2012

Statistical methods in disease risk analysis, disease testing and nutrition epidemiology

Hui Lin
Iowa State University

Statistical methods in disease risk analysis, disease testing and nutrition epidemiology

by

Hui Lin

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Alicia Carriquiry, Co-major Professor

Chong Wang, Co-major Professor

Kenneth Koehler

Dan Nordman

Derald Holtkamp

Iowa State University

Ames, Iowa

2013

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and writing this thesis. First and foremost, Dr. Alicia Carriquiry and Dr. Chong Wang for their guidance, patience and support throughout this research and the writing of this thesis. I would additionally like to thank Dr. Derald Holtkamp for his help and encouragement throughout my graduate study. I would also like to thank Dr. Kenneth Koehler and Dr. Dan Nordman for their effort and time. I would also like to thank my friends and family for their loving guidance and financial assistance during my school life. Without those people I would not have been able to complete this work.

ABSTRACT

The thesis is composed of three separated projects: disease risk scoring systems (chapter 2), statistical tests for proportion difference in one-to-two matched binary data (chapter 3) and bivariate measurement error model for nutrition epidemiology (chapter 4). In the first project, we propose to use group lasso algorithm for logistic regression to construct a risk scoring system for predicting disease in swine. We choose the penalty parameter for the group lasso through leave-one-out cross validation and use the area under the receiver operating characteristic curve as criterion. We show our proposed scoring system is superior to existing methods. The second project was originally motivated by the pooling of diagnostic tests. We proposed exact and asymptotic tests for one-to-two matched binary data. Unlike other existing methods, our procedure doesn't rely on a mutual independence assumption. The emphasis on dependence among observations from the same matched set is natural and appealing, as much in human health as it is in veterinary medicine. It can be applied to many kinds of diagnostic studies with a one-to-two matched data structure. Our method can also be generalized to one-to-N matched case in a straightforward manner. In the third paper we consider the problem of estimating the joint distribution of two correlated random variables where one of the variables is observed with error. DKM is first used to adjust the univariate measurement error. A Gaussian copula is then used to model the correlation structure between the two variables after error adjustment.

CHAPTER 1. GENERAL INTRODUCTION

Statistics plays an important role in the design, analysis and interpretation of studies related to medicine and public health. In this dissertation, we develop statistical tools to explore health related problems, estimate associations between risk factors and disease and draw appropriate conclusions based on possibly messy data. More specifically, the thesis is composed of three papers that discuss disease risk scoring systems, statistical tests for proportion differences in one-to-two matched binary data and bivariate measurement error model for nutrition epidemiology. While we consider specific areas of application, the methods we propose can be applied in multiple areas.

The first project (Chapter 2) is motivated by the need to develop a risk scoring system from survey data on risk factors for porcine reproductive and respiratory syndrome (PRRS). This is a disease with major impact on pork production and can be a serious financial problem for swine producers. We propose to use group lasso algorithm for logistic regression to construct a risk scoring system for predicting disease in swine. Group lasso provides an attractive approach to this research question because of its ability to achieve group variable selection and stabilize parameter estimates at the same time. We choose the penalty parameter for the group lasso through leave-one-out cross validation, using the criterion of the area under the receiver operating characteristic curve. We show that our proposed scoring system is superior to existing methods.

The second project (Chapter 3) was originally motivated by the common practice of pooling diagnostic tests. Matched observations with dichotomous responses commonly occur in medical and epidemiological studies. Although standard approaches exist for one-to-one paired binary data analyses, not much work has been produced for the case where we have one-to-two or one-to-N matched binary data. The existing Miettinen's test assumes that the multiple

observations from the same matched set are mutually independent. In this paper, we propose exact and asymptotic tests for one-to-two matched binary data. Our method is in markedly different from previously proposed methods in that we do not rely on the mutual independence assumption. The emphasis on dependence among observations from the same matched set is natural and appealing, in both human health and in veterinary medicine studies. The method we propose can be applied to many kinds of diagnostic studies that have a one-to-two matched data structure. Our methods can also be generalized to the one-to-N matched case in a straightforward manner.

Finally we consider the problem of estimating the joint distribution of two correlated random variables where one is observed with error (Chapter 4). An example in nutrition is estimation of the joint distribution of usual energy intake and usual micronutrient intake. While precise biomarkers for energy consumption are available, there are no reliable biomarkers of consumption for nutrients including vitamins and minerals (vitamin K is an exception). Yet, nutritionists are interested in estimating the distribution of usual intake of micronutrients per unit of caloric intake. This is denoted the nutrient density of the diet and involves estimation of the distribution of the ratio of two non-normal random variables, one of which is observed with measurement error. We develop an approach that combines a deconvolution kernel method (DKM) and the method of copulas to estimate the joint distribution of two non-normal variables where one is contaminated. DKM is first used to adjust the univariate measurement error. A Gaussian copula is then used to model the correlation structure between the two variables after error adjustment. We carried out a small simulation study to investigate whether the two-step method we propose is promising. At least in the context of our simulation, we found that the approach produces good results when the correlation between the two random variables is reasonably high. Our findings are tentative, however, and more research is needed before we can recommend the methodology for use broader.

CHAPTER 2. CONSTRUCTION OF DISEASE RISK SCORING SYSTEMS USING LOGISTIC GROUP LASSO: APPLICATION TO PORCINE REPRODUCTIVE AND RESPIRATORY SYNDROME SURVEY DATA

A paper published in Journal of Applied Statistics, Vol 40, No 4, 736-746

Hui Lin[a], Chong Wang[ab], Peng Liu[a] and Derald J. Holtkamp[b] [1]

Abstract

We propose to utilize the group lasso algorithm for logistic regression to construct risk scoring system for predicting disease in swine. This work is motivated by the need to develop a risk scoring system from survey data on risk factor for porcine reproductive and respiratory syndrome (PRRS), which is a major disease, production and financial problem for swine producers in nearly every country. The group lasso provides an attractive solution to this research question, because of its ability to achieve group variable selection and stabilize parameter estimates at the same time. We propose to choose the penalty parameter for the group lasso through leave-one-out cross validation, using the criterion of the area under the receiver operating characteristic curve. Survey data for 896 swine breeding herd sites in the United States and Canada completed between March 2005 and March 2009 is used to construct the risk scoring system for predicting PRRS outbreaks in swine. We show that our scoring system for PRRS significantly improves the current scoring system based on expert opinion. We also show our proposed scoring system is superior in terms of area under the curve to that developed by using multivariate logistic regression model selected based on variable significance.

[1a] *Department of Statistics, College of Liberal Arts and Sciences, Iowa State University, Ames, IA 50011, USA* ; [b] *Department of Veterinary Diagnostic and Production Animal Medicine, College of Veterinary Medicine, Iowa State University, Ames, IA 50011, USA*

2.1 Introduction

Risk scoring systems for predicting disease are widely used in medicine. Such scoring systems are usually derived from multivariate logistic regression models with disease as the response variable. Typical approaches in the literature select potential explanatory variables (risk factors) based on variable significance, with risk scores of selected variables assigned based on estimated regression coefficients (2; 13; 10). However, when the number of potential explanatory variables is large, such approaches may fail to produce a risk scoring system with the greatest power for predicting disease.

This paper is motivated by the need to develop a risk scoring system for porcine reproductive and respiratory syndrome (PRRS) based on survey data. PRRS, caused by the PRRS virus, is a major disease, production and financial problem for swine producers in nearly every country. PRRS costs the United States swine industry around $560 million annually (8). PRRS outbreaks in China caused pork prices to increase by 85 percent in 2006 (5). For breeding herds, costs of clinical outbreaks of PRRS result from lost production due to abortion, mummies, stillborns, pre-wean mortality and sow deaths and increased costs for treatment and control. Performance of observational studies to better understand the relative importance of risk factors for PRRS outbreaks have been limited by the availability of good data on a large set of farms over a relatively long period of time.

In human medicine, large datasets of information on risk factors, prevalence, incidence and clinical outcomes of disease are common. In veterinary medicine, until recently, there have been no parallel efforts to create epidemiological databases on a similar scale. The American Association of Swine Veterinarians (AASV) Production Animal Disease Risk Assessment Program (PADRAP) is a program through which a set of web-based risk assessment surveys are delivered(please visit: http://vdpambi.vdl.iastate.edu/padrap/default.aspx). It is used by veterinarians who are members of the AASV. Each of the surveys consists of a set of questions about potential risk factors for clinical outbreaks of PRRS in swine. Each question may have up to 6 possible responses. Members of the AASV use PADRAP to help producers systematically assess biosecurity factors that may be associated with clinical outcomes. As assessments

are performed by veterinarians they are added to the database of completed assessments.

Version 2 of the PRRS Risk Assessment for the Breeding Herd survey was introduced in 2005. The survey instrument was developed using expert opinion with the aid of the PRRS Risk Assessment Working Group composed of 21 veterinarians and researchers with expertise in PRRS. Initial estimates of the risk scores associated with each response were based on the consensus of expert opinion and equal weight is assigned to each question.

The aim of this study is to use the survey data that has been collected to develop a risk scoring system with 127 survey questions (categorical explanatory variables) that outperforms the current risk scoring system based on expert opinion when multivariate logistic regression is used in similar studies with variables selected by significance. " Quasi-complete-separation" may result when there are a large number of explanatory variables which makes estimation of the coefficients unstable. To stabilize the estimation of parameter coefficients, one popular approach is the lasso algorithm with l_1-norm penalty proposed by Tibshirani (12). Since the lasso algorithm can estimate some variable coefficients to be 0, it can also be used as a variable selection tool. For models with categorical survey questions (explanatory variables), however, original lasso algorithm only selects individual dummy variables instead of sets of the dummy variables grouped by question in the survey. Another disadvantage of applying the lasso method to grouped variables is that the estimates are affected by the way dummy variables are encoded. Thus the group lasso (16) method has been proposed to enable variable selection in linear regression models on groups of variables, instead of on single variables. For logistic regression models, the group lasso algorithm was first studied by Kim *et al.* (4). They proposed a gradient descent algorithm to solve the corresponding constrained problem, which does, however, depend on unknown constants. Meier *et al.* (7) proposed a new algorithm that could work directly on the penalized problem and its convergence property does not depend on unknown constants. The algorithm is especially suitable for high-dimensional problems. It can also be applied to solve the corresponding convex optimization problem in generalized linear models. The logistic group lasso involves selection of a penalty (tuning) parameter λ which can be determined by cross-validation. The group lasso estimator proposed by Meier *et al.* (7) for logistic regression has been shown to be statistically consistent, even with large number of

categorical predictors.

In this paper, we propose to use the logistic group lasso algorithm to construct risk scoring systems for predicting clinical PRRS outbreaks in swine herds. The paper is organized as follows. In Section 2.2, we introduce the multivariate logistic regression and the group lasso method for logistic regression to construct risk scoring systems for clinical PRRS outbreaks and we propose to use the second one. The penalty parameter λ for group lasso is selected through leave-one-out cross validation, using the criterion of the area under the receiver operating characteristic curve . In Section 2.3, we discuss the application to the PRRS survey data from 896 swine breeding herd sites in the United States and Canada. We show that our scoring system for PRRS is superior to both the current scoring system based on expert opinion and that developed by using logistic regression with model selection based on variable significance. Section 2.4 presents a simulation study to evaluate the performance of the multivariate logistic regression and the logistic group lasso method. We conclude with some discussion in Section 2.5.

2.2 Models for risk scoring systems

Consider risk scoring system construction using a sample of n observations, with information collected for G categorical predictors and one binary response variable for each observation. Let $x_{i,g}$ be the vector of dummy variables associated with the gth categorical predictor for the ith observation, where $i = 1, \cdots, n$, $g = 1, \cdots, G$. Let y_i (= 1, diseased; or 0, not diseased) be the binary response for the ith observation. Denote the degrees of freedom of the gth predictor by df_g, which is also the length of vector $x_{i,g}$.

2.2.1 Multivariate logistic regression model

Multivariate logistic regression has been used to construct risk scoring systems for predicting disease (2; 13; 10). Denote the probability of disease for ith subject by θ_i, the model can be formulated as

$$y_i \sim Bernoulli(\theta_i), \tag{2.1}$$

with

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \eta_\beta(x_i) = \beta_0 + \sum_{g=1}^{G} \mathbf{x}_{i,g}^T \beta_g, \tag{2.2}$$

where β_0 is the intercept and β_g is the parameter vector corresponding to the gth predictor.

Construction of risk scoring systems using logistic regression usually consists of two steps: selection among the G risk factors, and estimation of parameters for the selected factors. For model selection, statistical significance has been used as a criterion for inclusion and exclusion of risk factors (2; 13; 10). Some researchers use univariate logistic regression to screen factors by significance before putting them into a multivariate logistic regression model (2; 10), whereas others (13) don't. Traditional estimation of logistic parameters $\beta = (\beta_0^T, \beta_1^T, \beta_2^T, ..., \beta_G^T)^T$ is done through maximizing the log-likelihood

$$\begin{aligned} l(\beta) &= log[\prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i}] \\ &= \sum_{i=1}^{n}\{y_i log(\theta_i) + (1-y_i)log(1-\theta_i)\} \\ &= \sum_{i=1}^{n}\{\ y_i\eta_\beta(\mathbf{x}_i) - log[1+exp(\eta_\beta(\mathbf{x}_i))]\ \}. \end{aligned}$$

For logistic regression analysis with a large number of explanatory variables, complete- or quasi-complete-separation may result which makes the maximum likelihood estimation unstable (1).

