=50, and p=0.167.

	Theo.		1000 simulations				
θ	t_0	t_1	t_0	t_1	Nonskilled	T. PAUC	Esti. PAUC (MSE)
0.01	0.68	0.98	0.598	0.980	4.0%	0.3473	0.3664 (0.00094)
0.05	0.40	0.98	0.342	0.971	1.6%	0.5593	0.5871 (0.00210)
0.1	0	0.98	0.098	0.971	0.1%	0.7488	0.7335 (0.00363)
0.167	0	0.98	0.003	0.975	0.0%	0.8060	0.8034 (0.00582)
0.2	0	0.78	0.003	0.745	0.0%	0.5804	0.5822 (0.00457)
0.3	0	0.36	0.003	0.372	0.2%	0.2415	0.2511 (0.00193)
0.5	0	0.10	0.006	0.100	11.9%	0.0372	0.0471 (0.00033)
0.6	0	0.06	0.004	0.056	15.4%	0.0165	0.0233 (0.00015)

the disease rate was equal to the loss function.

When disease rate was increased to 0.5, Tables 48–50 show that lower θ ranging from 0.01 to 0.1 had very high nonskilled frequency (more than 10%). However, the nonskilled rate did not decrease for larger sample sizes (Table 49 and Table 50). The MSEs of PAUC estimates were very small. The MSEs decreased to almost zero when we increased the sample sizes, and the largest MSE happened when the disease rate was equal to the loss function.

Table 43: Comparison of the Theoretical Skillful Region and 1000 Simulation Studies, n_1=20, n_0=100, and p=0.167.

	Theo.		1000 simulations				
θ	t_0	t_1	t_0	t_1	Nonskilled	T. PAUC	Esti. PAUC (MSE)
0.01	0.80	0.99	0.728	0.990	3.5%	0.2452	0.2528 (0.00018)
0.05	0.35	0.99	0.358	0.985	1.3%	0.5669	0.5824 (0.00088)
0.1	0.06	0.99	0.079	0.986	0.0%	0.7814	0.7874 (0.00193)
0.167	0	0.996	0.001	0.988	0.0%	0.8194	0.8177 (0.00280)
0.2	0	0.75	0.001	0.762	0.0%	0.5804	0.5966 (0.00242)
0.3	0	0.39	0.001	0.384	0.0%	0.2517	0.2562 (0.00105)
0.5	0	0.11	0.002	0.110	1.7%	0.0477	0.0529 (0.00015)
0.6	0	0.05	0.002	0.059	3.3%	0.0205	0.0244 (0.00006)

Table 44: Comparison of the Theoretical Skillful Region and 1000 Simulation Studies, n_1=100, n_0=500, and p=0.167.

	Theo.		1000 simulations				
θ	t_0	t_1	t_0	t_1	Nonskilled	T. PAUC	Esti. PAUC (MSE)
0.01	0.944	0.998	0.876	0.998	3.3%	0.1164	0.1177 (0.00001)
0.05	0.370	0.998	0.368	0.995	0.1%	0.5848	0.5897 (0.00014)
0.1	0.058	0.998	0.063	0.997	0.0%	0.8055	0.8073 (0.00046)
0.167	0	0.996	0	0.995	0.0%	0.8270	0.8275 (0.00055)
0.2	0	0.774	0	0.774	0.0%	0.6087	0.6094 (0.00050)
0.3	0	0.396	0	0.394	0.0%	0.2600	0.2611 (0.00024)
0.5	0	0.120	0	0.119	0.0%	0.0536	0.0547 (0.00003)
0.6	0	0.066	0	0.065	0.0%	0.0236	0.0244 (0.00001)

To summarize the four disease rates' simulation studies, Figure 8 shows the MSEs of PAUC estimates decreased to zero in all disease rates when we increased (twofold and tenfold) the sample sizes.

2.5.2.1 Reasonable Regions of FPR

The choice of the region of FPR is controversial (Dodd and Pepe 2003). A reasonable region of FPR should be based on a clinically relevant summary measure of accuracy.

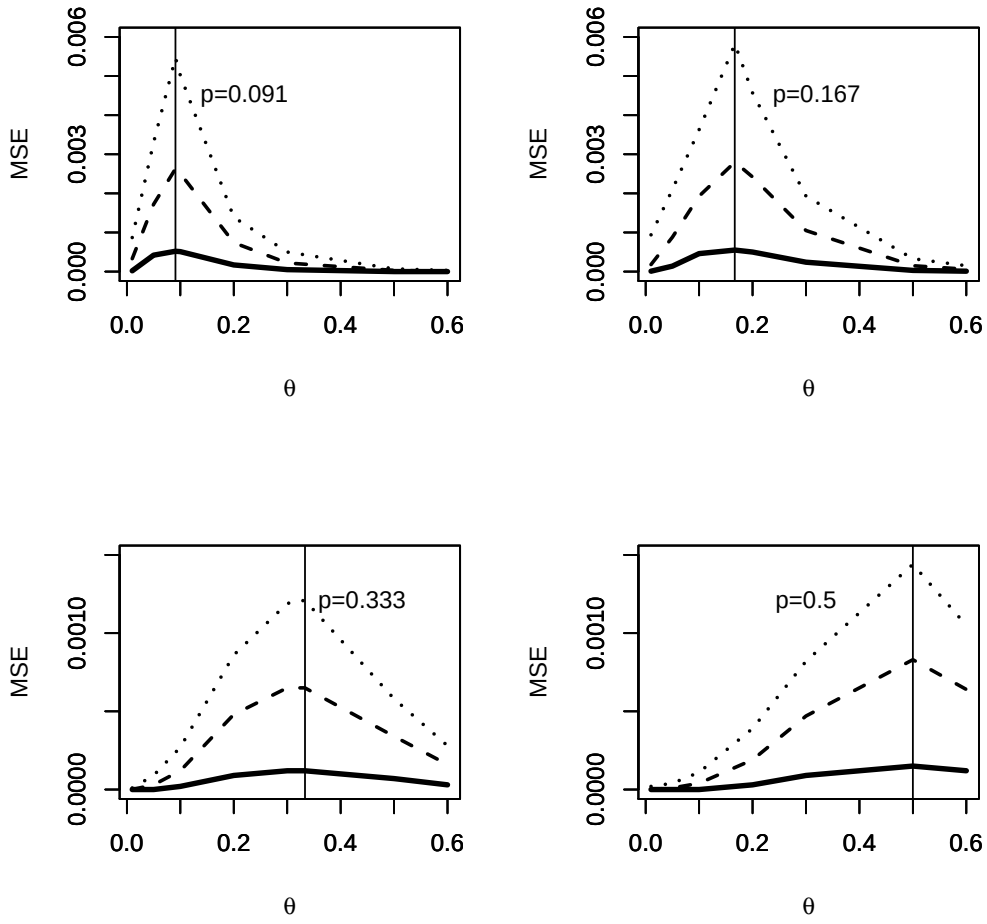


Figure 8: The MSE of PAUC by using the Skill Region. The dotted lines correspond to: $n_1 = 10$, $n_0 = 100$ (top left); $n_1 = 10$, $n_0 = 50$ (top right); $n_1 = 50$, $n_0 = 100$ (bottom left); $n_1 = 50$, $n_0 = 50$ (bottom right). In each case, the dashed line corresponds to a doubled sample size, and the wide solid line to a tenfold increase.

Table 45: Comparison of the Theoretical Skillful Region and 1000 Simulation Studies, n_1=50, n_0=100, and p=0.333.

	Theo.		1000 simulations				
θ	t_0	t_1	t_0	t_1	Nonskilled	T. PAUC	Esti. PAUC (MSE)
0.01	0.90	0.99	0.840	0.990	8.2%	0.1352	0.1373 (0.00001)
0.05	0.78	0.99	0.704	0.984	5.6%	0.2561	0.2613 (0.00009)
0.1	0.44	0.99	0.452	0.979	1.0%	0.4891	0.4980 (0.00027)
0.2	0.09	0.99	0.093	0.986	0.0%	0.7752	0.7805 (0.00086)
0.3	0	0.99	0.004	0.988	0.0%	0.8187	0.8204 (0.00119)
0.333	0	0.99	0	0.988	0.0%	0.8197	0.8208 (0.00193)
0.5	0	0.32	0	0.419	0.0%	0.2807	0.2839 (0.00058)
0.6	0	0.25	0	0.243	0.0%	0.1386	0.1422 (0.00028)

Table 46: Comparison of the Theoretical Skillful Region and 1000 Simulation Studies, n_1=100, n_0=200, and p=0.333.

	Theo.		1000 simulations				
θ	t_0	t_1	t_0	t_1	Nonskilled	T. PAUC	Esti. PAUC (MSE)
0.01	0.945	0.995	0.904	0.995	7.5%	0.0830	0.0837 (0.00000)
0.05	0.865	0.995	0.752	0.991	4.4%	0.2222	0.2252 (0.00003)
0.1	0.450	0.995	0.458	0.989	0.5%	0.4988	0.5042 (0.00012)
0.2	0.085	0.995	0.091	0.993	0.0%	0.7855	0.7884 (0.00048)
0.3	0.005	0.995	0.004	0.994	0.0%	0.8250	0.8258 (0.00065)
0.333	0	0.995	0	0.994	0.0%	0.8256	0.8262 (0.00065)
0.5	0	0.425	0	0.427	0.0%	0.2875	0.2887 (0.00034)
0.6	0	0.245	0	0.247	0.0%	0.1415	0.1429 (0.00016)

It might also depend on information about the cost associated with true- and false-positive diagnoses. In the Dodd and Pepe (2003) simulation studies, fixed regions were used: $(t_0, t_1) = \{(0, 0.1), (0, 0.2), (0.1, 0.2), (0.1, 0.3)\}$ as illustrations. Figure 9 shows the region of FPR in the setting of Dodd and Pepe (2003).

Figures 10–13 (large sample sizes) show the different skillful regions of FPR from simulation studies, based on the consideration of disease rates and loss function values. Note that when the disease rate is equal to the loss function, the skillful region of FPR is the whole region of FPR from 0 to 1 (simulation setting in Section 2.5.1, i.e.

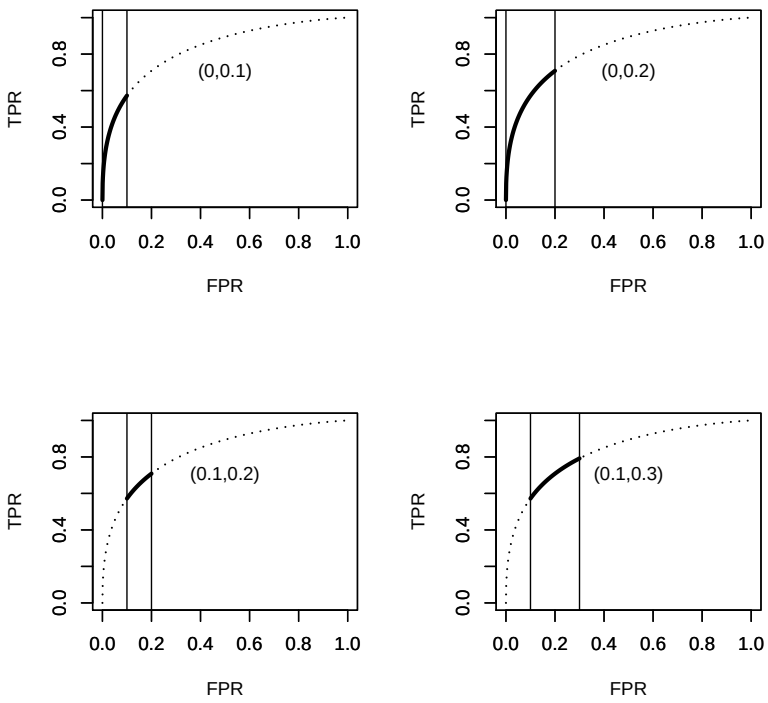


Figure 9: ROC Curve (the dotted line) and Fixed Regions (the wide solid line) of FPR for (0,0.1),(0,0.2),(0.1,0.2), and (0.1,0.3) from Dodd and Pepe (2003).

Table 47: Comparison of the Theoretical Skillful Region and 1000 Simulation Studies, n_1=500, n_1=1000, and p=0.333.

	Theo.		1000 simulations				
θ	t_0	t_1	t_0	t_1	Nonskilled	T. PAUC	Esti. PAUC (MSE)
0.01	0.986	0.999	0.974	0.999	7.3%	0.0228	0.0229 (0.00000)
0.05	0.865	0.999	0.833	0.996	1.7%	0.1584	0.1590 (0.00000)
0.1	0.467	0.999	0.467	0.998	0.1%	0.5061	0.5073 (0.00002)
0.2	0.085	0.999	0.086	0.999	0.0%	0.7954	0.7962 (0.00009)
0.3	0.003	0.999	0.003	0.999	0.0%	0.8300	0.8305 (0.00012)
0.333	0	0.999	0	0.999	0.0%	0.8304	0.8308 (0.00012)
0.5	0	0.432	0	0.432	0.0%	0.2923	0.2927 (0.00007)
0.6	0	0.251	0	0.251	0.0%	0.1445	0.1449 (0.00003)

Table 48: Comparison of the Theoretical Skillful Region and 1000 Simulation Studies, n_1=50, n_0=50, and p=0.5.

	Theo.		1000 simulations				
θ	t_0	t_1	t_0	t_1	Nonskilled	T. PAUC	Esti. PAUC (MSE)
0.01	0.90	0.98	0.829	0.980	15.4%	0.1256	0.1278 (0.00002)
0.05	0.90	0.98	0.803	0.978	14.9%	0.1452	0.1483 (0.00004)
0.1	0.78	0.98	0.682	0.968	10.5%	0.2466	0.2525 (0.00011)
0.2	0.40	0.98	0.395	0.963	2.4%	0.5124	0.5229 (0.00039)
0.3	0.14	0.98	0.155	0.970	0.0%	0.7190	0.7269 (0.00082)
0.5	0	0.98	0	0.976	0.0%	0.8078	0.8086 (0.00144)
0.6	0	0.62	0	0.606	0.0%	0.4490	0.4516 (0.00105)

$F(x) \geq G(x)$ for all values of x). Thus, the PAUC estimator has the skillful region of FPR (0,1) which is AUC. Therefore, we conclude that the Skill score is very helpful in determining the skillful region of FPR for the PAUC.

2.5.3 Summary for Simulation Studies

The simulation experiments show the mean skillful region of FPR is about the same as the theoretical skillful region. When the sample size increased, the mean skillful region of FPR converged to the theoretical skillful region of FPR. The MSEs of

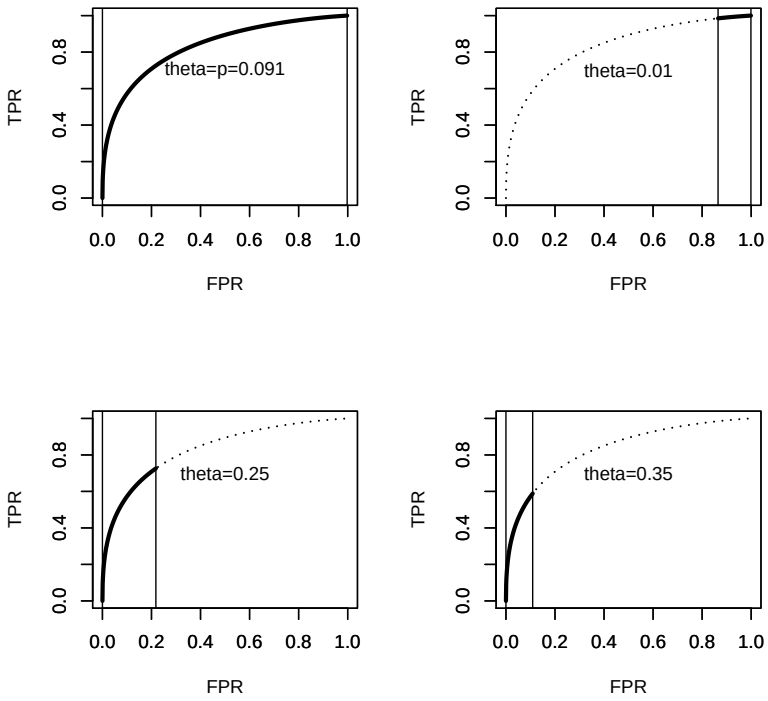


Figure 10: ROC Curve (the dotted line) and the Skillful Regions (the wide solid line) for θ=0.091, 0.01, 0.25, 0.35. n_1=100, n_0=1000, and p=0.091.

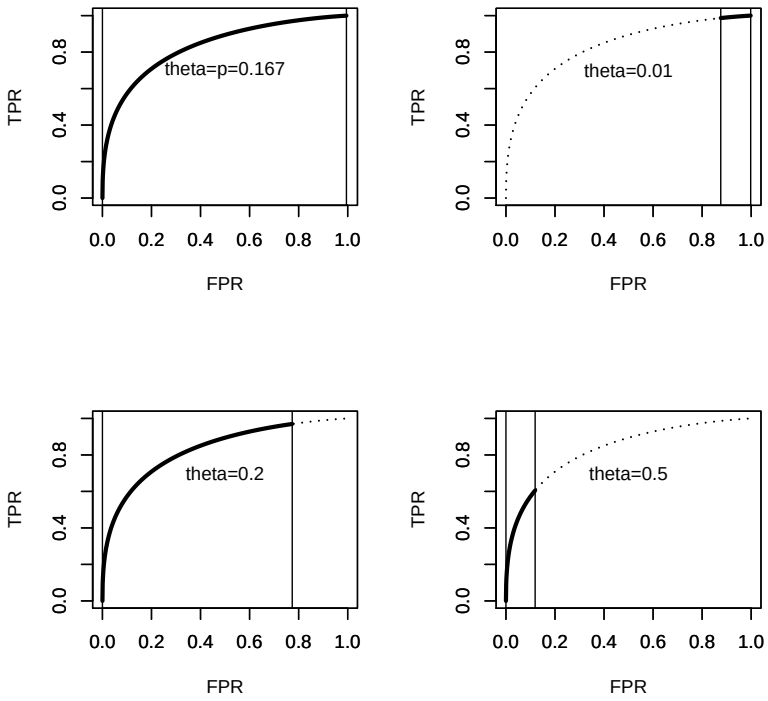


Figure 11: ROC Curve (the dotted line) and the Skillful Regions (the wide solid line) for θ=0.167, 0.01, 0.2, 0.5. n_1=100, n_0=500, and p=0.167.

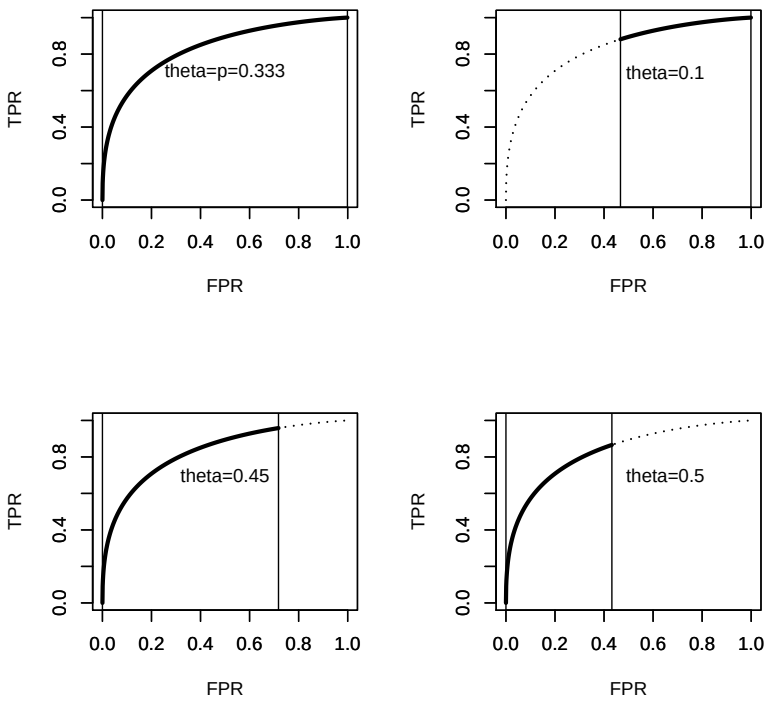


Figure 12: ROC Curve (the dotted line) and the Skillful Regions (the wide solid line) for θ=0.333, 0.1, 0.45, 0.5. n_1=500, n_0=1000, and p=0.333.

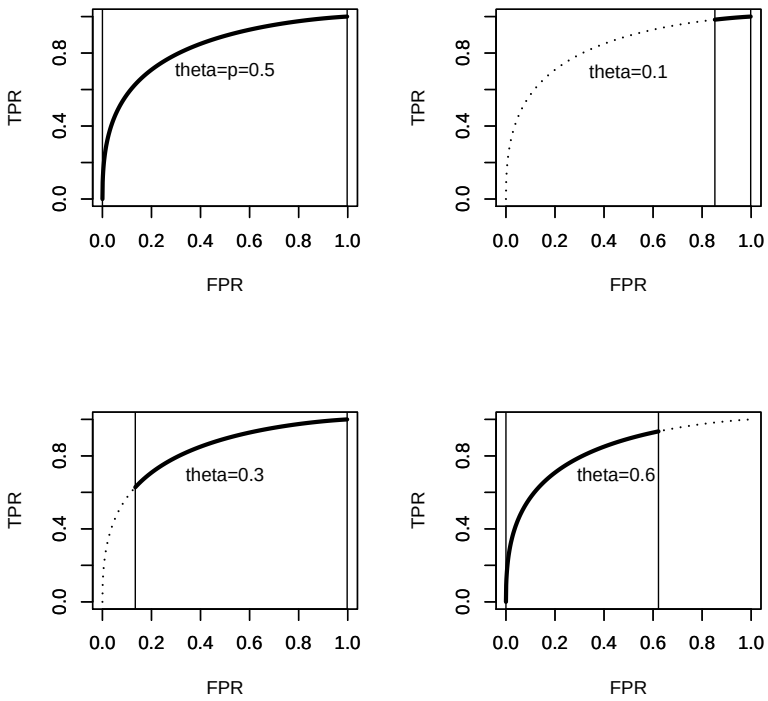


Figure 13: ROC Curve (the dotted line) and the Skill Region (the wide solid line) for θ=0.5, 0.1, 0.3, 0.6. n_1=500, n_0=500, and p=0.5.

Table 49: Comparison of the Theoretical Skillful Region and 1000 Simulation Studies, n_1=100, n_0=100, and p=0.5.