2.2.2 Group lasso for logistic regression

In this paper, we propose to perform model selection and parameter estimation for risk scoring system construction by using the group lasso algorithm of Meier *et al.* (7). Instead of maximizing the log-likelihood $l(\beta)$ in the maximum likelihood method, the logistic group lasso estimates are calculated by minimizing the covex function

$$S_\lambda(\beta) = -l(\beta) + \lambda \sum_{g=1}^{G} s(df_g) \parallel \beta_g \parallel_2, \tag{2.3}$$

where λ is a tuning parameter for the penalty and $s(\cdot)$ is a function to rescale the penalty. In lasso algorithms, selection of λ is usually determined by cross-validation using data. For $s(\cdot)$, we use the square root function $s(df_g) = df_g^{0.5}$ as suggested in Meier *et al.* (7).

Here we consider selection of the tuning parameter λ from a multiplicative grid of 148 values $\{0.96\lambda_{max}, 0.96^2\lambda_{max}, 0.96^3\lambda_{max}, ..., 0.96^{148}\lambda_{max}\}$, as in Meier *et al.* (7). Here λ_{max} is defined as

$$\lambda_{max} = \max_{g\in\{1,...,G\}} \left\{ \frac{1}{s(df_g)} \parallel \mathbf{x}_g^T(\mathbf{y} - \bar{\mathbf{y}}) \parallel_2 \right\}, \tag{2.4}$$

such that when $\lambda = \lambda_{max}$, only the intercept is in the model. When λ goes to 0, the model is equivalent to ordinary logistic regression.

The optimal value of λ is determined through leave-one-out cross validation, which is a special case of K-fold cross-validation with K being equal to n, the number of observations in the sample. In each fold, leave-one-out cross validation uses a single observation from the original sample as the validation data, and the remaining observations as the training data. This step is repeated until each observation in the sample is used once as the validation data. Predicted probabilities of disease are calculated and are compared to true observed disease status to assess the predictive power of model.

Three criteria may be used to select the optimal value of λ. The log-likelihood score used in Meier *et al.* (7) is taken as the average of log likelihood of the validation data over all cross-validation sets. Another one is the the maximum correlation coefficient in Yeo and Burge (15) that is defined as

$$\rho_{max} = max\{\rho_\tau | \tau \in (0,1)\}, \tag{2.5}$$

where $\tau \in (0,1)$ is a threshold to classify the predicted probability into a binary disease status and ρ_τ is the Pearson correlation coefficient between the true binary disease status and the predictive disease status with threshold τ.

The third criterion is through the Receiver Operating Characteristic (ROC) analysis. For a given λ value, each leave-one-out cross validation results in one pair of the predicted probability of disease and the true observed disease status for the validation data. In total, we get n such pairs from all leave-one-out cross validation. Given a cutoff value for the predicted probability of disease, we can calculate the true positive rate (sensitivity) and false positive rate (1-specificity) using the n pairs. Then when varying the cutoff value for the predicted probability of disease,

different pairs of true positive rate and false positive rate are generated. Plotting true positive rate versus false positive rate results in an ROC curve. Theoretically, cutoff values can be any values on the real line. The practical cutoff values are determined from resulting scores based on our data. The value of area under the ROC curve (AUC) as well as the confidence interval of AUC can be estimated through an approach proposed by DeLong *et al.* (3). One interpretation of AUC is that it is the probability for the case that a random diseased individual has larger predicted probability of disease than a random non-diseased individual (9) and it has been used to assess predictive power of risk scoring systems (2; 13; 10). We calculate the AUCs for all λs, and the value of λ with the largest AUC is chosen as the λ used in constructing the final scoring system.

2.3 Application to PRRS Data

In this section, we applied the proposed group lasso method to construct a scoring system for PRRS survey data of swine breeding herd sites in the United States and Canada.

2.3.1 Data Description

Surveys in the database completed between March 2005 and March 2009 were candidates for inclusion in the analysis. To avoid multiple surveys from a single swine breeding herd site, the study data set was limited to responses obtained from the most recently completed survey for each site. Surveys meeting these criteria were extracted from the database, and identity information was removed. Incomplete surveys were excluded.

The outcome of interest is whether a site is positive or not. Positive sites are sites with clinical PRRS outbreak in the 3 years prior to when the assessment was completed, negative sites otherwise. The information to determine the outcome was obtained from the survey. A clinical PRRS outbreak was described in the survey as an increase in one or more reproductive performance measures that exceeded normal variation with diagnostic confirmation of PRRS virus involvement.

Of the 896 sites in the United States and Canada included in the study, 499 (56%) became positive during the past 3 years. 127 survey questions were considered potential explanatory

variables in the analysis. The survey questions were first converted to dummy indicator variables. All of the responses for each survey question were defined as a group of variables.

2.3.2 Application of logistic group lasso

First, leave-one-out cross validation was used to choose tuning parameter λ, as described in Section 2.2.

For each λ in the grid $\{0.96\lambda_{max}, 0.96^2\lambda_{max}, 0.96^3\lambda_{max}, ..., 0.96^{148}\lambda_{max}\}$, the values of three evaluation criteria were calculated based on cross validation. The penalty parameter for final risk scoring system was selected to be the one that optimizes AUC.

The logistic group lasso based scoring system was compared with two other systems:

1. The current risk scoring system used in versions 2 of the PRRS risk assessment for the breeding herd that is based on expert opinion,

2. A risk scoring system based on multivariate logistic regression model selected by variable significance.

We constructed the significance based logistic model by following the method used by Van Zee *et al.* (13). Specifically, we used forward stepwise variable selection to construct the logistic regression model with 0.05 significant level. Leave-one-out cross validation was applied to the model construction by variable significance, in the same manner as described for logistic group lasso.

ROC curves are plotted for the three risk scoring systems. A point estimate as well as the 95% confidence interval for the AUC are provided. The estimated AUCs were compared by using the nonparametric approach of DeLong *et al.* (3) and p-values were calculated.

The R package "grplasso" (6) is used to perform group lasso logistic regression. Significance-based logistic model selection is performed using the LOGISTIC procedure in SAS. All other algorithms and calculations are programmed in the R language.

2.3.3 Results

2.3.3.1 Determination of penalty parameter λ

The AUC, maximum correlation coefficient and log-likelihood are calculated based on leave-one-out cross validation and are plotted against the penalty parameter λ in Figure 2.1. The trends for all three criteria are similar with a sharp increase for small values of λ and gradual decrease after reaching the maximum. The optimal values of λ selected to maximize the three criteria are 11.72, 4.22 and 11.72 for AUC, maximum correlation coefficient and log-likelihood respectively.

2.3.3.2 Logistic group lasso based PRRS risk scoring system

The penalty parameter maximizing AUC (i.e. $\lambda = 11.72$) from the leave-one-out cross validation was used for the group lasso estimation of the logistic regression parameters. Figure 2.2 shows the distributions of the predicted probabilities based on cross validation for both negative and positive farms. The predicted probability for positive farms is larger than that of negative farms in stochastic order. The actual risk score can take the value of the predicted probability, the linear predictor in the logistic regression model, or any strictly increasing function of the predicted probability. This is because the ROC curve for a predictor is invariate to such transformation.

In the resulting scoring system, 74 out of 127 survey questions were estimated with 0 coefficients and were excluded from the system. PADRAP questions target internal risks (bio-management of virus already present) and external risks (bio-exclusion of virus not present). A summary of the number of questions included in the final risk scoring system in each category of risk factors in the PRRS Risk Assessment for the Breeding Herd is shown in Table 2.1.

Three out of eight questions regarding internal risk factors remained in the scoring system, and they are all factors concerning characteristics of the herd. Fifty questions remained in external risk factor section out of the total 119 questions. In the external risks section, all of the 14 categories had at least one question remaining in the final scoring system, except that all 4 questions concerning facilities were excluded. Several categories had a large number of