	Theo.		1000 simulations				
θ	t_0	t_1	t_0	t_1	Nonskilled	T. PAUC	Esti. PAUC (MSE)
0.01	0.94	0.99	0.897	0.990	15.6%	0.0776	0.0783 (0.00000)
0.05	0.94	0.99	0.858	0.988	14.6%	0.1087	0.1101 (0.00001)
0.1	0.80	0.99	0.732	0.979	7.6%	0.2220	0.2254 (0.00004)
0.2	0.41	0.99	0.399	0.979	0.5%	0.5381	0.5445 (0.00019)
0.3	0.14	0.99	0.147	0.985	0.0%	0.7419	0.7464 (0.00047)
0.5	0	0.99	0	0.988	0.0%	0.8196	0.8198 (0.00083)
0.6	0	0.61	0	0.616	0.0%	0.4585	0.4596 (0.00064)

Table 50: Comparison of the Theoretical Skillful Region and 1000 Simulation Studies, n_1=500, n_0=500, and p=0.5.

	Theo.		1000 simulations				
θ	t_0	t_1	t_0	t_1	Nonskilled	T. PAUC	Esti. PAUC (MSE)
0.01	0.944	0.998	0.975	0.998	13.2%	0.0203	0.0204 (0.00000)
0.05	0.370	0.998	0.941	0.996	10.4%	0.0491	0.0493 (0.00000)
0.1	0.058	0.998	0.814	0.993	2.4%	0.1724	0.1730 (0.00000)
0.2	0	0.774	0.397	0.996	0.0%	0.5642	0.5656 (0.00003)
0.3	0	0.396	0.137	0.997	0.0%	0.7632	0.7642 (0.00009)
0.5	0	0.120	0	0.998	0.0%	0.8292	0.8295 (0.00015)
0.6	0	0.066	0	0.622	0.0%	0.4634	0.4638 (0.00012)

PAUC estimates were very small. In fact, the MSEs decreased to almost zero when we increased the sample sizes. The largest MSE happened when the disease rate was equal to the loss function.

The purpose for the optimal cutoff points is to maximize the TPR and minimize the FPR. However, we cannot achieve those two goals when the diseased group distribution and the healthy group distribution partially overlap. There will be a trade-off between larger TPR and larger FPR versus smaller TPR and smaller FPR.

The skill score combines the disease rate and cost of misclassification. Based on the Skill score, the skillful region of FPR varies according to the disease rates and

loss function values. Simulation studies showed that, when the disease rate is equal to the loss function, the whole region of the FPR is skillful. This will imply that PAUC over the skillful region equals the AUC. If the loss function is larger than the disease rate, the skillful region of FPR will have larger FPR and larger TPR. On the other hand, the loss function is smaller than the disease rate, the skillful region of FPR will have smaller FPR and smaller TPR.

Dodd and Pepe (2003) suggested the PAUC estimator is a more clinically relevant summary measure of accuracy, but the choice of the appropriate restricted region of the FPR may be controversial. Based on our simulations, we gave a reasonable explanation for how to choose the skillful region of FPR for the PAUC given the disease rates and loss function values.

2.6 Osteoporosis Study

Our motivating example involved DPXA osteoporosis data from 5,662 elderly women (Hans et al. 1996). Bone mineral density (BMD) is the "gold standard" in detecting osteoporosis, obtained by dual photon X-ray absorptiometry (DPXA) technology. This device takes a measurement directly at the femoral bone. A new ultrasonic device has been designed to measure broad and ultrasound attenuation (BUA) through the bone. Using the BUA measure, scientists can garner information not only on the bone density but also on its architecture and elasticity. Therefore, this ultrasound device can help detect osteoporosis in elderly women and hence can help prevent hip fractures, a major adverse outcome frequently resulting from osteoporosis. The particular device under investigation takes measurements at the heel of a patient. The new measurement is easier and more economical for patient.

Expert analysis can model osteoporosis as a function of multiple factors via statistical learning techniques (e.g., discriminant analysis, tree classifiers, neural networks). However, diagnostic tests must sometimes rely on single markers to present a simple and effective diagnostic tool for practitioners. Given this approach, we compare markers BMD and BUA for this task. We first compare the performance of markers BMD and BUA by the Skill score from Section 2.3. Then we study the relationship between diseased and healthy groups to construct the Skill scores for markers BMD and BUA. We use the KG model and mean-shift cases as examples.

2.6.1 Hip Fracture Data

The EPIDOS study group was recruited for a 2-year follow-up study between January 1992 and January 1994. There were 115 fractures recorded during two years (a 2.07% rate), and Figure 14 shows density plots for BMD, and BUA scores grouped by whether fractures occurred or not. We used the R function `density` to generate

kernel density estimates with a Gaussian smoothing kernel and the default (rule-of-thumb) bandwidth selection.

2.6.1.1 Fracture or No-Fracture Sample Test

From Table 51, both tests suggest that the two populations vary similarly and have significantly different mean values. Fractures, in these cases, occurred at lower values of the markers.

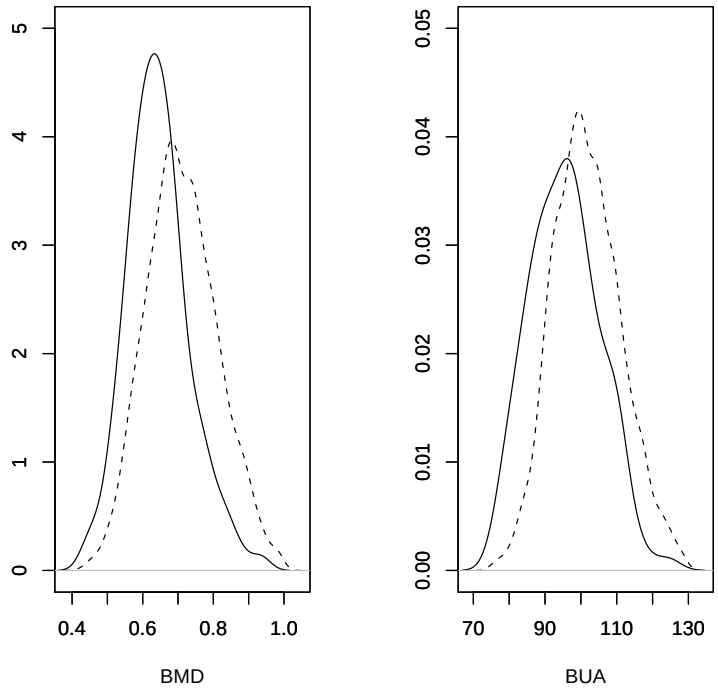


Figure 14: Density Plots. The solid line represents the hip fractures group and the dashed line is the non-hip-fracture group.

2.6.1.2 ROC curves

Figure 15 shows the ROC curves for the fracture data for two index of BMD and BUA. Two ROC curves are intersect at FPR about 0.2. When FPR less than 0.2, the ROC curve of the marker BUA is more concave than the ROC curve of the marker BMD,

Table 51: Two-Sample Test Results for Two Different Measures (BMD, and BUA) for Fracture Occurred or not.

		Equal Variance Test	Two Samples' Mean Test
Hypothesis		$H_0 : \sigma_0^2/\sigma_1^2 = 1$	$H_0 : \mu_0 - \mu_1 = 0$
BMD	95% CI	(1.02, 1.67)	(0.0507, 0.0835)
BUA	95% CI	(0.76, 1.25)	(4.4191, 7.9886)

otherwise the ROC curve of the marker BMD is more concave than the ROC curve of the marker BUA. The AUC statistic from marker BMD is 0.6963, and for testing the hypothesis of equal distributions (i.e., a 45° curve), the p-value corresponding to the Mann-Whitney test is 444141.5 ($5.38e^{-13}$). From DeLong et al. (1988), a variance estimator is:

$$Var(AUC) \doteq Var(P(Z_0 \geq Z_1))/n_1 + Var(P(Z_1 \geq Z_0))/n_0.$$

This method is intuitive and has reasonable results for variance. The 95% CI for the AUC is calculated to be (0.6113, 0.7812); the p-value is $2.95e^{-6}$. The AUC statistic from marker BUA is 0.6751 (0.5886, 0.7615); the p-value is $3.64e^{-5}$. While both AUC statistics are highly significant (in terms of p-value) for detecting differences in the populations, they are not significantly different from each other.

2.6.1.3 The Skill Score and Other Statistics

Because BMD represents the standard, we use it as our benchmark for the diagnosis of hip fracture probability and consider the corresponding skill plots. We consider three distinct values for θ: 0.01, 0.02 and 0.1. The results are plotted in Figure 16. Figure 17 shows only the skill plot and its confidence interval bounds under loss function $\theta = 0.01$, along with the density estimates of the two groups. The Skill plot has random movement (some variations) because the relationship between TPR and FPR changed over the cutoff points and we used empirical density functions to

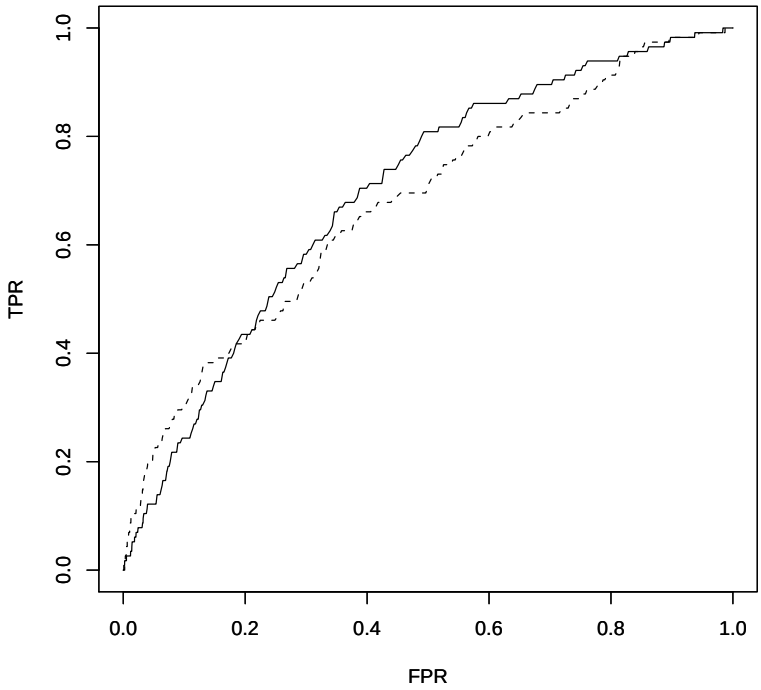


Figure 15: ROC Curves for BMD, and BUA for Hip Fractures. The solid line represents BMD and the dashed line is BUA.

estimate the Skill Score. It could be adjusted by using a local Kernel Smoother. From the Skill plot, we see that the test is skillful corresponding to values of $c \geq 0.688$. The skillful region ($\hat{K}(c) \geq 0$) corresponds to the set FPRs $\in (0.4278, 1)$. The skillful region of FPR corresponding to PAUC is:

$$A(0.4278, 1) = \int_{0.4278}^{1} ROC(t)dt = 0.5136.$$

To estimate the variance of PAUC based on DeLong et al. (1988),

$$\begin{aligned} Var(A(t_0, t_1)) &= Var(P(Z_0 \geq Z_1, Z_0 \in (u_0, u_1)))/n_1 + Var(P(Z_1 \geq Z_0, Z_0 \in (u_0, u_1)))/n_0 \\ &\approx A(t_0, t_1)(1 - A(t_0, t_1))/n_1 + A(t_0, t_1)(1 - A(t_0, t_1))/n_0, \end{aligned}$$

where $t_0 = P(X \geq u_1 | Y = 0), t_1 = P(X \geq u_0 | Y = 0)$. The corresponding 95% confidence interval of the PAUC is $(0.4213, 0.6059)$; the p-value is 0.0128. The statistic of no information of diagnosis test under FPR $\in (0.4278, 1)$ is 0.4085, which is not included in the 95% confidence interval. Therefore, the PAUC of BMD rejects the no information hypothesis under $\theta = 0.01$ and significance level 0.05.