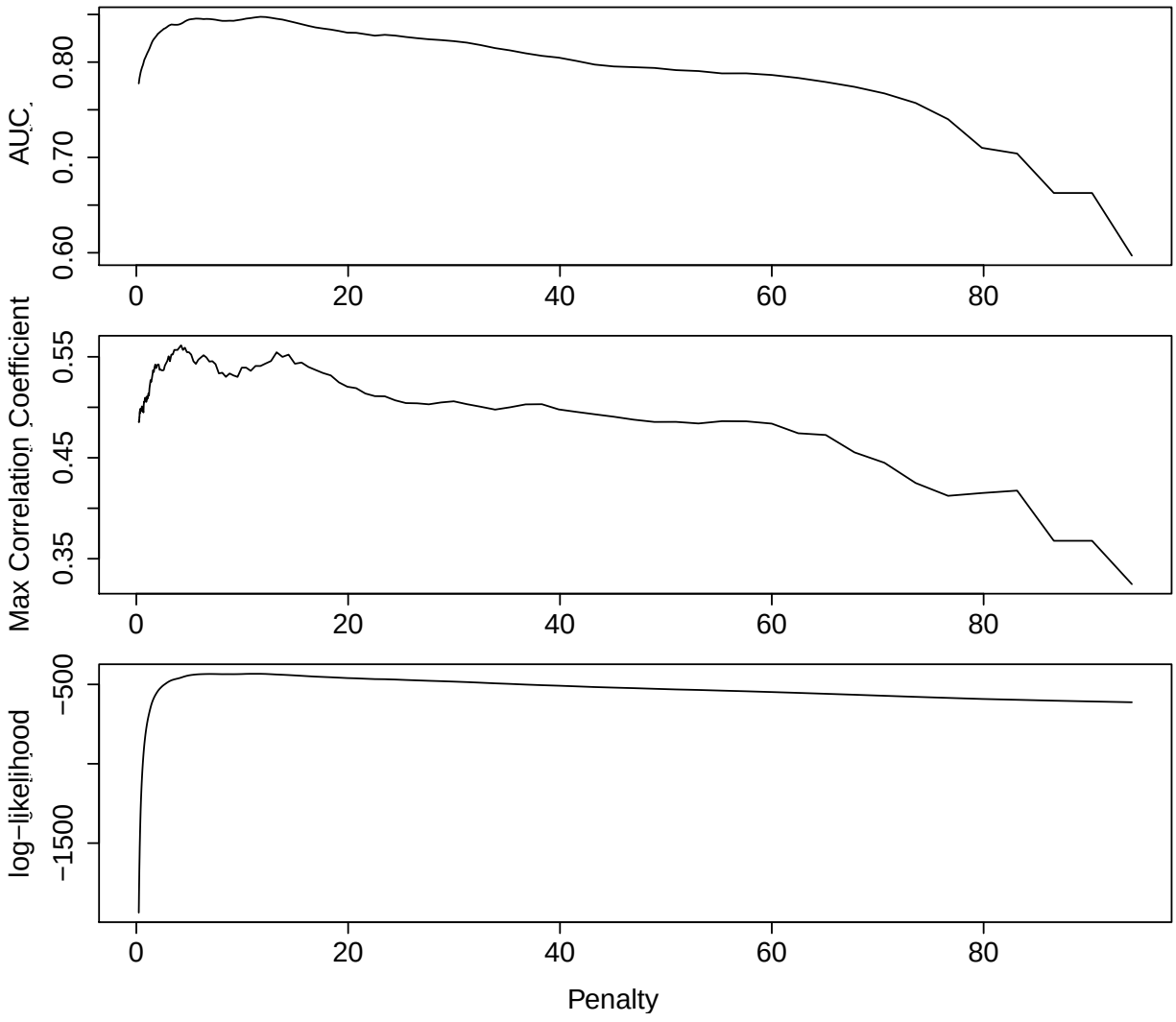


Figure 2.1 Three criteria for choice of penalty parameter λ

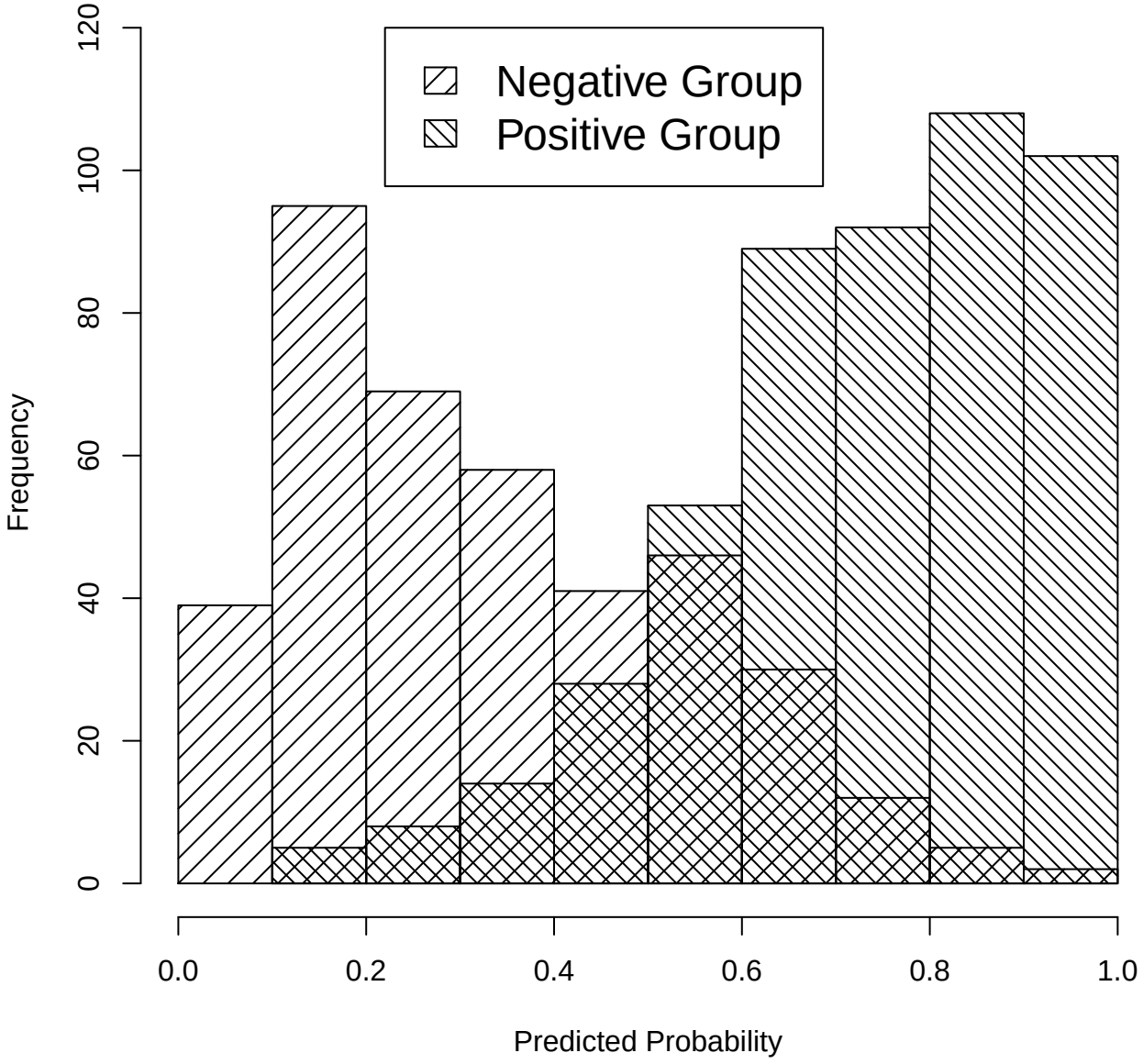


Figure 2.2 Distributions of estimated probabilities for both negative and positive groups

questions removed. In particular, 8 of 12 (66.7%) questions concerning entry of animals into the breeding herd, 18 of 31 (58.1%) questions concerning entry of semen into the breeding herd, 16 of 29 (55.2%) questions concerning transportation of live animals, and 10 of 13 (76.9%) questions concerning neighboring pig farms were excluded.

2.3.4 Comparison among risk scoring systems

The ROC curves for the three risk scoring systems are plotted in Figure 2.3. The ROC curves for the two scoring systems based on logistic regression analyses of the data were constructed using the results of leave-one-out cross validation. The ROC curve of logistic group lasso apparently dominates the other two scoring systems.

Point and 95% interval estimates of AUC are reported in Table 2.2. The risk scoring system based on the logistic group lasso has the largest AUC = 0.848. This AUC estimate is significantly higher than those based on either expert opinion (AUC = 0.696, p-value < 0.001) or logistic regression model selected by variable significance (AUC = 0.807, p-value < 0.001).

2.4 Simulation Study

A simulation study was performed to demonstrate the performance of group lasso logistic regression and compare it to ordinary forward stepwise logistic regression. We simulated 800 farms (i.e. n=800) and 120 survey questions (i.e. G=120) in each dataset, mimicking the real PADRAP data that motivates this paper. There were three possible answers for each question. The outbreak status for the i^{th} farm is generated from a $Bernoulli(1,\ p_i)$ distribution with p_i being a function of the question answers: $ln(\frac{p_i}{1-p_i}) = \beta_0 + \sum_{g=1}^{G} \mathbf{x_{i,g}^T}\beta_{\mathbf{g}}$, where β_0 is the intercept, $\mathbf{x_{i,g}}$ is a three dimentional indication vector for question answer and $\beta_{\mathbf{g}}$ is the parameter vector corresponding to the g^{th} predictor. Three types of questions were considered regarding their effects on the outcome. The first forty survey questions were important questions such that the coefficients of the three answers to these questions were all different:

$$\beta_{\mathbf{g}} = (1, 0, -1) \times \gamma, g = 1, ..., 40,$$

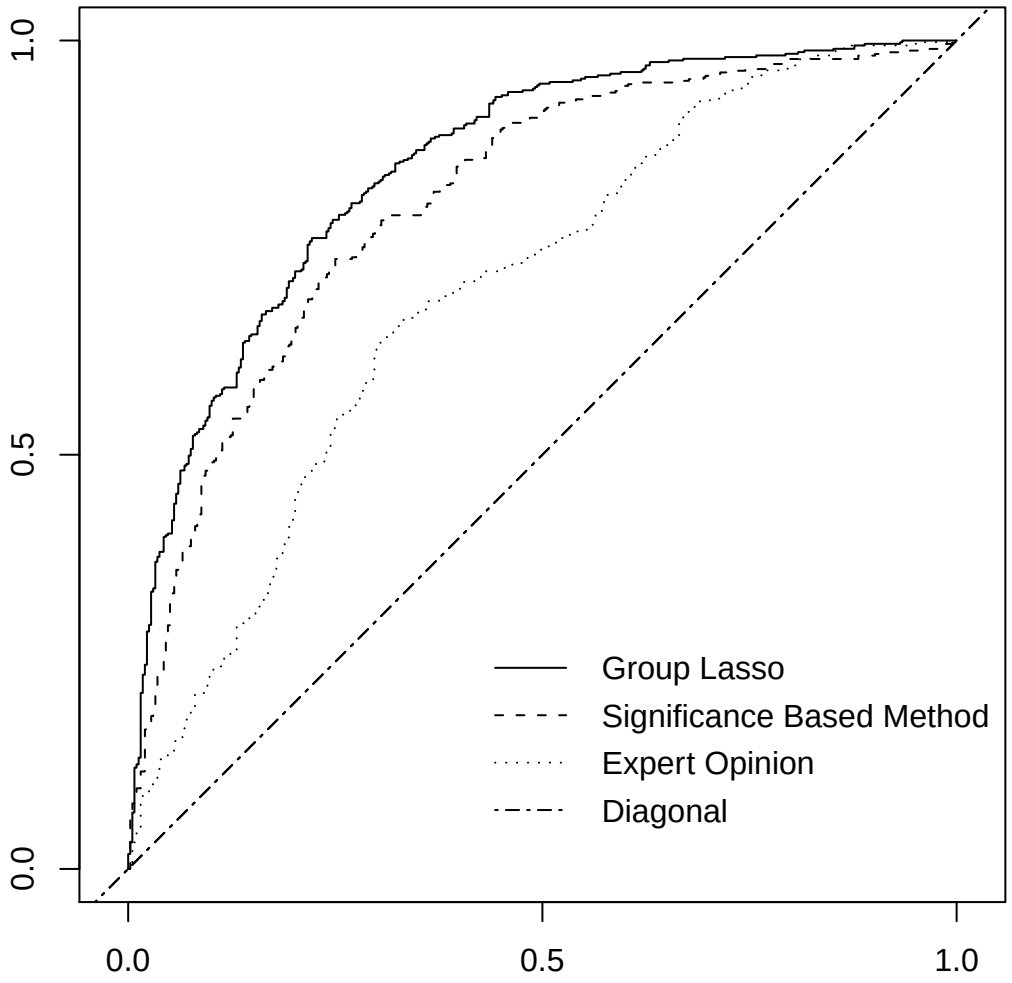


Figure 2.3 ROC curves for three risk scoring systems

where the coefficient γ in the above simulation was set to control the strength of the questions' effect on the outcome. The second forty survey questions were also important questions but only one answer had a coefficient that was different from the other two answers:

$$\beta_{\mathbf{g}} = (1, 0, 0) \times \gamma, g = 41, ..., 80.$$

The last forty survey questions were unimportant questions such that all three answers had the same coefficients:

$$\beta_{\mathbf{g}} = (0, 0, 0) \times \gamma, g = 81, ..., 120.$$

The baseline coefficient β_0 was set to be $-\frac{40}{3}\gamma$ so that on average a farm has 50% of chance to have an outbreak. In this simulation study, we considered the situations where $\gamma = 0.1, 0.25, 0.5, 1$ and 2. For each value of γ, 20 datasets were simulated. We applied the logistic group lasso procedure described in Section 2.2 and the forward stepwise logistic regression to fit the model for each simulated data and calculate the AUC for each fitted model.

Results for the simulation study are shown in Table 2.3. The mean AUC is increasing with the value of γ for both methods. The Wilcoxon signed-rank test result in the last column of Table 2.3 shows that AUC's from group lasso are significantly larger than those from logistic regression, especially for $\gamma \geq 0.25$.

2.5 Discussion

The risk scoring system for disease developed using the logistic group lasso algorithm significantly improves upon the current risk scoring system based on expert opinion for predicting whether a swine breeding site experienced a PRRS outbreak. The simuation study explores the performance of the scoring systems with different settings of coefficients. The logistic group lasso based scoring system is superior to the scoring system constructed through logistic regression selected by variable significance.

One advantage of group lasso is that it can be used as variable selection tool. It not only helps to find important explanatory factors in predicting the response variables but also identifies questions that could be removed from the survey without affecting the survey's ability

for classifying herds according to whether they report clinical PRRS outbreaks in the previous 3 years.

Seventy-four of the 127 questions analyzed were excluded from the final risk scoring system based on logistic group lasso. The questions in the survey were assigned to the internal and external risk sections, in part, on the basis of possible routes of transmission of the PRRS virus. That questions in all except one of the external risk sections were included in the risk scoring system suggests that nearly all of the routes of transmission that were considered in Version 2 of the PRRS Risk Assessment for the Breeding Herd survey are important enough that excluding them would result in risk scoring system that performed significantly worse. This is consistent with the body of research demonstrating the importance of multiple routes by which PRRS virus is transmitted (17). The analysis and results demonstrate how a program like PADRAP, that is supported by a professional association and used by a community of veterinarians, can generate valuable data that contributes to our understanding of the relative importance of risk factors and areas of risk factors for clinical outcomes. The results may also be used to decrease the reliance upon expert opinion to identify questions that should remain in the survey and those that may be eliminated to iteratively increase the value of the program and the data.

Table 2.1 Summary of number of questions in the final risk scoring system by category of risk factors

Category of risk factors	Questions		Dummy Variables	
INTERNAL RISKS	Included	Total	Included	Total
Circulation Risk				
Characteristics of the herd	3	4	9	11
Characteristics of the site	0	2	0	5
Management practices	0	2	0	9
Total	***3***	***8***	***9***	***25***
EXTERNAL RISKS				
Pig Related				
Entry of replacement animals into the breeding herd	4	12	18	40
Entry of semen into the breeding herd	13	31	47	104
Non-Pig Related				
Transportation of live animals	13	29	38	71
Transportation of feed	1	1	2	2
Employee and service vehicles	1	2	3	6
Disposal of dead animals and waste management	2	8	3	10
Employees and visitors	5	9	15	19
Entry of supplies	1	1	3	3
Facilities	0	4	0	11
Biovectors	1	1	2	1
Density of pig farms in the area	3	3	10	10
Neighboring pig farms	3	13	12	28
Distance to pork industry infrastructure	2	4	5	11
Topography and forestation of surrounding area	1	1	3	3
Total	***50***	***119***	***161***	***319***

Table 2.2 AUC estimations for three risk scoring systems

Model Names	AUC	95% CI
Group Lasso	0.848	(0.822, 0.873)
Significance Based Method	0.807	(0.773, 0.841)
Expert Opinion	0.696	(0.661, 0.731)

Table 2.3 Simulation study result with various values of coefficient γ. Reported are mean and standard deviation of AUC for both methods, mean difference and p value from Wilcoxon signed rank test.

Coefficient γ	Group Lasso (mean±sd)	Logistic Regression (mean±sd)	p value
0.1	0.57 ± 0.03	0.54 ± 0.06	0.040
0.25	0.71 ± 0.02	0.64 ± 0.04	< 0.001
0.5	0.91 ± 0.03	0.78 ± 0.03	< 0.001
1	0.92 ± 0.01	0.82 ± 0.02	< 0.001
2	0.95 ± 0.01	0.84 ± 0.02	< 0.001

CHAPTER 3. EXACT AND ASYMPTOTIC STATISTICAL TESTS FOR DIFFERENCE IN PROPORTIONS OF ONE-TO-TWO MATCHED BINARY VARIABLES

Abstract Matched observations with dichotomous responses commonly occur in medical and epidemiological researches. Although standard approaches exist for one-to-one paired binary data analyses, there is not much work for one-to-two or one-to-N matched binary data in the current statistical literature. The existing Miettinen's test assumes that the multiple observations from the same matched set are mutually independent. In this paper, we propose exact and asymptotic tests for one-to-two matched binary data. Our method is markedly different from existing methods in that ours does not rely on a mutual independence assumption. The emphasis on dependence among observations from the same matched set is natural and appealing, as much in human health as it is in veterinary medicine. It can be applied to many types of diagnostic studies with one-to-two matched data structure. Our methods can be generalized to one-to-N matched case in a straightforward manner.