The odds of the PAUC (Λ) defined in Eq.(27) equals 8.7675. The odds of no information of diagnosis test under FPR $\in (0.4278, 1)$ is 2.4953. Transforming the interval estimation method for the PAUC described earlier, a 95% confidence interval for the odds of PAUC is $(2.7925, \infty)$. Therefore, the odds of PAUC of BMD reject the no information assumption.

Although the BUA provides information about the bone micro-architecture and elasticity that are not provided by the BMD, the PAUC is not significant. The BUA skill plots are plotted in Figure 18, and Figure 19 shows only the skill plot for $\theta = 0.01$, From the Skill plot, we see that the test is skillful corresponding to values of $c \geq 103.4$. The skillful region corresponds to the set FPRs $\in (0.5819, 1)$. The skillful region of FPR corresponding to PAUC is:

$$A(0.5819, 1) = \int_{0.5819}^{1} ROC(t)dt = 0.3799.$$

The corresponding 95% confidence interval of the PAUC is $(0.2902, 0.4695)$; the p-value is 0.1410. The statistic of no information of diagnosis test under FPR $\in (0.5819, 1)$ is 0.3308, which is included in the 95% confidence interval. Therefore, the PAUC of BUA did not reject the no information hypothesis under $\theta = 0.01$ and significance level 0.05.

The odds of the PAUC (Λ) defined in Eq.(27) equals 9.9459. The odds of no information of diagnosis test under FPR $\in (0.5819, 1)$ is 3.7840. Transforming the interval estimation method for the PAUC described earlier, a 95% confidence interval for the odds of PAUC is $(2.708, \infty)$. Therefore, the odds of PAUC of BUA did not reject the no information assumption. Figure 16 shows the skillful region of FPR under $\theta = 0.01$.

When we use the fixed region of FPR from Dodd and Pepe (2003), for example $(t_0, t_1) = (0, 0.2)$. Figure 21 shows the fixed region of FPR, (0,0.2). From the ROC curves in Figure 21, the PAUC of BUA (0.0591) is a little higher than the PAUC of BMD (0.0487).

2.6.1.4 Optimal Cutoff Point

In Section 1.1, we describe the choice of the cutoff points in general. By applying the Skill score, the maximum Skill score will be another good choice for the optimal cutoff point. Here we compare two rules to find the optimal cutoff point with the maximum Skill score: the overall misclassification probability, and the expected cost of misclassification.

From the rule of the overall misclassification probability, setting the slope of the ROC curve equal to $(1 - p)/p$ leads us to find the optimal cutoff point. Since the

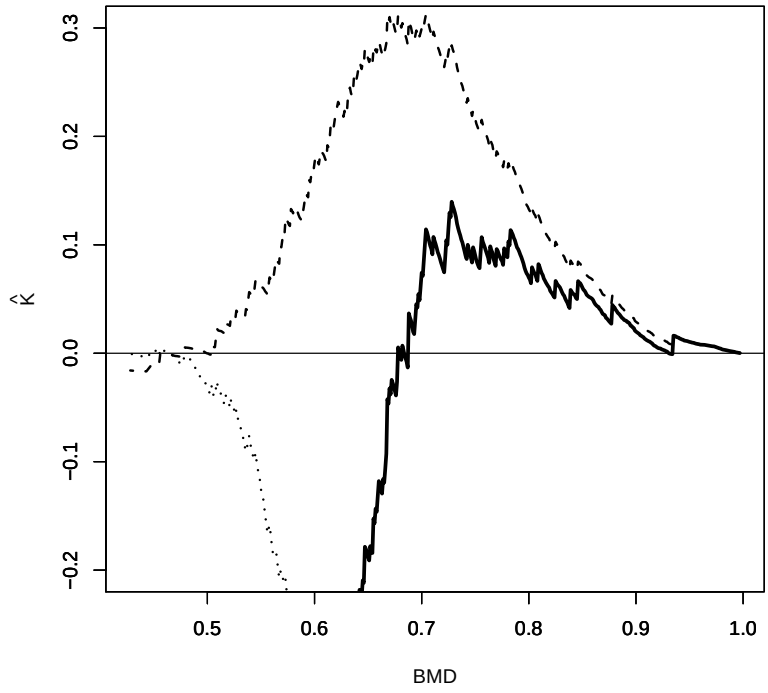


Figure 16: The Skill Plot of BMD for Hip Fractures under θ= 0.01, 0.02 and 0.1. The wide solid line represents θ= 0.01, the dashed line is θ= 0.02 and the dotted line is θ= 0.1.

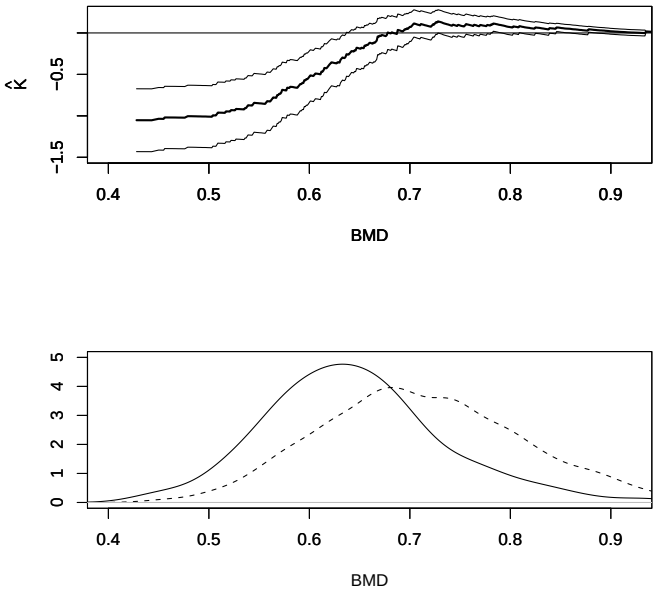


Figure 17: The Skill Plot and 95 % CI and Density Plots of BMD for Hip Fractures under θ= 0.01. The wide solid line is the Skill score and the solid lines are the confidence intervals in the upper plot. The solid line represents hip fractures group and the dashed line is non-hip-fracture group in the lower plot.

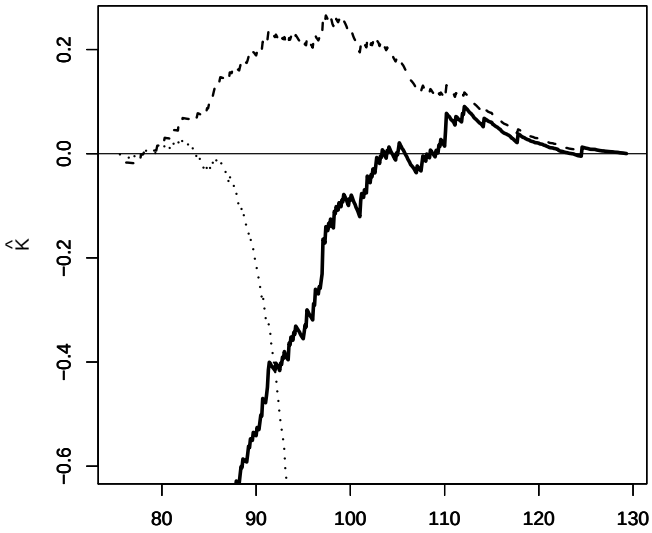


Figure 18: The Skill Plot of BUA for Hip Fractures under θ= 0.01, 0.02 and 0.1. The wide solid line represents θ= 0.01, the dashed line is θ= 0.02 and the dotted line is θ= 0.1.

fracture rate is 0.0207, $(1-\hat{p})/\hat{p} = 48.235$. Table 52 shows five possible optimal cutoff points that corresponding slope of BMD's ROC curve to 48.235. Observe that there is more than one optimal cutoff point.

Table 52: The Optimal Cutoff Points of BMD under the Rule of the Overall Misclassification Probability.

Cutoff Point	FPR	TPR
0.455	0.0022	0.0087
0.503	0.0119	0.0261
0.507	0.0139	0.0348
0.516	0.0184	0.0522
0.537	0.0324	0.0870

Table 53 shows five possible optimal cutoff points that corresponding slope of BUA's ROC curve to 48.235. As with BMD, here there is more than one optimal cutoff point.

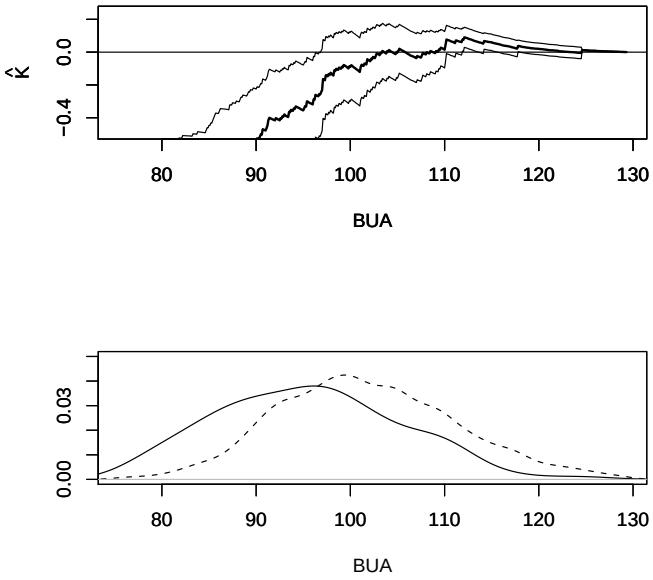


Figure 19: The Skill Plot and 95% CI and Density Plots of BUA for Hip Fractures under θ= 0.01. The wide solid line is the Skill score and the solid lines are the confidence intervals in the upper plot. The solid line represents the hip fracture group and the dashed line is non-hip-fracture group in the lower plot.

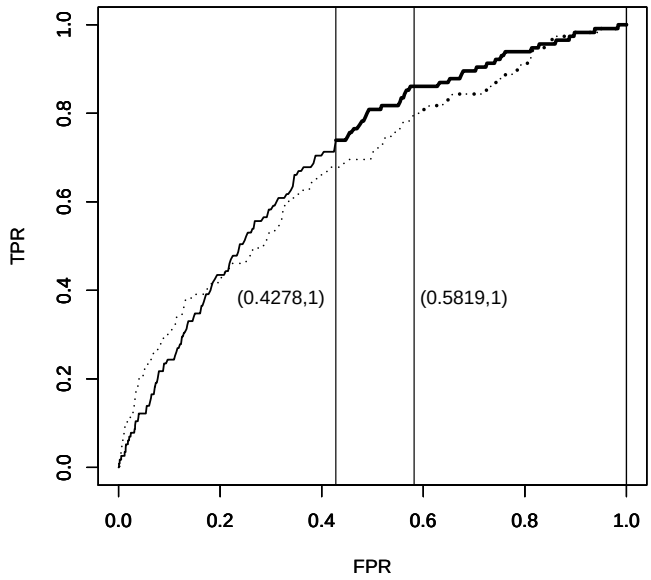


Figure 20: ROC Curve of the BMD (the solid line) and its Skillful Region (the wide solid line) and ROC Curve of the BUA (the dotted line) and its Skillful Region (the wide dotted line) for θ=0.01.