3.1 Introduction

Matched observations with dichotomous responses commonly occur in medical and epidemiological researches. Although standard approaches exist for one-to-one paired binary data analyses, not much research on one-to-two or one-to-N matched binary data has been published.

Our research was originally motivated by the pooling of diagnostic tests. Often, testing units one-by-one is inefficient, especially when disease prevalence is sufficiently low. The concept of screening pooled samples originated during the second world war to detect syphilis in US soldiers (21). It has aroused significant amount of attention and been used successfully in

various applications. Many studies have demonstrated the successful use of pooling strategies on HIV testing (22) (30) (27) (31). Budget reduction is an important issue which limits the number of tests so that the derived estimates can be imprecise. One way to overcome budget limitations and improve the accuracy of estimates is pooled testing. Vansteelandt *et al.* (29) showed that a good design can severely reduce cost. An applied example in Vansteelandt *et al.* (29) showed that using test pools with an average of seven units reduced cost by 44 percent with virtually no loss in precision. In some circumstances, the advantages of pooling include earning more accuracy as well(22).

For the one-to-one case, McNemar (24) developed a test of marginal homogeneity in a 2×2 table that is applicable to pair-matched observations or a cohort measured twice on a variable with binary outcome. Bennett and Underwood (19) compared exact power with the non-central Chi-square approximation for sample sizes of 10, 20 and 40 and found the Chi-square approximation to be adequate. Miettinen (25) derived the asymptotic power for testing the difference between cases and controls with dichotomous response in the case of one to one and one to R matching. Stephen (28) derived the exact power based on Miettinen's work and compared it to the asymptotic power of the test. However, Miettinen's test assumes that the multiple observations from the same matched set are mutually independent conditioned on the pair. This assumption is hard to hold for pooling test data where the pooled sample is of course dependent of the individual samples being pooled. Furthermore, the independence assumption can be assessed statistically using Fisher's exact test and our data show significant evidence of dependence.

We proposed exact and asymptotic tests for one-to-two matched binary data. Our methods fit a more general situation that does not assume that observations from the same subject are mutually independent. It can be applied to many types of diagnostic studies with one-to-two matched data structures besides dual sample pooling, such as one-to-two case control studies. It is important to understand the properties of matching designs so as to be able to make the best use of them. Our methods can be generalized to one-to-N matched cases. For clarity of presentation we establish basic concepts, terminology and notation in Section 3.2. We illustrate the exact test procedure and an asymptotic test procedure in Section 3.3 and

Section 3.4, respectively. In Section 3.5, we demonstrate the merits of our tests through a simulation study. In Section 3.6, we applied our methods on two practical situations that fail to have the independence required by Miettinen's test. Discussion follows in Section 3.7.

3.2 Basic concepts, Terminology and Notation

Assume we have n subjects going through two strategies of test. By saying one-to-two we mean there is one binary observation taken from each subject under strategy 1 and two binary observations taken from the same subject under strategy 2. In the paper, we use upper case letters to denote random variables and lower case letters to denote observed realizations. We denote the set of three observations from subject j and its realization as:

$$(Y_{1j}, Y_{2j1}, Y_{2j2}) \; and \; (y_{1j}, y_{2j1}, y_{2j2})$$

respectively, where Y_{1j} denotes the observation under strategy 1 while Y_{2j1} and Y_{2j2} denote observations under strategy 2 with $j = 1, .., n$.

For the j^{th} matched group a realization $(y_{1j}, y_{2j1}, y_{2j2})$ is obtained for the random response vector $(Y_{1j}, Y_{2j1}, Y_{2j2})$. The value of the response variable Y is either 0 or1. And $p_1 = Pr\{Y_{1j} = 1\}$ (probability a subject under strategy 1 has test result 1) and $p_2 = Pr\{Y_{2j1} = 1\} = Pr\{Y_{2j2} = 1\}$ (probability a subject under strategy 2 has test result 1). The object of the study is to make inferences about

$$\delta = p_1 - p_2,$$

and test the null hypothesis

$$H_0 : \delta = 0$$

We consider the multinomial distribution of the response vector (X_{1j}, X_{2j}) where $X_{1j} = Y_{1j}$ and $X_{2j} = Y_{2j1} + Y_{2j2}$. There are six possible realizations and denote $Z_{kl}^{(j)} = I(X_{1j} = k,\ X_{2j} = l)$ with $k = 0, 1$ and $l = 0, 1, 2$. It is a multinomial distribution with $Z_{kl}^{(j)} \sim multi - Bernoulli(p_{kl})$ where $p_{kl} = E[Z_{kl}^{(j)}]$ is invariant across units denoted by j. The cell counts for a total of n sets. $Z_{kl} = \sum_{j=1}^{n} Z_{kl}^{(j)}$ has a multinomial distribution with $Z_{kl} \sim multinomial(n, p_{kl})$.

Table 3.1 Outcome for Subject j

	Test 2		
Test 1	$Z_{12}^{(j)}$	$Z_{11}^{(j)}$	$Z_{10}^{(j)}$
	$Z_{02}^{(j)}$	$Z_{01}^{(j)}$	$Z_{00}^{(j)}$

Table 3.2 Counting Table for n Sets of Observations

		Test 2			
		2	1	0	Total
Test 1	1	Z_{12}	Z_{11}	Z_{10}	$n_{1.}$
	0	Z_{02}	Z_{01}	Z_{00}	$n_{0.}$
	Total	$n_{.2}$	$n_{.1}$	$n_{.0}$	n

3.2.1 Miettinen Exact Test

Miettinen proposed an exact test for this matching design under the following two assumptions:

1. the n vectors $(Y_{1j}, Y_{2j1}, Y_{2j2})$ are independently and identically distributed, and that
2. Y_{1j}, Y_{2j1}, Y_{2j2} are mutually independent conditionally on $(\pi_1, \pi_2) = (\pi_{1j}, \pi_{2j})$ where $p_1 = E(\pi_1), p_2 = E(\pi_2)$.

Miettinen (25) proposed an exact test based on the multinomial formulation. Conditioning on $S_1 = Z_{10} + Z_{01}$ and $S_2 = Z_{11} + Z_{02}$, Z_{10} and Z_{11} have independent binomial distributions. Under H_0,

$$Z_{10} \sim Binomial(S_1, \frac{1}{3});$$
$$Z_{11} \sim Binomial(S_2, \frac{2}{3}).$$

The computation of the p-value for hypothesis testing is: $p = Pr(Z_{10} + Z_{11} \geq z_{10} + z_{11} = v)$ i.e.

$$p = \sum_{k_1+k_2 \geq v} \binom{s_1}{k_1} \left(\frac{1}{3}\right)^{k_1} \left(\frac{2}{3}\right)^{s_1-k_1} \binom{s_2}{k_2} \left(\frac{2}{3}\right)^{k_2} \left(\frac{1}{3}\right)^{s_2-k_2} \tag{3.1}$$

When Test 1 and Test 2 results are biologically related, as in a pooling test senario, the assumption of independence between Test 1 and Test 2 may not be reasonable. Paired test analysis methods such as McNemar's test do not generally require independence between paired

test results. In the following sections, we discuss statistical tests without the conditional independence assumption.

3.3 Random Exact Test

3.3.1 Test Statistic

Let $R_{kl}^{(j)} \mid Z_{kl}^{(j)} \sim Bin(Z_{kl}^{(j)}, \frac{1}{2})$ and $R_{kl} = \sum_{j=1}^{n} R_{kl}^{(j)}$. The marginal probability is $Pr\{R_{kl}^{(j)} = 1\} = Pr\{R_{kl}^{(j)} = 1 \mid Z_{kl}^{(j)} = 1\}Pr\{Z_{kl}^{(j)} = 1\} = \frac{p_{kl}}{2}$. So for $k \neq k'$ *or* $l \neq l'$,

$$Pr\{Z_{k'l'}^{(j)} + R_{kl}^{(j)} = 2\} = Pr\{Z_{k'l'}^{(j)} = 1, R_{kl}^{(j)} = 1\} = 0 \tag{3.2}$$

$$\begin{aligned} &Pr\{Z_{k'l'}^{(j)} + R_{kl}^{(j)} = 1\} = Pr\{Z_{k'l'}^{(j)} = 1, R_{kl}^{(j)} = 0\} + Pr\{Z_{k'l'}^{(j)} = 0, R_{kl}^{(j)} = 1\} \\ &= Pr\{Z_{k'l'}^{(j)} = 1\}Pr\{R_{kl}^{(j)} = 0 \mid Z_{k'l'}^{(j)} = 1\} + Pr\{R_{kl}^{(j)} = 1\}Pr\{Z_{k'l'}^{(j)} = 0 \mid R_{kl}^{(j)} = 1\} \\ &= Pr\{Z_{k'l'}^{(j)} = 1\} + Pr\{R_{kl}^{(j)} = 1\} = p_{k'l'} + \tfrac{p_{kl}}{2} \end{aligned} \tag{3.3}$$

$$Pr\{Z_{k'l'}^{(j)} + R_{kl}^{(j)} = 0\} = 1 - Pr\{Z_{k'l'}^{(j)} + R_{kl}^{(j)} = 1\} = 1 - (p_{k'l'} - \frac{p_{kl}}{2}) \tag{3.4}$$

Then we have $\sum_{j=1}^{n}(Z_{k'l'}^{(j)} + R_{kl}^{(j)}) = Z_{k'l'} + R_{kl} \sim Bin(n, p_{k'l'} + \frac{p_{kl}}{2})$, for any $(k,l) \neq (k^{'}, l^{'})$.

$$\begin{aligned} \delta &= E(X_{1j}) - \tfrac{E(X_{2j})}{2} \\ &= (p_{12} + p_{11} + p_{10}) - (p_{12} + p_{02} + \tfrac{1}{2}p_{11} + \tfrac{1}{2}p_{01}) \\ &= p_{10} - \tfrac{1}{2}p_{01} + \tfrac{1}{2}p_{11} - p_{02} \end{aligned} \tag{3.5}$$

Denote $S = Z_{10} + R_{11} + Z_{02} + R_{01}$ and $p_s = p_{10} + \frac{p_{11}}{2} + p_{02} + \frac{p_{01}}{2}$. Under H_0: $p_{10} + \frac{1}{2}p_{11} = p_{02} + \frac{1}{2}p_{01}$, we have $Z_{10} + R_{11} \mid S \sim Bin(S, \frac{1}{2})$. A two-sided Random Exact Test is done through the following three steps:

1. Random sample $r_{11} \mid\sim Bin(z_{11}, \frac{1}{2})$ *and* $r_{01} \sim Bin(z_{01}, \frac{1}{2})$
2. Denote $s_1 = max(z_{10} + r_{11}, z_{02} + r_{01})$, $s_2 = min(z_{10} + r_{11}, z_{02} + r_{01})$ and $s = z_{10} + r_{11} + z_{02} + r_{01}$.
3. Calculate p-value by $Pr\{x \leq s_2 \text{ or } x \geq s_1\}$ with $x \sim Bin(s, \frac{1}{2})$.

Due to the randomization of r_{11}, the procedure can give different answers for the exact same data. We can avoid the arbitrariness of randomization while keeping the beautiful theory of these procedures by a simple change of viewpoint to what is called a "fuzzy p-value" advanced by Geyer & Meeden (2005) (20). Different from conventional p-values, fuzzy p-values are random variables interpreted as p-values. In terms of the random exact test illustrated above, r_{11} is called a latent variable and the p-value calculated in step 3 is refered to as a latent p-value. The latent p-value would be a p-value if the values of the latent variable were observed. The exact test employing the notion of a fuzzy p-value uses simulations of the latent under the null hypothesis. It provides an expression of both the strength and the uncertainty of the evidence against the null hypothsis.

3.3.2 Power of the Random Exact Test

For fixed δ, p_1, p_{12} and p_{11},

$$p_{01} = 2 * p_1 * (1 - p_1) - p_{11} \tag{3.6}$$

$$p_{02} = p_1^2 - p_{12} \tag{3.7}$$

$$p_{10} = p_1 + \delta - p_{12} - p_{11} \tag{3.8}$$

$$p_{00} = (1 - p_1)^2 - (p_1 + \delta - p_{12} - p_{11}) \tag{3.9}$$

We have shown that $Z_{10} + R_{11} \sim Bin(n, p_{10} + \frac{p_{11}}{2})$ and $Z_{02} + R_{01} \sim Bin(n, p_{02} + \frac{p_{01}}{2})$.

$$p_s = p_{10} + \frac{p_{11}}{2} + p_{02} + \frac{p_{01}}{2} = 2p_1 + \delta - p_{11} - 2p_{12} \tag{3.10}$$

$$\frac{p_{10} + \frac{p_{11}}{2}}{p_{10} + \frac{p_{11}}{2} + p_{02} + \frac{p_{01}}{2}} = \frac{p_1 - p_{12} - \frac{p_{11}}{2} + \delta}{2p_1 + \delta - p_{11} - 2p_{12}} \equiv q \tag{3.11}$$

Then $S \sim Bin(N, p_s)$ and $Z_{10} + R_{11} \mid S \sim Bin(S, q)$. The unconditional power can be obtained as the expectation of the conditional power. The power expression of the exact binomial test is (3.12).

$$
\begin{aligned}
&Pr\{Z_{10}+R_{11}\le u_{\alpha/2} \text{ or } Z_{10}+R_{11}\ge u_{1-\alpha/2}\}\\
&=\sum_{S=0}^{n}\binom{n}{S}p_s^S(1-p_s)^{n-S}\sum_{\substack{Z_{10}+R_{11}\le l_{\alpha/2},\\ Z_{10}+R_{11}\ge u_{\alpha/2}}}\binom{S}{Z_{10}+R_{11}}q^{Z_{10}+R_{11}}(1-q)^{S-(Z_{10}+R_{11})}
\end{aligned}
\tag{3.12}
$$

where $l_{\alpha/2}\doteq max\{n \mid \sum_{x=0}^{n}\binom{S}{x}(\frac{1}{2})^S\le\frac{\alpha}{2}\}$ and $u_{\alpha/2}\doteq min\{n \mid \sum_{x=n}^{S}\binom{S}{x}(\frac{1}{2})^S\le\frac{\alpha}{2}\}$.

3.4 Asymptotic Test

Denote $T^{(j)}=(Z_{10}^{(j)}+\frac{Z_{11}^{(j)}}{2})-(Z_{02}^{(j)}+\frac{Z_{01}^{(j)}}{2})$.

$$
E[T^{(j)}]=p_{10}+\frac{p_{11}}{2}-p_{02}-\frac{p_{01}}{2}=\delta \tag{3.13}
$$

$$
\begin{aligned}
Var[T^{(j)}] &= E[T^{(j)2}]-\{E[T^{(j)}]\}^2\\
&= E[\{(Z_{10}^{(j)}+\tfrac{Z_{11}^{(j)}}{2})-(Z_{02}^{(j)}+\tfrac{Z_{01}^{(j)}}{2})\}^2]-\delta^2\\
&= E[(Z_{10}^{(j)}+\tfrac{Z_{11}^{(j)}}{2})^2+(Z_{02}^{(j)}+\tfrac{Z_{01}^{(j)}}{2})^2-2(Z_{10}^{(j)}+\tfrac{Z_{11}^{(j)}}{2})(Z_{02}^{(j)}+\tfrac{Z_{01}^{(j)}}{2})]-\delta^2
\end{aligned}
\tag{3.14}
$$

Since at most one of $\{Z_{10}^{(j)},Z_{11}^{(j)},Z_{02}^{(j)},Z_{01}^{(j)}\}$ is non-zero, $Z_{kl}^{(j)}Z_{k'l'}^{(j)}=0$ if $k\ne k'$ or $l\ne l'$. Also we have $Z_{kl}^{(j)2}=Z_{kl}^{(j)}$. Then (3.14) can be written as (3.15).

$$
Var[T^{(j)}]=E[Z_{10}^{(j)2}+\frac{Z_{11}^{(j)2}}{4}+Z_{02}^{(j)2}+\frac{Z_{01}^{(j)2}}{4})]-\delta^2=p_{10}+\frac{p_{11}}{4}+p_{02}+\frac{p_{01}}{4}-\delta^2 \tag{3.15}
$$

Since observations from different subjects are independent, we have:

$$
\mu=E\sum_{j=1}^{n}T^{(j)}=n\delta \tag{3.16}
$$

$$
\sigma^2=Var\sum_{j=1}^{n}T^{(j)}=n(p_{10}+\frac{p_{11}}{4}+p_{02}+\frac{p_{01}}{4}-\delta^2) \tag{3.17}
$$

Under H_0, By CLT, $\frac{\sum_{j=1}^{n}T^{(j)}}{\sqrt{Var(\sum_{j=1}^{n}T^{(j)})}}=\frac{(Z_{10}+\frac{Z_{11}}{2})-(Z_{02}+\frac{Z_{01}}{2})}{\sqrt{n(p_{10}+\frac{p_{11}}{4}+p_{02}+\frac{p_{01}}{4})}}$ is asymptotic standard normal when n is large. The asymptotic test is to compare the following test statistics to $N(0,1)$.

$$\frac{(Z_{10}+\frac{Z_{11}}{2})-(Z_{02}+\frac{Z_{01}}{2})}{\sqrt{z_{10}+\frac{z_{11}}{4}+z_{02}+\frac{z_{01}}{4}}} \tag{3.18}$$

When $\delta \neq 0$, $\frac{(Z_{10}+\frac{Z_{11}}{2})-(Z_{02}+\frac{Z_{01}}{2})-n\delta}{\sqrt{n(p_{10}+\frac{p_{11}}{4}+p_{02}+\frac{p_{01}}{4}-\delta^2)}}$ is asymptotic standard normal . The power with respect to δ is a function of the mean and variance of the test statistic

$$\beta = 2\Phi(\frac{\phi_{\alpha/2}\sqrt{n(p_{10}+\frac{p_{11}}{4}+p_{02}+\frac{p_{01}}{4})}-n\delta}{\sqrt{n(p_{10}+\frac{p_{11}}{4}+p_{02}+\frac{p_{01}}{4}-\delta^2)}}) \tag{3.19}$$

where $\phi_{\alpha/2}$ is the $\alpha/2$ lower quantile of standard normal distribution and $\Phi(\cdot)$ is the cumulative density function of standard normal distribution. It is time consuming to compute (3.12) and (3.19) for large n. Therefore, in the following simulation study, we estimate the power of exact and asymptotic tests through Monte Carlo sampling.

3.5 Simulation

We conducted a Monte Carlo study to examine type one error levels and power of the proposed statistical tests. In particular, we took $n = 10, 20, 30, 50, 100, 200, 300$. Since the exact power depend on the individual binomial or multinomial parameters, the arbitary choice of a further parameter is necessary. We gave arbitary values for p_1, p_{12} , p_{11} and δ, then the rest parameters are determined by solving (3.6) - (3.9). We consider different parameterizations according to (3.20) - (3.24).

$$\delta = 0.05 * (h-1), \ \ where\ h = 1, 2, 3, 4, 5 \tag{3.20}$$

For each δ value in (3.20), we simulate sample from four different settings:

$$Setting\ 1: \ \ p_1 = 0.3, p_{12} = 0.01, p_{11} = 0.01 \tag{3.21}$$

$$Setting\ 2: \ \ p_1 = 0.3, p_{12} = 0.08, p_{11} = 0.15 \tag{3.22}$$

$$Setting\ 3: \ \ p_1 = 0.4, p_{12} = 0.15, p_{11} = 0.24 \tag{3.23}$$

$$Setting\ 4: \ p_1 = 0.4, p_{12} = 0.11, p_{11} = 0.03 \tag{3.24}$$

Note that the 4^{th} setting can only get 3 δ values (i.e $\delta = 0, 0.05, 0.1$) due to the support of parameter ($[0, 1]$). For each case, $M = 2000$ simulations were performed. Function "rmultinom()" in R is used to simulate multinomial samples. As mentioned, it is computationally infeasible to calculate (3.12) for large n. Therefore, the power for each test on each sample set was estimated with 2000 simulations.

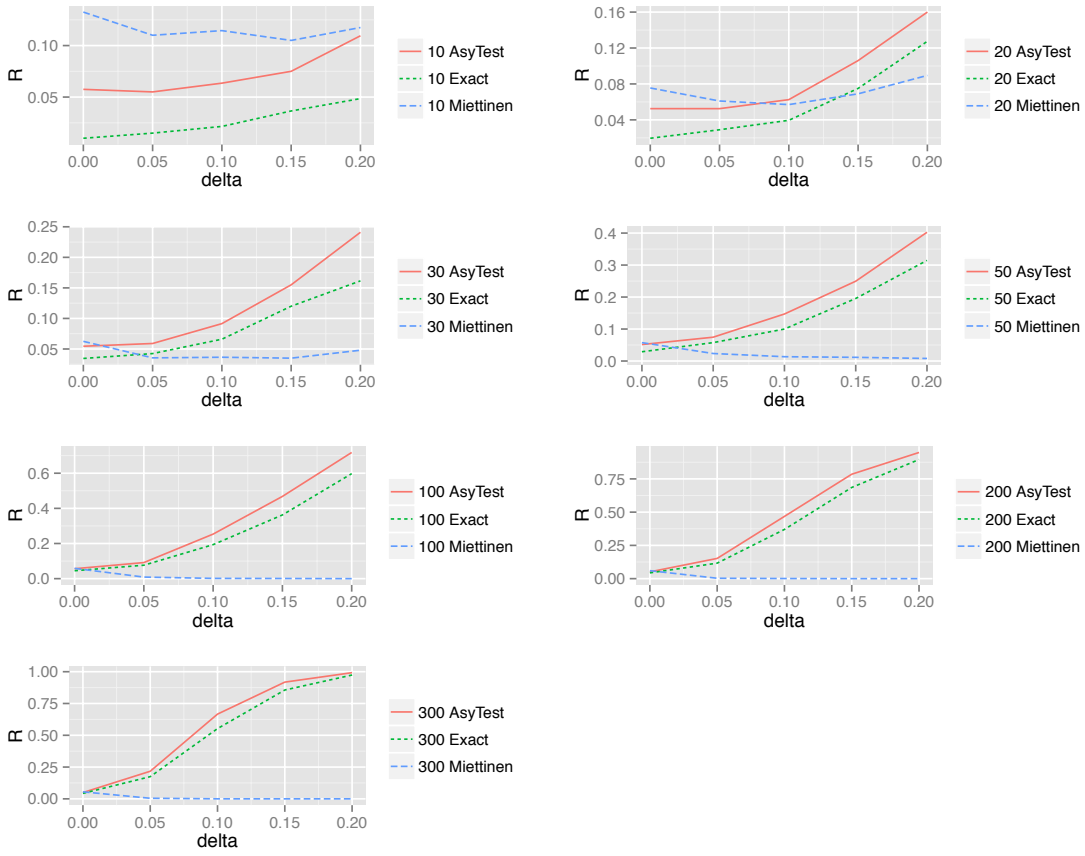


Figure 3.1 Comparison of exact test power and asymptotic test power for setting 1 of one–to-two case. The numbers in the legend indicate the sample size. AsyTest indicates asymptotic test; Exact indicates exact test.

The resulting power values of the exact test and the asymptotic test for different parameterizations are shown in Figure 3.1 to 3.4. From the results we can see that the Miettinen's

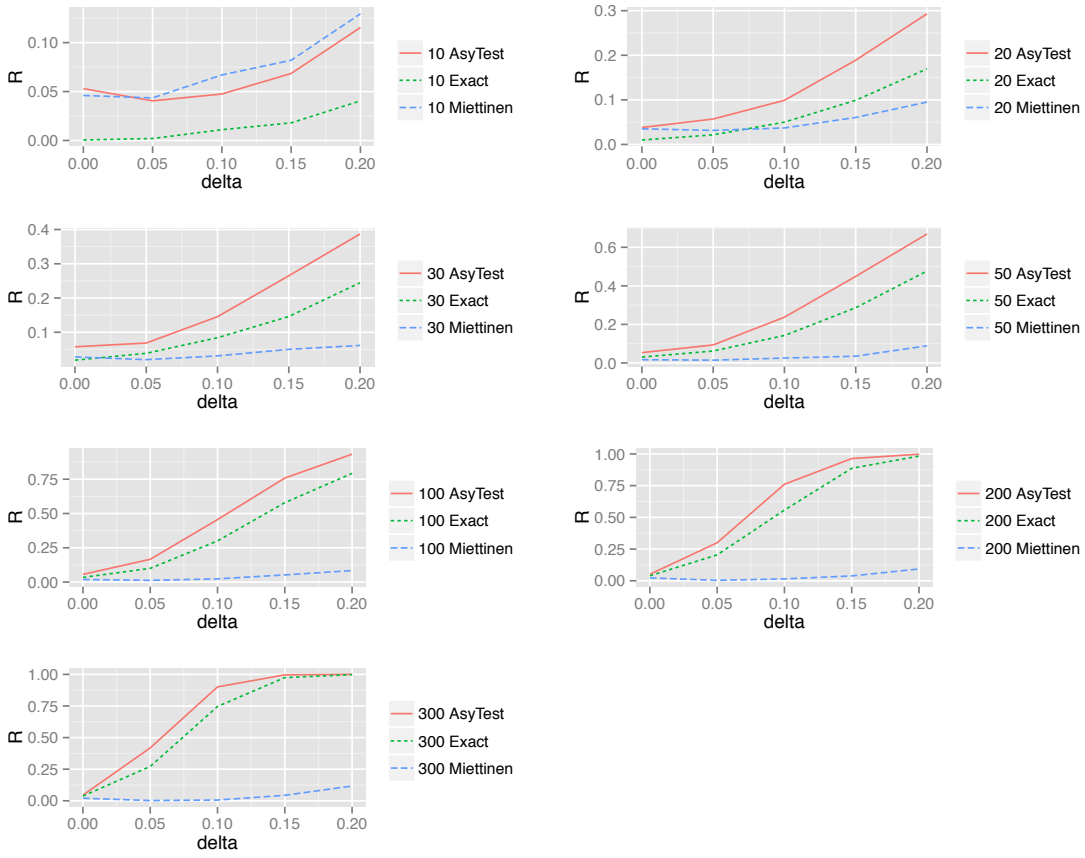


Figure 3.2 Comparison of exact test power and asymptotic test power for setting 2 of one–to-two case. The numbers in the legend indicate the sample size. AsyTest indicates asymptotic test; Exact indicates exact test.