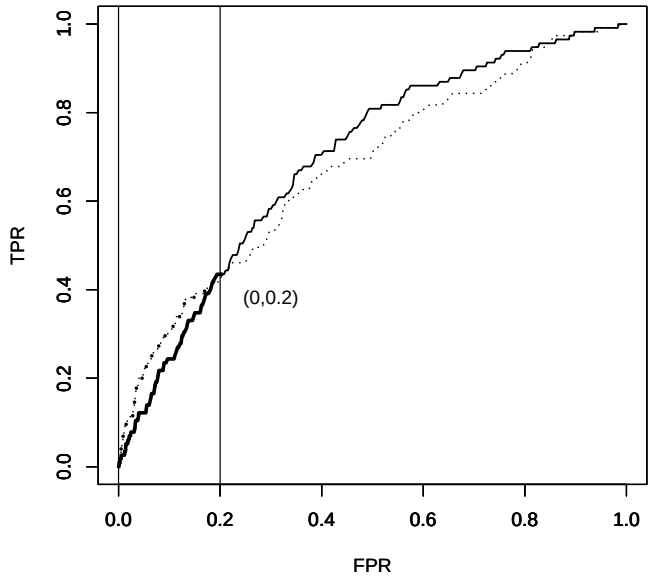


Figure 21: ROC Curve of the BMD (the solid line) and BUA (the dotted line) and Fix Region of the FPR (BMD is the wide solid line, and BUA isthe dotted line).

Table 53: The Optimal Cutoff Points of BUA under the Rule of the Overall Misclassification Probability.

Cutoff Point	FPR	TPR
78.0	0.0032	0.0174
79.4	0.0047	0.0261
82.1	0.0124	0.0870
85.3	0.0323	0.1478
124.5	0.9870	0.9913

From the rule of the expected cost of misclassification, setting the slope of the ROC curve equal to $\theta(1-p)/p/(1-\theta)$ leads us to find the optimal cutoff point. Therefore $\theta(1-\hat{p})/\hat{p}/(1-\theta) = 0.48$, under $\theta = 0.01$. Both BMD and BUA cannot find the corresponding slope of its ROC curve to 0.48. There is no solution under the rule of the expected cost of misclassification.

Two general rules of deciding the optimal cutoff point gave us either more than one optimal cutoff points or no solution. Here we try to use the Skill score to help find the optimal cutoff point. The BMD also achieves a greater maximum Skill score than the BUA. At $c_0 = 0.73$, the skill score for the BMD is maximized with $\hat{K}_{\hat{p}=0.02,\theta=0.01}(c_0) = 0.1397$. A 95% confidence interval for $K_{p=0.02,\theta=0.01}(c_0)$, based on the asymptotic results from Section 3.1, is (-0.0006, 0.2800). The maximum skill score of BMD is not very significant. Note that the FPR and TPR corresponding to this cutoff point ($c_0 = 0.728$) are 0.5747 and 0.8609, respectively.

For the BUA, the skill score is maximized at $c_1 = 112.1$, where $\hat{K}_{\hat{p}=0.02,\theta=0.01}(c_1) = 0.0907$ with a narrower 95% confidence interval of (0.0294, 0.1519). The maximum skill score of BUA is statistically significant. For the BUA, both the FPR (0.8558) and TPR (0.9739) are higher than the corresponding BMD rates. Searching the skillful region of FPR, 124.5 is one of the optimal cutoff points under the rule of the expected cost of misclassification and also inside the skillful region of BUA. Therefore, we can

distinguish 124.5 as better than other possible optimal cutoff points under the rule of the expected cost of misclassification.

2.6.2 Relationship between Hip Fracture and Non-Fracture Groups

While both the BMD and BUA exploit noticeable differences between patients who experienced hip fractures and those who do not, they are dissimilar in the way they communicate this result. Besides the differences between FPR and TPR, there is another important difference between how the two measures relate the two populations. As it turns out, a more interesting structure can be identified using the BUA measurements. This structure is revealed with the KG model, which fits the BUA measurements but not those for the BMD. We showed the results in Section 2.6.2.1. Another special case is the mean-shift model between hip fracture and non-hip-fracture groups in Section 2.6.2.2.

2.6.2.1 KG model

With $F(c) \leq G(c)$, the KG model stipulates that $F(c) = G^{\beta}(c)$ for some $\beta \geq 1$. The measure of KG model fit can be based on first estimating β at different marker values (c) in $G_{n_1}(c) = F_{n_0}(c)^{1/\hat{\beta}}(c)$, where

$$\hat{\beta}(c) = \frac{log\left(F_{n_0}(c)\right)}{log\left(G_{n_1}(c)\right)}.$$

A (1-α)% CI for $\beta(c)$, derived by Hollander et al. (2003), is constructed as

$$\hat{\beta}(c) \pm \frac{Z_{\alpha/2}}{\sqrt{n_1}} \sqrt{\frac{\left(-log(F_{n_0}(c))\right)^2 \cdot \hat{\sigma}_G^2(c)}{\left(-log(G_{n_1}(c))\right)^4} + \frac{\hat{\sigma}_F^2(c)}{\left(-log(G_{n_1}(c))\right)^2}},$$

where

$$\hat{\sigma}_G^2(c) = \sum_{c_i < c} \frac{I(Z_1(c_i) > 0)}{Z_1^2(c_i)} \quad \text{and} \quad Z_1(c_i) = \sum_{i=1}^{n_1} I(Z_{1i} \geq c_i)$$

are the number of subjects at risk before c_i. Also,

$$\hat{\sigma}_F^2(c) = \sum_{c_i<c} \frac{I(Z_0(c_i) > 0)}{Z_0^2(c_i)}, \quad \text{and} \quad Z_0(c_i) = \sum_{i=1}^{n_0} I(Z_{0i} \geq c_i).$$

If we plug in the BMD data, for example, we estimate β as

$$\hat{\beta} = \frac{\Sigma_i^{n_1} \hat{\beta}(c_i)}{n_1} = 2.398,$$

with a 95% CI for β of (2.263, 2.533). The left side of Figure 22 shows the EDFs for BMD measurements associated with hip fractures G_{n_1} as well as the non-fracture set F_{n_0}. In this case, $\hat{G}(c)$ is the KG model estimator based on $F_{n_0}^{1/2.398}(c)$. On the right side of Figure 22, the same plots are shown for the BUA measurements, where $\hat{\beta} = 2.224$ and a 95% CI for β is (2.094, 2.354).

What the graph implies is that the shape of G_{n_1} and $\hat{G}$ are similar for the BUA measurements, but not for the BMD measurements. This, of course, supports the application of the KG model for the BUA. The Q-Q plots in Figure 23 lend further support to this claim. While the underlying EDFs look dissimilar on the left-hand Q-Q plot, the plot on the right suggests a better fit between $G_{n_1}(c)$ and $\hat{G}(c) = F_{n_0}^{\hat{\beta}^{-1}}(c)$.

Using the KG model information with $\hat{\beta} = 2.224$, the marker BUA test's AUC is 0.6898, a skillful region is $(0.7408, 1)$, and PAUC$(0.7408, 1)$= 0.2453. The corresponding 95% confidence interval of the PAUC is $(0.1658, 0.3248)$; the p-value is 0.3136. The odds of PAUC equals 17.6475 for the BUA. The corresponding 95% confidence interval of the odds of the PAUC is $(1.7766, \infty)$. There is no information for classifying correct groups under the BUA skillful region $(0.7408, 1)$ if PAUC and odds of PAUC are 0.2256 and 6.7161, respectively. Therefore, both the PAUC of BUA and the odds of PAUC of BUA did not reject the no information assumption. One advantage for using KG model is it smooths the Skill plot, showed in Figure 24.

In summary, both markers (BMD and BUA) comparing hip fracture and non-fracture groups are significantly different. However, the BMD marker did not fit

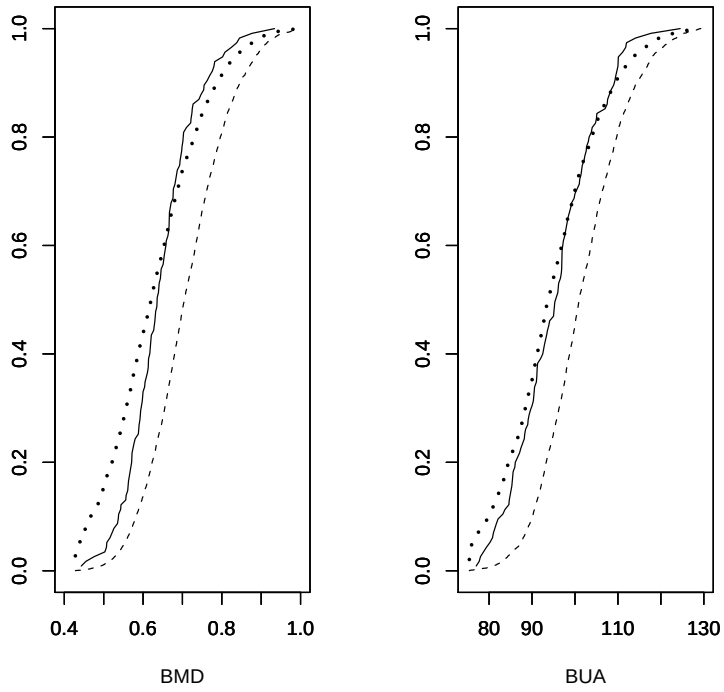


Figure 22: The EDFs of BMD and BUA for Fracture Occurred or not. The dotted line presents the EDF of hip fractures occurred, the dashed line is no hip fractures occurred, and the solid line is using KG model to estimate EDF of hip fractures occurred group. For BMD, $\hat{G}(c) = F_{n_0}^{1/2.398}(c)$. For BUA, $\hat{G}(c) = F_{n_0}^{1/2.224}(c)$.

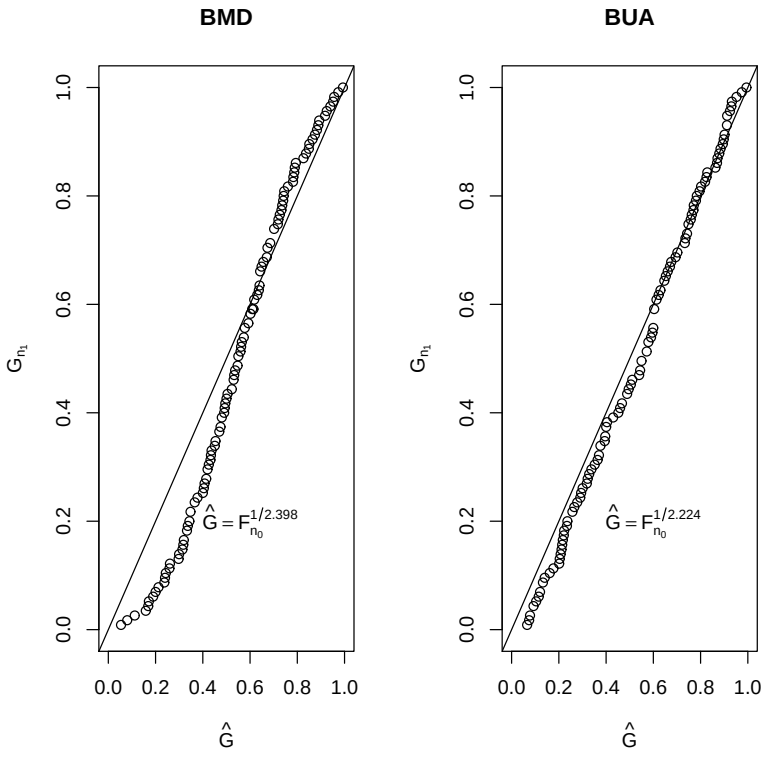


Figure 23: Q-Q Plot for BMD and BUA.

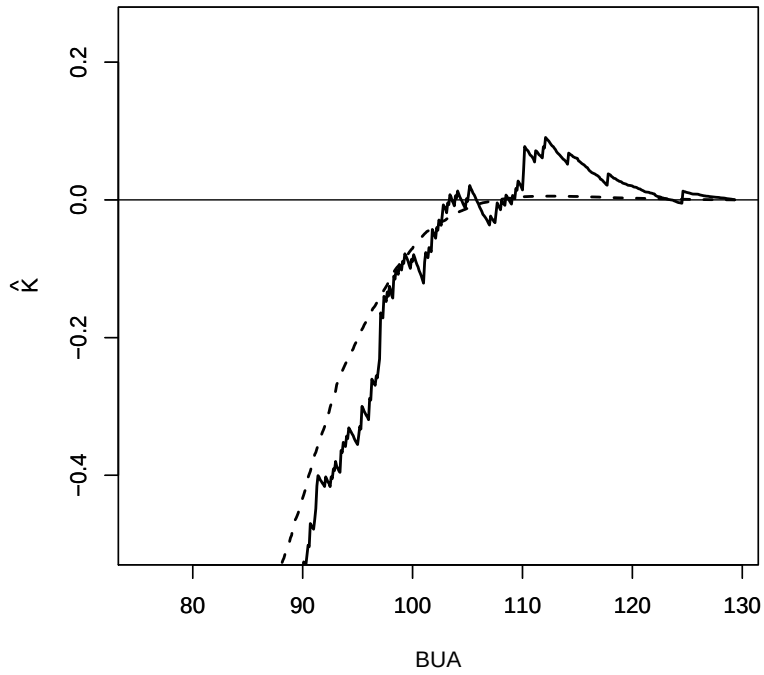


Figure 24: The Skill Plot for BUA (the solid line) Estimating by the Original and KG Model (the dashed line).

well by using the KG model to describe hip fracture distribution from non-fracture groups. BUA is from new ultrasonic device fit well by using the KG model to describe hip fracture distribution from non-fracture groups. By using BUA marker, we have a better model to simply the relationship of hip fracture and non-fracture distributions.

2.6.2.2 The Mean-Shift Case

Figure 14 shows that $F(c)$ and $G(c)$ might have differ by a location shift, $G(c) = F(c+\eta)$, where $\hat{\eta}$ is the difference of the mean of non-disease group and disease group. In the BMD data, we estimate $\hat{\eta}$ as 0.0671. The left side of Figure 25 shows the EDFs for BMD measurements associated with hip fractures G_{n_1} as well as the non-fracture set F_{n_0}. In this case, $\hat{G}(c)$ is the mean-shift estimator based on $F_{n_0}(c + 0.0671)$. On the right side of Figure 23, the same plots are shown for the BUA measurements, where $\hat{\eta} = 6.2038$.

What the graph implies is that the shape of G_{n_1} and $\hat{G}$ are similar for the BUA measurements. It is acceptable for the BMD measurements; there is a difference around the tail part. The Q-Q plots in Figure 25 show the underlying EDFs look dissimilar on the left-hand Q-Q plot, the plot on the right suggests a better fit between $G_{n_1}(c)$ and $\hat{G}(c) = F_{n_0}(c + 6.2038)$.

Using the mean-shift information with $\hat{\eta} = 0.0671$, the marker BMD has the skillful region of FPR $(0.5670, 1)$ and PAUC$(0.5670, 1)$= 0.3997. The corresponding 95% confidence interval of the PAUC is $(0.3092, 0.4902)$; the p-value is 0.0951. The odds of PAUC(Λ) equals 12.003 for the BMD. There is no information for classifying correct groups under the region $(0.5670, 1)$. The corresponding 95% confidence interval of the odds of the PAUC is $(2.499, \infty)$. There is no information for classifying correct groups under the BUA skillful region $(0.5670, 1)$ if PAUC and odds of PAUC are 0.3387 and 3.6189, respectively. Therefore, both the PAUC of BMD and the odds of

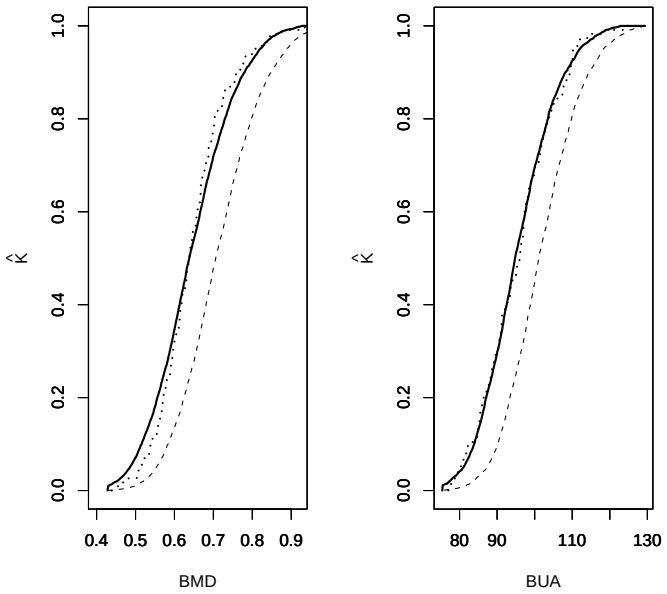


Figure 25: The EDFs of BMD and BUA for Fracture Occurred or not. The dotted line presents the EDF of hip fractures occurred, the dashed line is no hip fractures occurred, and the solid line is a mean-shift special case to estimate EDF of hip fractures occurred group. For BMD, $\hat{G}(c) = F_{n_0}(c + 0.0671)$. For BUA, $\hat{G}(c) = F_{n_0}(c + 6.2038)$.

PAUC of BMD did not reject the no information assumption.

For marker BUA with an estimate $\hat{\eta} = 6.2038$, the skillful PAUC region is $(0.5819, 1)$ and PAUC$(0.5819, 1)$= 0.3772. The corresponding 95% confidence interval of the PAUC is $(0.2877, 0.4667)$; the p-value is 0.1542. The odds of PAUC equals 9.222. The corresponding 95% confidence interval of the odds of the PAUC is $(2.2063, \infty)$. There is no information for classifying correct groups under the region $(0.5819, 1)$ if PAUC and the odds of PAUC are 0.3307 and 3.7835, respectively.

The left side of Figure 26 shows the Skill plot for BMD estimated by the original EDFs and mean-shift model. Although the Q-Q plot of the original EDF and a mean-shift model look similar, the Skill plot for BMD estimated by the original EDFs and a mean-shift model are not a very close match when BMD is larger than 0.65. The right side of Figure 26 shows the Skill plot for BUA estimated by the original and the mean-shift model is a good fit. Compared to Figure 23, the mean-shift model fitting are very similar under using KG model for marker BUA.

In summary, both markers (BMD, and BUA) comparing hip fracture and non-fracture groups are useful measures for the osteopetrosis diagnostics. However, the BUA has the added advantage of fitting a more refined model that relates the disease and non-disease groups.

2.6.3 Summary of Example

We demonstrated how to apply the skill test for the hip fracture or diagnosis. Our intuition and simulation studies suggest that the ROC curve will make more sense if the test is skillful. Similarly, by integrating the ROC curve only where the test is skillful, the AUC statistic might be more sensible than AUC measurements based on arbitrary cutoff points. So there are several conclusions based on our analysis of the osteopetrosis data, as follows.

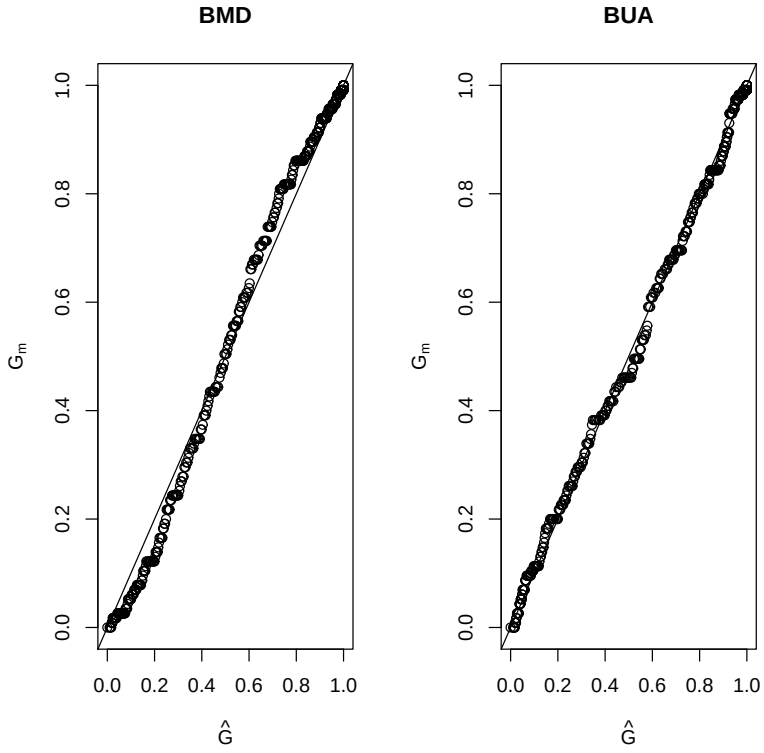


Figure 26: Q-Q Plot for BMD and BUA. For BMD, $\hat{G}(c) = F_{n_0}(c + 0.0671)$. For BUA, $\hat{G}(c) = F_{n_0}(c + 6.2038)$.

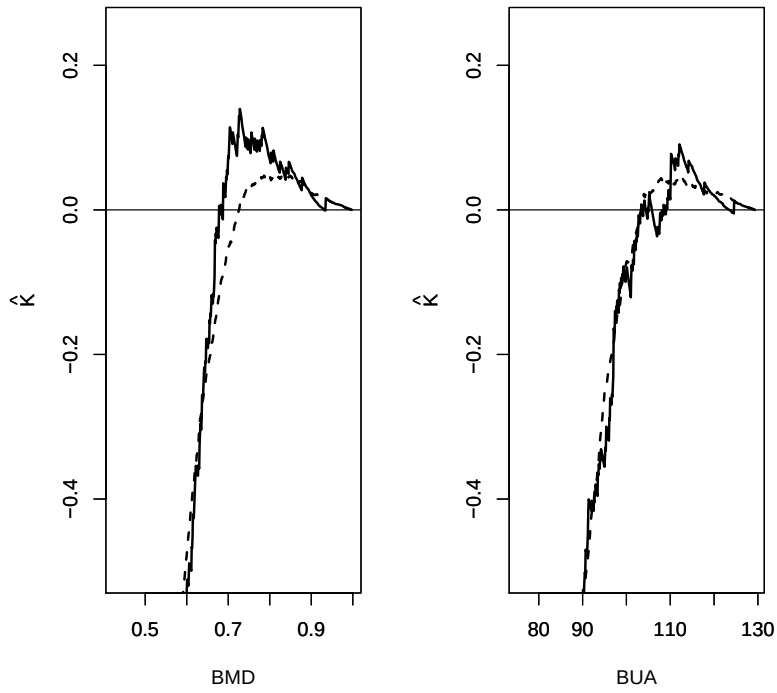


Figure 27: The Skill Plot for BMD and BUA Estimating by the Original and Mean-Shift Models.