Test does not work here. The asymptotic test consistently dominates the others for all settings. Though for small sample size ($n \leq 30$) the performance drops, our proposed tests perform well as long as the size is larger than 50. As δ increases, the power increases steadily for the exact binomial test and asymptotic test as we would expect. Miettnen's test nearly has no power even the sample size is large except for setting 3. The reason may be that the design of parameterization for setting 3 approximate the mutually independent situation the most. A full table of parameterization strategy for setting 3 under non-hypothesis is shown in table A.3.

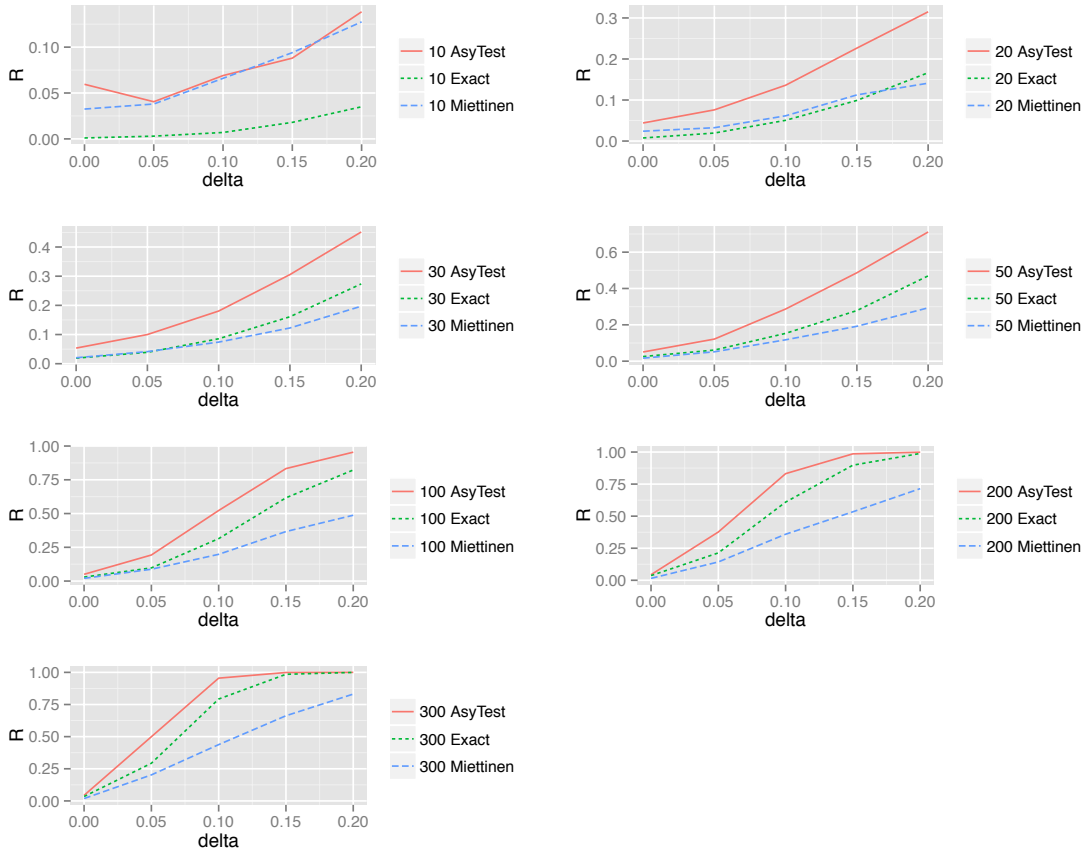


Figure 3.3 Comparison of exact test power and asymptotic test power for setting 3 of one–to-two case. The numbers in the legend indicate the sample size. AsyTest indicates asymptotic test; Exact indicates exact test.

3.6 Application Examples

3.6.1 Dual Sample Pooling Test

Salmonella enteric serovar Enteritidis (SE) has emerged in the past 30 years as a leading cause of human salmonellosis in the United States (18; 26). If SE is isolated from the environment of chicken houses, then eggs from SE-positive houses must be tested. Testing eggs for SE requires a large sample size as only a small proportion are contaminated in an infected flock. Therefore, environmental sampling is the primary means by which flocks are monitored for SE. Environmental (or egg) testing has traditionally been carried out using bacterial cul-

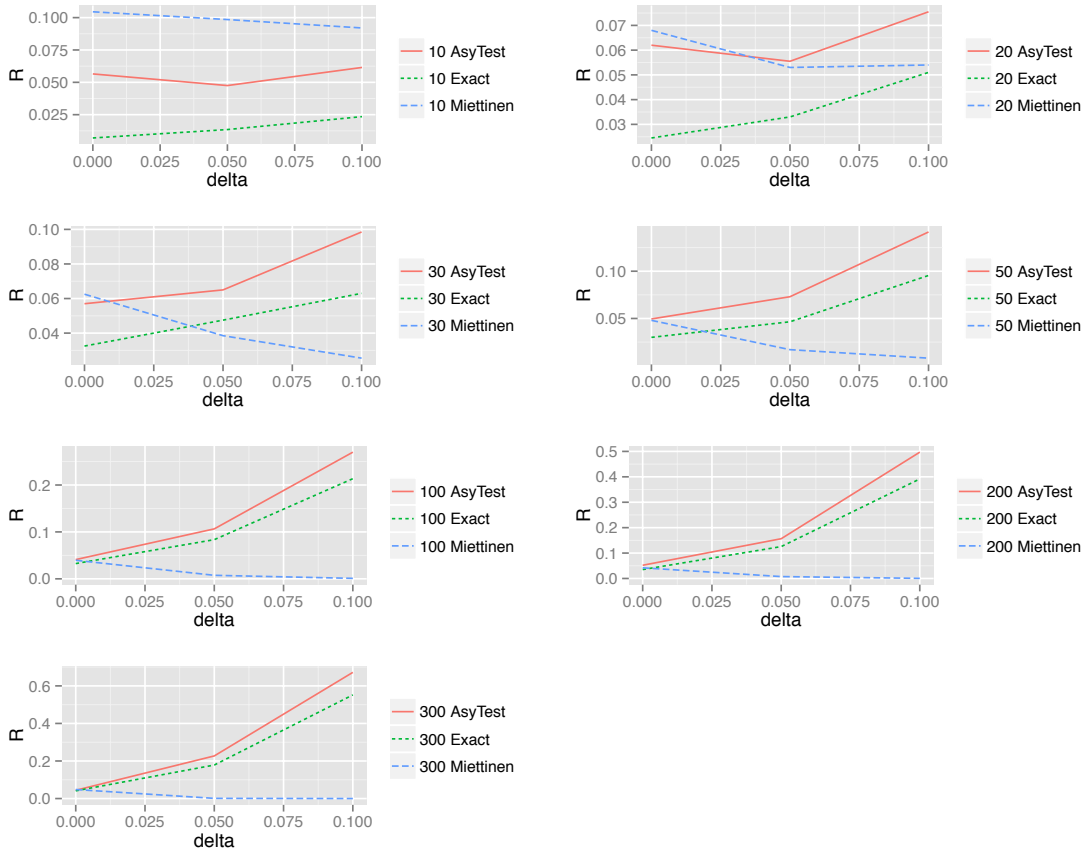


Figure 3.4 Comparison of exact test power and asymptotic test power for setting 4 of one–to-two case. The numbers in the legend indicate the sample size. AsyTest indicates asymptotic test; Exact indicates exact test.

turing which is the standard by which all other tests are compared. Bacteriological culturing typically requires 5 to 7 days before results are obtained. Real-time polymerase chain reaction (RT PCR) is one testing method that has been developed to decrease the time required for testing. The cost of testing associated with the implementation of U.S. Food and Drug Administration (FDA) 's Final Rule has placed a substaincial burden on producers. Sample pooling is one strategy to reduce costs and labor. The aim of the study is to examine the validity of an SE-specific RT PCR in pooled samples. The provisionally approved National Poultry Improvement Plan (NPIP) modified semisolid Rappaport-Vassiliadis (MSRV) method as the gold standard. RT PCR results from pool sizes of two were compared with single sample testing. A

total of 208 environmental field samples were collected from three commercial layer houses on the same site. Houses were previously found to be positive for SE by culture at the ISU VDL. Each house contained twelve rows of cages with three tiers of cages within each row. Flocks within each house consisted of adult laying hens. Gauze drag swabs pre-soaked with sterilized milk mile were used to sample egg belt sections from each tier of cages within each row and from fecal material on support beams directly under the cage section sampled. Samples were taken every fifty feet along the length of the house. Swabs were put into Whirl-Pak bags and transported on ice to the Iowa State University Veterinary Diagnostic Laboratory for testing. After incubation, 1 ml aliquots were removed from the enrichment broth of field environmental samples for RT PCR analysis. Sets of pooled samples were prepared from these aliquots so that each individual sample was represented once and randomly assigned to a pooled set of 2 samples (208 individual, 104 pools of two). In this example, the pooling test is test 1 and the single test is test 2 with n=104. The counting results are presented in Table 7 .

Table 3.3 Counting Table for Dual Pooling Test

		Test 2			
		2	1	0	Total
Test 1	1	0	7	0	7
	0	0	0	97	97
	Total	0	7	97	104

A Fisher's exact test for independence results in a p-value of 4.707×10^{-11}, indicating convincing evidence of dependency between tests in the table. Thus Miettinen's test should not be applied in this situation because it is derived under the independence assumption. The probability that the fuzzy p-value is less than 0.05 is only 0.06. The median fuzzy p-value is 0.25 and the 95% quantile is 1. The result indicates no evidence against H_0.

3.6.2 Pen-based oral fluid specimens for influenza A virus detection

Christa K. Goodell et al. used a matched design in their influenza A virus (IVA) monitoring study. For IAV detection, the traditional ante mortem Nasal Swabs (NS) specimen is difficult and expensive to get because it is necessary to select, restrain, and swab individual pigs. Alternatively, oral fluids (OF), a specimen new to swine diagnostics but well-characterized in

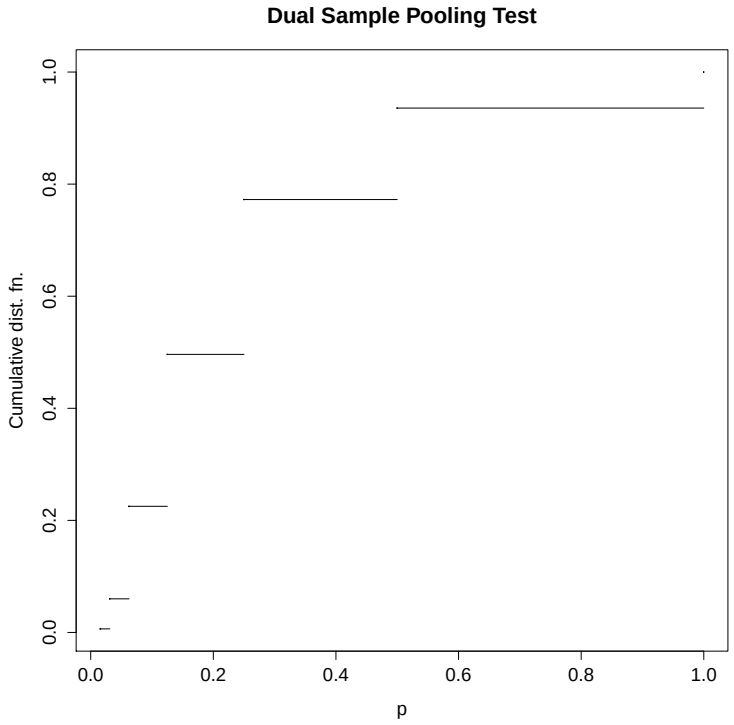


Figure 3.5 Cumulative distribution functions of the fuzzy p-values for Dual Sample Pooling Test based on 2000 iterations

human diagnostics, is easy to collect because pigs naturally investigate their environment by chewing. The question is to compare the probability of detecting IAV in OF and NS specimens collected from vaccinated pigs. IAV vaccinated pigs were inoculated with subtypes H1N1 or H3N2. Pen-based oral fluid samples were collected day post inoculation. There were two pigs in each pen. The OF and NS samples were tested in the laboratory with results to be either negative or positive. Each OF sample from one pen matches with two NS samples from two individual pigs in the same pen. The data are as follows:

Table 3.4 Counting Table for influenza A virus detection

		NS			
		2	1	0	Total
OF	1	114	28	29	171
	0	2	7	42	51
	Total	116	35	81	222

A Fisher's exact test for independence results in a p-value $< 2.2 \times 10^{-16}$, indicating con-

vincing evidence of dependency between tests in the table. Thus Miettinen's test should not be applied in this situation because it is derived under the independence assumption. The whole distribution of the fuzzy p-value is concentrated below 0.05. The asymptotic test p-value is 2.71×10^{-9}. There is very strong evidence for difference between positive rates of the two tests. OF is better than SN in terms of both convenience and sensitivity. This example is not a pooling test as the first example, however, the data also has a matched scheme.