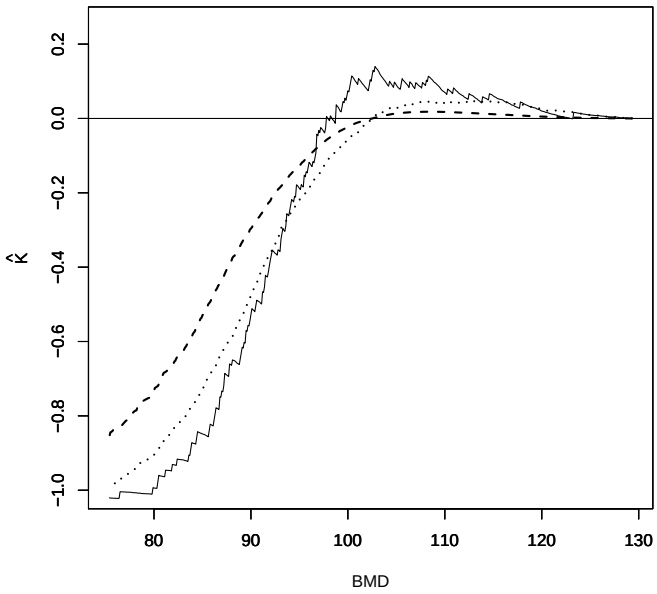


Figure 28: The Skill Plot for BMD Estimated by the Original, KG model and Mean-Shift Model. The solid line presents the original EDFs, the dashed line is the KG model, and the dotted line is the mean-shift model.

The density plots for fracture or no-fracture were overlapped. From the fracture or no-fracture two sample test, it showed the means of both markers, BMD and BUA, are significantly different under significance level $\alpha = 0.05$.

The AUC statistic from the marker BMD is 0.6963 and the 95% CI of AUC of the marker BMD is (0.6113, 0.7812). The AUC statistic from the marker BUA is 0.6751 which is slightly lower than that from the marker BMD. However, it is not significantly different. The 95% CI of AUC of the marker BUA is (0.5886, 0.7615). Note that both AUC statistics are significantly higher than 0.5, which is the the benchmark index of the AUC statistic corresponding to a case for which the FPR is equal to the TPR for any cutoff.

When the FPR is less than 0.2, the ROC curve of BUA is more concave than the ROC curve of BMD; otherwise, the ROC curve of the marker BMD is more concave than the ROC curve of the marker BUA. This means that if we used FPR regions

(0,0.1) or (0,0.2)—these are two suggestions in Dodd and Pepe (2003)—the PAUC of the marker BUA would be higher than the PAUC of the gold standard marker BMD. We recommend using the Skill score to decide the FPR region by considering the misclassification cost (loss).

For the marker BMD, the Skill plot by using the loss function $\theta = 0.01$ showed that the skillful value of BMD is larger than 0.688. The skillful region corresponds to the set FPRs $\in (0.4278, 1)$. The skillful region of FPR corresponding to PAUC is 0.5136. The corresponding 95% confidence interval of the PAUC is $(0.4213, 0.6059)$; the p-value is 0.0128. Therefore, the PAUC of BMD rejects the no information hypothesis under $\theta = 0.01$ and significance level 0.05. The odds of the PAUC of BMD is 8.7675, and a 95% CI for the odds of PAUC is $(2.7925, \infty)$. The odds of PAUC of BMD rejects the no information assumption.

Although the BUA provides information about the bone micro-architecture and elasticity that are not provided by the BMD, the PAUC is not significant. The skillful value of BUA larger than 103.4. The skillful region corresponds to the set FPRs $\in (0.5819, 1)$. The skillful region of FPR corresponding to PAUC is 0.3799. The corresponding 95% confidence interval of the PAUC is $(0.2902, 0.4695)$; the p-value is 0.1410. The PAUC of BUA did not reject the no information hypothesis under $\theta = 0.01$ and significance level 0.05. The odds of the PAUC of BUA is 9.9459, and a 95% confidence interval for the odds of PAUC of BUA is $(2.708, \infty)$ including the odds of no information of diagnosis test under FPR $\in (0.5819, 1)$ is 3.7840. Therefore, the odds of PAUC of BUA did not reject the no information assumption.

There have been several methods to find the optimal cutoff points in the past (Section 2.1.1.). We applied two rules: (1) minimizing the overall misclassification probability, and (2) minimizing the expect cost of misclassification. For rule (1), we found there are not unique solutions for the optimal cutoff points for the marker BMD and BUA. There are five optimal cutoff points for each of the markers BMD

and BUA. For rule (2), we found there is no solution for the optimal cutoff points for the marker BMD and BUA. By using the skillful region of the FPR, we might narrow down the optimal cutoff points in the situations that have more than one optimal cutoff point. For example, 124.5 of the marker BUA is one of the optimal cutoff points under rule (1), and it also is in the skillful region of BUA. Therefore, we can distinguish 124.5 as better than other possible optimal cutoff points under the rule (1). However, we failed to combine rule (1) and the skillful region of BMD to identify the better optimal cutoff points; no optimal value of the BMD is larger than 0.688.

Therefore, the maximum Skill score is a very good choice for deciding the optimal cutoff point. Under $\theta = 0.01$, the marker BMD (0.73) maximized the skill score for the BMD with a value of 0.1397, a 95% CI for the maximized skill score of BMD is (-0.0006, 0.2800). The maximum skill score of BMD is not significantly different from "not skillful," though it is nearly significant. Note that the FPR and TPR corresponding to this cutoff point 0.728 are 0.5747 and 0.8609, respectively.

Under $\theta = 0.01$, the skill score is maximized at 112.1 for the marker BUA, where $\hat{K}_{\hat{p}=0.02,\theta=0.01}(c_1) = 0.0907$ with a narrower 95% CI of (0.0294, 0.1519). The maximum skill score of BUA is significantly different from "not skillful." For the BUA, the FPR and TPR corresponding to this cutoff point 112.1 are 0.8558 and 0.9739, respectively.

While both the BMD and BUA exploit noticeable differences between patients who experienced hip fractures and those who do not, they are dissimilar in the way they communicate this result. Besides the differences between FPR and TPR, there is another important difference between how the two measures relate the two populations. As it turns out, a more interesting structure can be identified using the BUA measurements. This structure is revealed with the KG model, which fits the BUA measurements but not those for the BMD. The structure of mean-shift model between hip fracture and non-hip-fracture groups fits well on both markers BMD and BUA.

Applying the relation between hip fracture and non-hip-fracture groups might allow us the advantage of fitting refined or simplified estimations of the density functions, the ROC curves, the AUC, the skill scores, the PAUC, and the odds of PAUC.

2.7 Concluding Remarks and Future Work

Accurate diagnosis of disease, i.e. correctly determining whether a patient has a disease or not, is a critical technical aspect of health care. Screening and diagnostic tests are familiar and ever-evolving tools of modern medicine. Accurate early detection by screening is essential for early disease treatment.

Here we summarize our theoretical results on ROC curves, AUC, PAUC, and the odds of PAUC. We tried to understand more about accurate diagnosis in this work. We applied the Skill score to evaluate binary events by giving skill/accuracy or value. It is very intuitive that only skillful cutoff points should be used. The Skill plot provided the concept of PAUC, which is a very good idea for how to set the partial region by considering the loss and cost. Asymptotic properties of the Skill score have been provided. Simulation studies of PAUC indicate the advantage of using the Skill score and its unbiasedness and consistency properties. A motivating example from osteoporosis shows the importance of the Skill score in disease diagnosis. We also studied the relationship between diseased and non-diseased groups and how to find the optimal cutoff points for disease diagnosis.

The diagnosis technique used will affect people's decisions greatly by the associated factors and accuracy of the method. There are still very broad areas for further studies:

Diagnostic test results may be much more complicated, involving several components. Do Skill scores have a role to play in determining how to combine different sources of information to optimize diagnostic accuracy? Many papers used (logistic) regression of ROC curves for combining covariate variables of classification (Dodd and Pepe 2003, and Pepe et al. 2006). It will be very useful to consider other covariate variables in the Skill plot for more accurate diagnosis.

Inspired by the second author in Arcones, et al. (2002), we will study the effect of stochastic precedence on the Skill plot. Typically, diagnostic tests assume $F(x) \geq G(x)$, which defines stochastic dominance. This happens to be the stronger stochastic ordering between distributions. Stochastic precedence is a weaker and more realistic order restriction. The interesting topics would be: what if the stochastic inequality is not so strong? Moreover, is the ROC ineffective under this condition?

A new idea for the Skill score of Eq.(18): compare a random guess with an expert forecast, instead of comparing a naive guess with an expert forecast.

REFERENCES

[1] Arcones, M. A., Kvam, P. H., and Samaniego, F. J. (2002), "Nonparametric Estimation of a Distribution Subject to a Stochastic Precedence Constraint," *Journal of the American Statistical Association*, Vol. 97, 170-182.

[2] Atkinson, A. (1969), "The Use of Residuals as a Concomitant Variable," *Biometrika*, Vol. 56, 33-41.

[3] Baker, S. G. (2000), "Identifying Combinations of Cancer Markers for Further Study as Triggers of Early Intervention," *Biometrics*, Vol. 56, 1082-1087.

[4] Bamber, D. (1975), "The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph," *Journal of Mathematical Psychology*, Vol. 12, 387-415.

[5] Bartlett, M. S. (1955), *An Introduction to Stochastic Processes*, Cambridge: Cambridge University Press.

[6] Bartlett, M. S. (1967), "Inference and Stochastic Processes," *Journal of the Royal Statistical Society, Series A*, Vol. 130, 457-478.

[7] Bartlett, M. S. (1968), "A Further Note on Nearest Neighbour Models," *Journal of the Royal Statistical Society, Series A*, Vol. 131, 579-580.

[8] Bartlett, M. S. (1978), "Nearest Neighbour Models in the Analysis of Field Experiments," *Journal of the Royal Statistical Society, Series B*, Vol. 40, 147-174.

[9] Basu, S. and Reinsel, G. C. (1993), "Properties of the Spatial Unilateral First-order ARMA Model," *Advances in Applied Probability*, Vol. 25, 631-648.

[10] Basu, S. and Reinsel, G. C. (1994), "Regression Models with Spatially Correlated Errors," *Journal of the American Statistical Association*, Vol. 89, 88-99.

[11] Beaton, A. E. and Tukey, J. W. (1974), "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," *Technometrics*, Vol. 16, 147-185.

[12] Besag, J. E. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society, Series B*, Vol. 36, 192-236.

[13] Besag, J. E. and Higdon, D (1999), "Bayesian Analysis of Agricultural Field Experiments," *Journal of the Royal Statistical Society, Series B*, Vol. 61, 691-746.

[14] Besag, J. E. and Kempton, R. A. (1986), "Statistical Analysis of Field Experiments Using Neighboring Plots," *Biometrics*, Vol. 42, 231-251.

[15] Besag, J. E. and Kooperberg, C. (1995), "On Conditional and Intrinsic Autoregressions," *Biometrika*, Vol. 82, 733-746.

[16] Briggs, W. M. and Ruppert, D. (2005), "Assessing the Skill of Yes/No Predictions," *Biometrics*, Vol. 61, 799-807.

[17] Briggs, W. M. and Zaretzki, R. (2008), "The Skill Plot: a Graphical Technique for Evaluating Continuous ", *Biometrics*, Vol. 64, 250-256.

[18] Brumback, L. C., Pepe, M. S. and Alonzo, T. A. (2006), "Using the ROC Curve for Gauging Treatment Effect in Clinical Trials," *Statistics in Medicine*, Vol. 25, 575-590.

[19] Bustos, O. H., Fraiman, R., and Yohai, V.J. (1984), "Asymptotic Behavior of the Estimates Based on Residual Autocovariances for ARMA Models," In Robust and Nonlinear Times Analysi, Ed. J. Franke, W. Hardle, and D. Martin, 26-49. New York: Springer-Verlag.

[20] Bustos, O. H. and Yohai, V. J. (1986), "Robust Estimates for ARMA Models," *Journal of the American Statistical Association*, Vol. 81, 155-168.

[21] Cai, T., Pepe, M. S., Zheng, Y., Lumley, T. and Jenny, N. S. (2006), "The Sensitivity and Specificity of Markers for Event Times," *Biostatistics*, Vol. 7, 182-197.

[22] Casella, G. and Berger, R. L. (1990), *Statistical Inference*, Duxbury, Belmont.

[23] Chambless, L. E. and Diao, G. (2006), "Estimation of Time-dependent Area under the ROC Curve for Long-term Risk Prediction," *Statistics in Medicine*, Vol. 25, 3474-3486.

[24] Chang, I., Tiao, G. C., and Chen, C. (1988), "Estimation of Time Series Parameters in the Presence of Outliers," *Technometrics*, Vol. 30, 193-204.

[25] Chellappa, R. and Jain, A. (1993), *Markov Random Fields: Theory and Application*, Academic Press: New York.

[26] Claeskens, G., Jing, B.-Y., Peng, L., and Zhou, W. (2003). "Empirical Likelihood Confidence Regions for Comparison Distributions and ROC Curves," *The Canadian Journal of Statistics*, Vol. 31, 173-190.

[27] Cressie, N. A. C. (1993), *Statistics for Spatial Data*, John Wiley & Sons, Inc.

[28] Cullis, B. R. and Gleeson, A. C. (1991), "Spatial Analysis of Field Experiments–an Extension to Two Dimensions," *Biometrics*, Vol. 47, 1449-1460.

[29] Cullis, B. R., Lill, W. J., Fisher, J. A., Read, B. J., and Gleeson, A. C. (1989), "A New Procedure for the Analysis of Early Generation Variety Trials," *Applied Statistics*, Vol. 38, 361-375.

[30] Davis J. C., Gyurcsik, R. S., Lu, J.-C., and Hughes-Oliver J.M. (1996), "A Robust Metric for Measuring Within-Wafer Uniformity." *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part C*, Vol. 19, 283-289.

[31] Denby, L. and Martin, R. D. (1979), "Robust Estimation of the First-Order Autoregressive Parameter," *Journal of the American Statistical Association*, Vol. 74, 140-146.

[32] Dodd, L. E. and Pepe, M. S. (2003), "Partial AUC Estimation and Regression," *Biometrics*, Vol. 59, 614-623.

[33] Eicker, F. (1963), "Central Limit Theorems for Families of Sequences of Random Variables," *The Annals of Mathematical Statistics*, Vol. 34, 439-446.

[34] Eicker, F. (1966), "A Multivariate Central Limit Theorem for Random Linear Vector Forms," *The Annals of Mathematical Statistics*, Vol. 37, 1825-1828.

[35] Fox, A. J. (1972), "Outliers in Time Series," *Journal of the Royal Statistical Society, Series B*, Vol. 43, 350-363.

[36] Gleeson, A. C. and Cullis, B. R. (1987), "Residual Maximum Likelihood (REML) Estimation of a Neighbour Model for Field Experiments," *Biometrics*, Vol. 43, 277-287.

[37] Grau, E. A. (2000), *Robust Estimation of Autocorrelation Parameters in the AR(1)×AR(1) model*, PhD Thesis, North Carolina State University, Raleigh, NC.

[38] Grau, E. A., and Lu, J.-C. (2004), *Robust Estimation in a Spatial Unilateral Autoregressive model*, Technical Report, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

[39] Green, D. M, and Swets, J. A. (1966), *Signal Detection Theory and Psychophysics*, New York: Wiley.

[40] Hall, P., Hyndman, R. J. and Fan, Y. (2004), "Nonparametric Confidence Intervals for Receiver Operating Characteristic Curves," *Biometrika*, Vol. 91, 743-750.

[41] Hanley, J. A. and McNeil, B. J. (1982), "The Meaning and Use of the Area under a Receiver Operating Characteristic(ROC) Curve," *Radiology*, Vol. 143, 29-36.

[42] Hans, D., Dargent-Molina, P., Schott, A. M., Sebert, J. L., Cormier, C., Kotxki, P. O., Delmas, P. D., Pouilles, J. M., Breart, G., and Meunier, P. J. (1996), "Ultrasonographic Heel Measurements to Predict Hip Fracture in Elderly Women: the EPIDOS Prospective Study," *The Lancet*, Vol. 348, 511-514.

[43] Heckerling, P. S. (2001), "Parametric Three-way Receiver Operating Characteristic Surface Analysis Using Mathematica," *Medical Decision Making*, Vol. 21, 409-417.

[44] Hollander, M., Laird, G. and Song, K.-S. (2000), "Nonparametric Inference for the Proportionality Function in the Random Censorship Model," *Journal of Nonparametric Statistics*, Vol. 15, 151-169.

[45] Hsieh, F. and Turnbull, B. W. (1996), "Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve," *The Annals of Statistics*, Vol. 24, 25-40.

[46] Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, Vol. 35, 73-101.

[47] Hughes-Oliver, J. M., Lu, J.-C., Davis, J. C., and Gyurcsik, R. S. (1998), "Achieving Uniformity in a Semiconductor Fabrication Process Using Spatial Modeling," *Journal of the American Statistical Association*, Vol. 93, 36-45.

[48] Jager, G. (2001), "Determining the Area under the Receiver Operating Characteristic Curve for a Binary Diagnostic Test: Best Is Not Always Ideal," *Medical Decision Making*, Vol. 21, 512-514.

[49] Jureckova J. and Sen P. K. (1996), *Robust Statistical Procedures: Asymptotics and Interrelations*, John Wiley & Sons, Inc.

[50] Kempton, R. A. and Howes, C. W. (1981), "The Use of Neighboring Plot Values in the Analysis of Variety Trials," *Applied Statistics*, Vol. 30, 59-70.

[51] Koziol, J. A. and Green, S. B. (1976), "A Cramer-von Mises Statistic for Randomly Censored Data," *Biometrika*, Vol. 63, 465-474.

[52] Lange, K. L, Little, R. J. A., and Taylor, J. M. G. (1989), "Robust Statistical Modeling Using the T Distribution", *Journal of the American Statistical Association*, Vol. 84, 881-896.

[53] Lee, C.-S. and Rawlings, J. O. (1982), "Design of Experiments in Growth Chambers - Uniformity Trials in the North Carolina State University Phytotron," *Crop Science*, Vol. 22, 551-558.

[54] Li, S. Z (1995), *Markov Random Field Modeling in Computer Vision*, Springer-Verlag.

[55] Little, R. J. A. and Robin, D. B. (2002), *Statistical Analysis with Missing Data*, John Wiley & Sons, Inc.

[56] Lloyd, C. J. and Yong Z. (1999), "Kernel Estimators of the ROC Curve Are Better Than Empirical," *Statistics and Probability Letters*, Vol. 44, 221V228.

[57] Martin, R. J. (1979), "A Subclass of Lattice Processes Applied to a Problem in Planar Sampling," *Biometrika*, Vol. 66, 209-217.

[58] Martin, R. J. (1982), "Some Aspects of Experimental Design and Analysis when Errors Are Correlated," *Biometrika*, Vol. 69, 597-612.

[59] Martin, R. J. (1990), "The Use of Time-series Models and Methods in the Analysis of Agricultural Field Trials," *Communications in Statistics–Theory and Methods*, Vol. 19, 55-81.

[60] Martin, R. J. (1996), "Some Results on Unilateral ARMA Lattice Processes," *Journal of Statistical Planning and Inference*, Vol. 50, 395-411.

[61] McIntosh, M. W. and Pepe, M. S. (2002), "Combining Several Screening Tests: Optimality of the Risk Score," *Biometrics*, Vol. 58, 657-664.

[62] Mesenbrink, P., Lu, J.-C., McKenzie, R., and Taheri, J. (1994), "Characterization and Optimization of a Wave-Soldering Process," *Journal of the American Statistical Association*, Vol. 89, 1209-1217.

[63] Mossman, D. (1999), "Three-way ROCs," *Medical Decision Making*, Vol. 19, 78-89.

[64] Mozer, J. B. and Briggs, W. M. (2003), "Skill in Real-time Solar Wind Shock Predictions," *Journal of Geophysical Research*, Vol. 108, 1-9.

[65] Muijtjens, A. M. M., Van Luijk, S. J. and Van der Vleuten, C. P. M. (2006), "ROC and Loss Function Analysis in Sequential Testing," *Advances in Health Sciences Education*, Vol. 11, 5-17.

[66] Neter, J., Kutner, M. H., Wasserman, W., and Nachtsheim, C. J. (1996), *Applied Linear Regression Models*, McGraw-Hill/Irwin Inc.

[67] Ojeda, S. M., Vallejos, R. O., and Lucini, M. M. (2002), "Performance of Robust RA Estimator for Bidimensional Autoregressive Models," *Journal of Statistical Computation and Simulation*, Vol. 72, 47-62.

[68] Papadakis (1937), "Methode Statistique Pour Des Experiences Sur Champ. Institut d'Amelioration des Plantes a Thessaloniki (Grece)," *Bulletin Scientifique*, No. 23.

[69] Pepe, M. S. (2000), "An Interpretation for the ROC Curve and Inference Using GLM Procedures," *Biometrics*, Vol. 56, 352-359.

[70] Pepe, M. S. (2000), "Receiver Operating Characteristic Methodology," *Journal of the American Statistical Association*, Vol. 95, 308-311.

[71] Pepe, M. S. (2003), *The Statistical Evaluation of Medical Tests for classification and Prediction*, Oxford University Press: New York.

[72] Pepe, M. S., Cai, T. and Longton, G. (2006), "Combining Predictors for Classification Using the Area under the Receiver Operating Characteristic Curve," *Biometrics*, Vol. 62, 221-229.

[73] Pepe, M. S. and Longton, G. (2005), "Standardizing Diagnostic Markers to Evaluate and Compare Their Performance," *Epidemiology*, Vol. 16, 598-603.

[74] Pretty, I. A. and Maupome, G. (2004), "A Closer Look at Diagnosis in Clinical Dental Practice: Part 2. Using Predictive Values and Receiver Operating Characteristics in Assessing Diagnostic Accuracy," *Journal of the Canadian Dental Association*, Vol. 70, 313-316.

[75] Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, John Wiley & Sons, Inc.

[76] Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields, Theory and Applications*, Chapman & Hall.

[77] Serfling, R. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, Inc.

[78] Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005), "ROCR: Visualizing Classifier Performance in R," *Bioinformatics*, Vol. 21, 3940-3941.

[79] Swets, J. A. and Pickett, R. M. (1982), *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*, Academic Press: New York.

[80] Whittle, P. (1954), "On Stationary Processes in the Plane," *Biometrika*, Vol. 41, 434-449.

[81] Zimmerman, D. L. and Harville, D. A. (1989), "On the Unbiasedness of the Papadakis Estimator and Other Nonlinear Estimators of Treatment Contrasts in Field-Plot Experiments," *Biometrika*, Vol. 76, 253-259.

[82] Zimmerman, D. L. and Harville, D. A. (1991), "A Random Field Approach to the Analysis of Field-Plot Experiments and other Spatial Experiments," *Biometrics*, Vol. 47, 223-239.

[83] Zou, K. H., Hall, W. J., and Shapiro, D. E. (1997), "Smooth Non-parametric Receiver Operating Characteristic (ROC) Curves for Continuous Diagnostic Tests," *Statistics in Medicine*, Vol. 16, 2143-2156.