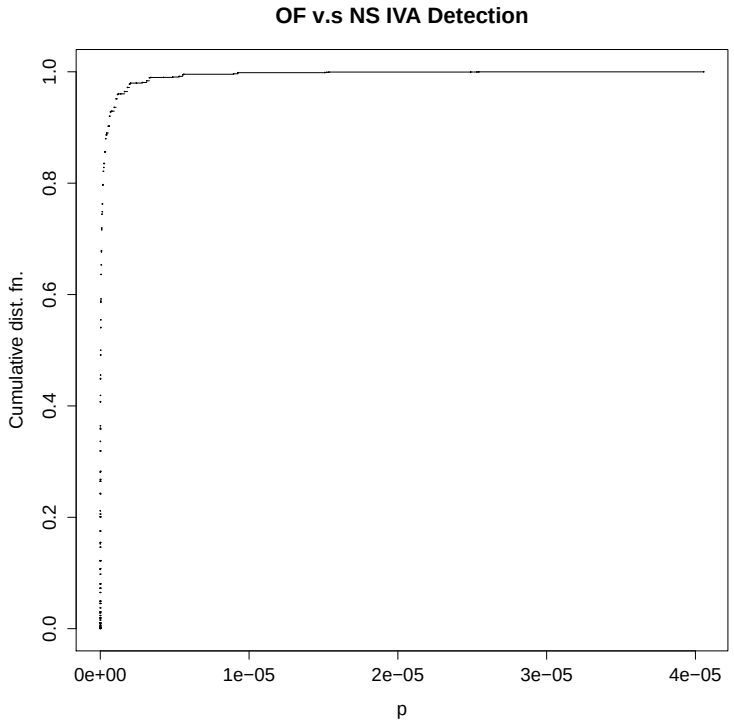


Figure 3.6 Cumulative distribution functions of the fuzzy p-values for Pen-based oral fluid specimens for influenza A virus detection based on 2000 iterations

3.7 Discussion and Conclusion

The simulation esults of above work show that Miettinen's test performs poorly when the multiple observations from the same matched set are dependent. Except for very small numbers of matched sets, in general, the results support that both exact and asymptotic test have good power and control type one error well. Asymptotic test out performs the exact test by effectiveness and computational speed. The estimated power for the asymptotic test based

on 2000 simulated data sets is very close to the calculated results from the power function. The tests proposed in the present work have rather wide applicability in medical and other research. Both the exact and the asymptotic versions of our proposed statistical tests can be generalized from 1-to-2 to 1-to-N matched data. The generalization of the test can be found in A.2. A related question arise: does the exact and asymptotic test remain accurate for $N > 2$? It is a question worthy of future investigation.

CHAPTER 4. MEASUREMENT ERROR IN A BIVARIATE MODEL – APPLICATION IN NUTRITION EPIDEMIOLOGY

Abstract

We consider the problem of estimating the joint distribution of two correlated random variables where one is observed with error. An interest in human nutrition is to estimate the joint distribution of usual energy intake and usual micronutrient intake. While precise biomarkers for energy consumption are available, there are no reliable biomarkers of consumption for nutrients including vitamins and minerals (vitamin K is an exception). Yet, nutritionists are interested in estimating the distribution of usual intake of micronutrients per unit of caloric intake. This is denoted the nutrient density of the diet and involves estimation of the distribution of the ratio of two non-normal random variables, one of which is observed with measurement error. We develop an approach that combines a deconvolution kernel method (DKM) and the method of copulas to estimate the joint distribution of two non-normal variables where one is contaminated. DKM is first used to adjust the univariate measurement error. A Gaussian copula is then used to model the correlation structure between the two variables after error adjustment. We carried out a small simulation study to investigate whether the two-step method we propose is promising. At least in the context of our simulation, we found that the approach produces good results when the correlation between the two random variables is reasonably high. Our findings are tentative, however and more research is needed before we can recommend the methodology for use more broadly.

4.1 Introduction

We consider estimation of a bivariate non-normal distribution given pairs of observations where one of the variables is contaminated with measurement error. This problem, which falls in the general category of a deconvolution problem, arises frequently in applications since, in practice, we often find that variables of interest are subject to measurement error.

Our work is motivated by the need to estimate a bivariate usual intake distribution. There has been a lot of work in this area, but research has mostly focused on the univariate case (see below for a review of some of the literature). For practical reasons – cost and respondent burden among them – intake data are collected from individuals in a sample of the population for only a few days per individual. Even though intake information for each individual in the sample is limited, epidemiologists and nutritionists are typically interested in the long-run average intake, denoted usual intake, and in particular, in the distribution of usual intakes in the population. Given an estimate of the distribution of usual intakes, it is then possible to estimate, for example, the proportion of the population whose intakes fall below a threshold such as the estimated average requirement (EAR). Excessive intakes, as in the case of cholesterol and sodium, are also of interest, and the proportion of the population with high intakes of a nutrient can also be assessed from the usual intake distribution.

One simple approach to estimate the distribution of usual intakes is to use the distribution of observed individual mean intakes as the estimator. However, even if we were to assume that the observed intake is unbiased for usual intake, an individual's mean daily intake for a particular dietary component has a variance that contains some within-individual variability. Thus, the variance of the observed means is inflated by the day-to-day variability in daily intake. Because of this, using the distribution of the mean of a few days as an estimate of the usual intake distribution can lead to erroneous inference regarding dietary status.

In the univariate case, adjustment for measurement error can be formulated as the problem of estimating the distribution of a random variable that is observed with error. In 1986, the National Research Council (NRC, 1986) proposed a simple measurement error model to describe the relation between the observed daily intake for person i on day j, Y_{ij} and the unobservable

usual intake for that person, y_i. In their formulation,

$$Y_{ij} = y_i + u_{ij},$$

where $y_i \sim N(\mu, \sigma_y^2)$ and $u_{ij} \sim N(0, \sigma_u^2)$. The measurement error u_{ij} is assumed to be independent of the unobservable usual intakes y_i and also of each other, within a person. Under the model, $y_i = E(Y_{ij}|i) = E(\bar{Y}_i|i)$, where $\bar{Y}_i$ is the observed mean intake of the ith person calculated over r_i daily intake observations. Further, $Var(\bar{Y}_i) = \sigma_y^2 + \sigma_u^2/r_i$.

The NRC suggested estimating y_i using a best linear unbiased predictor (BLUP) and then estimating the usual intake distribution as the distribution of those BLUPS. Since observed daily intakes Y_{ij} are typically non-normal, the NRC proposed that the model be fit after log-transforming the daily intakes. Nusser et al. (49) revisited this problem and recognized that estimating $f(y)$ is a deconvolution problem. They proposed an approximation to the deconvolution estimate of $f(y)$ that assumed that a univariate transformation of Y_{ij} into the normal scale implies that both y_i and u_{ij} are also normally distributed. In the normal scale, Nusser et al. (1996) fitted the simple measurement error model, estimated the unobservable, normal-scale usual intakes $\ddot{y}_i$ and then, using a suitable back-transformation, obtained the estimated distribution of the y in the original scale.

While the model described above is simple, the areas in which the model can be applied are multiple and include astronomy, biology, chemistry, economy and public health (34), (48). Estimation of the density of a univariate non-normal random variable with measurement error has been extensively studied. Mendelsohn and Rice (1982) (47) presented an example of estimation of a density given observations contaminated with normal error. Stefanski(1990) (54) considered estimation of a continuous bounded probability density when observations from the density are contaminated by additive measurement errors having a known distribution. These studies have focused on the univariate case. An exception is a recent paper by Zhang et al. (52) in which the authors propose a method for estimating a highly multivariate distribution when only short-term measurements are available. Overall, however, there is little work published for the case where the density of interest is multivariate.

We consider the problem of estimating the joint distribution of two correlated random vari-

ables where one of the variables is observed with error. An example in nutrition is estimation of the joint distribution of usual energy intake and usual micronutrient intake. While precise biomarkers for energy consumption are available (e.g., doubly-labeled water, Trabulsi and Schoeller, 2001), there are no reliable biomarkers of consumption for nutrients including vitamins and minerals (vitamin K is an exception). Yet, nutritionists are interested in estimating the distribution of usual intake of micronutrients per unit of caloric intake. This is referred to as the nutrient density of the diet and involves estimation of the distribution of the ratio of two non-normal random variables, one of which is observed with measurement error.

The main objective of this paper is to explore whether the method of copulas can be used to estimate the densities of two non-normal random variables when one is contaminated by normal measurement error. In our set-up, we do not observe the marginal distributions of the two variables, but have access to independent replicate observations, at least of the contaminated variable. While the unobservable bivariate distribution is of interest, we focus on estimation of the density of functions of the two random variables, and in particular, of the ratio of the two random variables. In summary, we develop an approach that combines a deconvolution kernel method (DKM) and the method of copulas to estimate the joint distribution of two non-normal variables where one is contaminated by normal measurement error. DKM is first used to adjust the univariate measurement error. A Gaussian copula is then used to model the correlation structure between the two variables after error adjustment.

This paper is organized as follows. In the next section we describe the model and introduce some notation. We also discuss the methods we propose in this same section. A simulation study is presented in Section 4.3. We investigate the performance of the algorithm we propose in this section, with emphasis on the accuracy with which we can estimate the density of the ratio of the two random variables. Section 4.4 includes a discussion of our findings, and gives some directions for future work.

4.2 Bivariate random measurements with error in one margin

Suppose that we obtain two measurements on the i^{th} sample person on the j^{th} measurement occasion. Let X_{1ij} and X_{2ij} denote the observed values for the i^{th} subject on the j^{th} occasion,

where $i = 1, ..., n;\ j = 1, ..., r_i$. For simplicity, we assume $r_i = r$ for all i. Suppose that X_{1ij} is an almost noise-free measurement of the usual value x_{1i} but that X_{2ij} measures x_{2i} with non-negligible error. A simple model in this case is

$$\begin{bmatrix} X_{1ij} \\ X_{2ij} \end{bmatrix} = \begin{bmatrix} x_{1i} \\ x_{2i} + \epsilon_{2ij} \end{bmatrix},$$

$$\begin{bmatrix} x_{1i} \\ x_{2i} \end{bmatrix} \sim \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right] \quad \text{and} \quad \epsilon_{2ij} \sim N(0, \sigma_\epsilon^2).$$

In this model, (x_{1i}, x_{2i}) are assumed to be independent of ϵ_{2ij}. For a given person i, we also assume that the measurement errors $\epsilon_{2ij}, \epsilon_{2ij'}$ are independent. We wish to estimate $f(x_1, x_2)$ when we have at least one observation for x_1 and more than one independent replicate of X_{2i}. We make no distributional assumptions about X_2 but will assume that the measurement error ϵ is normally distributed.

Let f_W, f_X and f_ϵ denote the densities of X_{2ij}, x_{2i} and ϵ_{2ij}. We propose a method for estimating $f(x_1, x_2)$ that consists of the following steps:

1. We first use the independent replicates $X_{2i1}, ..., X_{2ir}$ to obtain a moment estimator $\hat{\sigma}_\epsilon$ for the measurement error variance σ_ϵ. Then we have that $\hat{f}_\epsilon = \phi(0, \hat{\sigma}_\epsilon)$ where $\phi(\mu, \sigma)$ is the normal density.

2. We then adjust for the measurement error in X_2 using a kernel deconvolution method to estimate the density of x_{2i}, denoted $\hat{f}_X$.

3. We use a copula approach to estimate the conditional density $\hat{f}_{X_{1ij}|X_{2ij}}$.

4. Finally, we draw pairs (x_{1i}, x_{2i}) from their estimated joint density as follows:

   (a) Simulate $\epsilon^*_{2ij} \sim \hat{f}_\epsilon$ and $x^*_{2i} \sim \hat{f}_X$ and compute $X^*_{2ij} = x^*_{2i} + \epsilon^*_{2ij}$, i.e. simulate observations contaminated with error.

   (b) Draw X^*_{1ij} from $\hat{f}_{X_{1ij}|X_{2ij}}$ with $X_{2ij} = X^*_{2ij}$

   (c) Calculate $x^*_{1i} = \frac{1}{r} \sum_{j=1}^{r} X^*_{1ij}$

(d) Repeat a large number of times M to get pairs $(x^*_{1m}, x^*_{2m})_{m=1,\ldots,M}$.

In the next sections, we describe these steps in more detail.

4.2.1 Deconvolution estimator of $f_{X_2}(x_2)$

Let φ_W, φ_X and φ_ϵ denote the characteristic functions of X_{2ij}, x_{2i} and ϵ_{2ij}. Let f_W, f_X and f_ϵ be probability density functions of X_{2ij}, x_{2i} and ϵ_{2ij}, respectively. By the inversion formula,

$$f_X(x) = \frac{1}{2\pi}\int e^{-itx}\varphi_X(t)dt = \frac{1}{2\pi}\int e^{-itx}\frac{\varphi_W(t)}{\hat{\varphi}_\epsilon(t)}dt, \tag{4.1}$$

where we have omitted the subscript 2 to simplify notation. A kernel estimator of $\varphi_W(t)$ is given by

$$\hat{\varphi}_W(t) = \int e^{itw}\hat{f}_W(w)dw \tag{4.2}$$

where $\hat{f}_W(w) = \frac{1}{nh}\sum_{j=1}^n K(\frac{w-W_j}{h})$ is the conventional kernel density estimator of f_W and $K(\cdot)$ is a symmetric probability kernel with finite variance. The resulting estimator of f_X based on $\hat{\varphi}_W(t)$ is the deconvolution kernel density estimator (54)

$$\hat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^n L(\frac{x-W_i}{h}), \tag{4.3}$$

where

$$L(z) = \frac{1}{2\pi}\int e^{-itz}\frac{\varphi_K(t)}{\hat{\varphi}_\epsilon(\frac{t}{h})}dt$$

is called the deconvoluting kernel and is such that φ_K is compactly supported and is the characteristic function of the kernel $K(\cdot)$. The parameter h is the bandwidth parameter. The distribution estimator $\hat{F}_X$ of F_X is thus defined as the integral of $\hat{f}_X$ over $(-\infty, x]$:

$$\hat{F}_X(x) = \frac{1}{2} + \frac{1}{2\pi n}\sum_{j=1}^n \int \frac{sin(t(x-W_j))\varphi_K(ht)}{t\hat{\varphi}_\epsilon(t)}dt. \tag{4.4}$$

We chose a standard kernel function for normal errors, a second-order kernel whose characteristic function has a compact and symmetric support (Fan, 1992) given by

$$K(x) = \frac{48cos(x)}{\pi x^4}(1 - \frac{15}{x^2}) - \frac{144sin(x)}{\pi x^5}(2 - \frac{5}{x^2}). \tag{4.5}$$

The characteristic function of the second-order kernel is given by:

$$\varphi_K(t) = (1 - t^2)^3 I_{[-1,1]}(t). \tag{4.6}$$

The resulting deconvolution kernel when we assume normal errors is therefore:

$$L_1(x) = \frac{1}{\pi} \int_0^1 cos(tx)(1 - t^2)^3 e^{\frac{\sigma^2 t^2}{2h^2}} dt. \tag{4.7}$$

The unknown bandwidth parameter h is difficult to determine from the data. There have been at least three different major approaches proposed to estimate the bandwidth parameter. The cross-validation approach proposed by Habbema, Hermans and Van Der Broek (43) while simple to formulate, has been shown to produce highly variable results (32). An alternative is what is known as 'plug-in' methods, of which there is a wide variety discussed in the literature (46; 32). The approach discussed in Delaigle and Gijbels (32) is based on an asymptotic approximation to the mean integrated squared error (MISE), which we describe below. A third approach to estimating the bandwidth, is also based on the MISE, but instead of relying on an asymptotic approximation of the MISE, it relies on a bootstrap approximation to the MISE (33). Here, we select the bandwidth h by minimizing the asymptotic approximation to the mean integrated error, as in the 'plug-in' method. The (MISE) is defined by

$$MISE(h) = E \int (\hat{f}_X(x, h) - \hat{f}_X(x))^2 dx. \tag{4.8}$$

Stefanski and Carrol (54) showed that an estimate of the MISE is given by:

$$\hat{MISE}(h) = \frac{1}{2\pi nh} \int \frac{|\varphi_K(t)|^2}{|\varphi_\epsilon(\frac{t}{h})|^2} dt + \frac{h^4}{4} R(f_X^{''}) \int x^2 K(x) dx, \tag{4.9}$$

where $R(f_X^{''}) = \int [f_X^{''}(x)]^2 dx$. If we were to assume that x_{2i} is normal, $R(\hat{f}_X^{''}) = 0.375 \hat{\sigma}_X^{-5} \pi^{-\frac{1}{2}}$ where $\hat{\sigma}_X = \sqrt{\hat{\sigma}_W^2 - \hat{\sigma}_\epsilon^2}$, $\hat{\sigma}_W^2$ is the sample variance of X_{2ij} and $\hat{\sigma}_\epsilon^2 = (\sum_{i=1}^n \sum_{j=1}^r (X_{2ij} - \bar{x}_{2i.})^2)(n(r-1))^{-1}$. The plug-in selection of h is the value of the bandwidth that minimizes $\hat{MISE}(h)$.

4.2.2 A copula approach to conditional density estimation

Once we have estimated the marginal densities of x_1 and x_2, we can use the method of copulas to approximate their joint distribution. The history of the copula traces back to Frechet (36). Formally, a copula is a bi-(or multi) variable distribution function whose marginal distribution functions are uniform on the interval [0,1]. Suppose that we have a g-dimensional random vector $(Z_1, Z_2, ..., Z_g)$ with continuous marginal cumulative distribution functions $F_i(z) = P[Z_i \leq z]$. If we apply the probability integral transform to each marginal, the vector

$$\begin{pmatrix} U_1 & U_2 & ... & U_g \end{pmatrix} = \begin{pmatrix} F(z_1) & F(z_2) & ... & F(z_g) \end{pmatrix}$$

has marginal distributions that are uniform. The copula of the vector Z is then defined as the joint cumulative distribution function of the vector U. More formally,

Definition 4.2.1. A g-dimensional copula $C : [0,1]^g \longrightarrow [0,1]$ is a cumulative distribution function with uniform marginals.

Sklar (53) proved the following fundamental result:

Theorem 4.2.2. ***(Sklar1959)*** *Consider a g-dimensional cdf H with marginals $F_1, ..., F_g$. There exists a copula C, such that*

$$H(x_1, ..., x_g) = C(F_1(x_1), ..., F_g(x_g)) \tag{4.10}$$

for all $x_i \in \bar{R}$. If F_i is continuous for all $i = 1, ..., g$ then C is unique; otherwise C is uniquely determined only on $RanF_1 \times \cdots \times RanF_g$, where $RanF_i$ denotes the range of the cdf F_i .

This theorem gives a representation of a multivariate c.d.f as a function of each univariate c.d.f. In other words, the copula function captures the dependence structure among the components irrespective of the marginal distributions.

We estimate the conditional density $f_{x_{1i}|X_{2ij}}$ using a copula. By Theorem 4.2.2, we have that

$$H(X_{1ij}, X_{2ij}) = C(F_1(X_{1ij}), F_2(X_{2ij})), \tag{4.11}$$

where F_1 and F_2 are marginal cumulative density functions of $X_{1i}$and X_{2ij} and H is joint cumulative density function of X_{1i} and X_{2ij}. Then the joint probability density function is:

$$h(X_{1ij}, X_{2ij}) = \frac{\partial^2 H(X_{1ij}, X_{2ij})}{\partial X_{1ij} \partial X_{2ij}} = \frac{\partial^2 C(F_1(X_{1ij}), F_2(X_{2ij}))}{\partial X_{1ij} \partial X_{2ij}} = f_1(X_{1ij}) f_2(X_{2ij}) c(F_1(X_{1ij}), F_2(X_{2ij})), \tag{4.12}$$

and the conditional distribution of x_1 given x_2 is given by

$$f_{x_{1i}|X_{2ij}} = \frac{h(X_{1ij}, X_{2ij})}{f_2(X_{2ij})} = f_1(X_{1ij}) c(F_1(X_{1ij}), F_2(X_{2ij})). \tag{4.13}$$

We used a Gaussian copula to model the correlation structure between X_{1ij} and X_{2ij}, so that

$$C_\rho^{Ga}(F_1(X_{1ij}), F_2(X_{2ij})) = \int_{-\infty}^{\Phi^{-1}(F_1(X_{1ij}))} \int_{-\infty}^{\Phi^{-1}(F_2(X_{2ij}))} \frac{1}{2\pi\sqrt{1-\rho^2}} exp\{-\frac{s^2 - 2\rho st + t^2}{2(1-\rho^2)}\} ds dt. \tag{4.14}$$

The following corrected rank-based estimate was used to estimate the marginal cumulative distribution functions of X_{1ij} and X_{2ij} (55):

$$\hat{F}(x^{(k)}) = \frac{r(k) - 0.326}{n + 0.348}, \tag{4.15}$$

where $r(k)$ is the rank of the k^{th} observation in a vector of observations $\mathbf{x}$. A pseudo(partial)-likelihood for ρ is (Genest et al., 1995):

$$\tilde{ln}(\rho) = \sum_{i=1,\ldots,n; j=1,\ldots,r} ln C_\rho^{Ga}(\hat{F}_1(X_{1ij}), \hat{F}_2(X_{2ij})). \tag{4.16}$$

To estimate ρ we find the value that maximizes the equation (4.16):

$$\hat{\rho} = \underset{\rho}{argsup}\{\tilde{ln}(\rho)\}. \tag{4.17}$$

4.3 Simulation study

We carried out a simulation study to assess the performance of the method we propose to estimate the bivariate density of x_{1i}, x_{2i}. We first generate x_2 from a non-normal distribution as described below. To ensure that the simulated observations are positive, we assume that the additive measurement error model holds after a log transformation of the observations. We generated identically distributed x_{2i} according to (4.18).

$$x_{2i} \sim 2 * \Gamma(5, 2) + \chi(12). \tag{4.18}$$

We considered three different structures for the correlation between x_{1i}, x_{2i}. Under the first correlation structure, x_{1i} and x_{2i} are highly correlated. Under the third structure, x_{1i} and x_{2i} are almost uncorrelated. In the third case, knowing the value of X_{2ij} does not provide much information about the value of x_{1i}. More precisely, the three conditional distributions from which we draw the value of x_{1i} are

1. $x_{1i}|x_{2i} \sim \chi^2(x_{2i})$
2. $x_{1i}|x_{2i} \sim \Gamma(5, 1) + \sqrt{x_{2i}}$
3. $x_{1i}|x_{2i} \sim e^{\Gamma(3,5)} + sin(x_{2i})$.

A graphic illustration of the joint distributions of samples from the three schemes is shown in Figure 4.1 for a single realization.

The simulated observations X_{2ij} are then contaminated by either normal or t errors. Recall that the deconvolution kernel estimator is based on a normal error assumption, so we wished to explore whether the approach we propose is to robust to departures from the normality assumption for the measurement errors. The errors in the study are generated as:

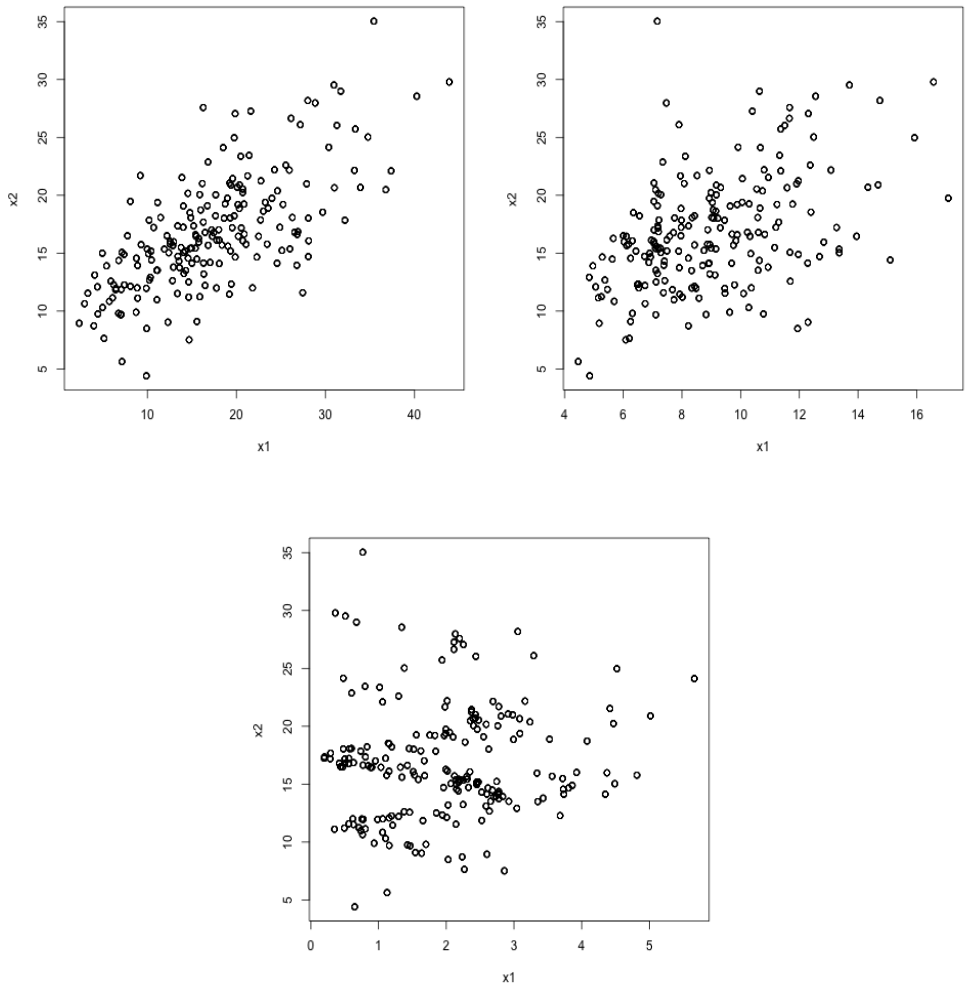
1. $\epsilon_{2ij} \sim N(0, 0.5)$
2. $\epsilon_{2ij} \sim t_3$.

The contaminated observations are then calculated as in (4.19).

$$X_{2ij} = x_{2i} * e^{\epsilon_{2ij}}. \tag{4.19}$$

Finally, we varied the number of individuals and the number of independent replicates available for each individual. We considered the case where we had $n = 200$ individuals, each with $r = 7$ independent replicate observations and the case where we had $n = 350$ individuals, each with $r = 4$ replicate observations. Overall, we considered 12 scenarios and the entire simulation study was repeated 15 times. Except where noted, all results presented below are averaged over the 15 simulation replicates.

Figure 4.1 Joint distribution for simulated x_{1i} and x_{2i}; top-left : $x_{1i}|x_{2i} \sim \chi^2(x_{2i})$; top-right: $x_{1i}|x_{2i} \sim \Gamma(5,1) + \sqrt{x_{2i}}$; bottom: $x_{1i}|x_{2i} \sim e^{\Gamma(3,5)} + sin(x_{2i})$



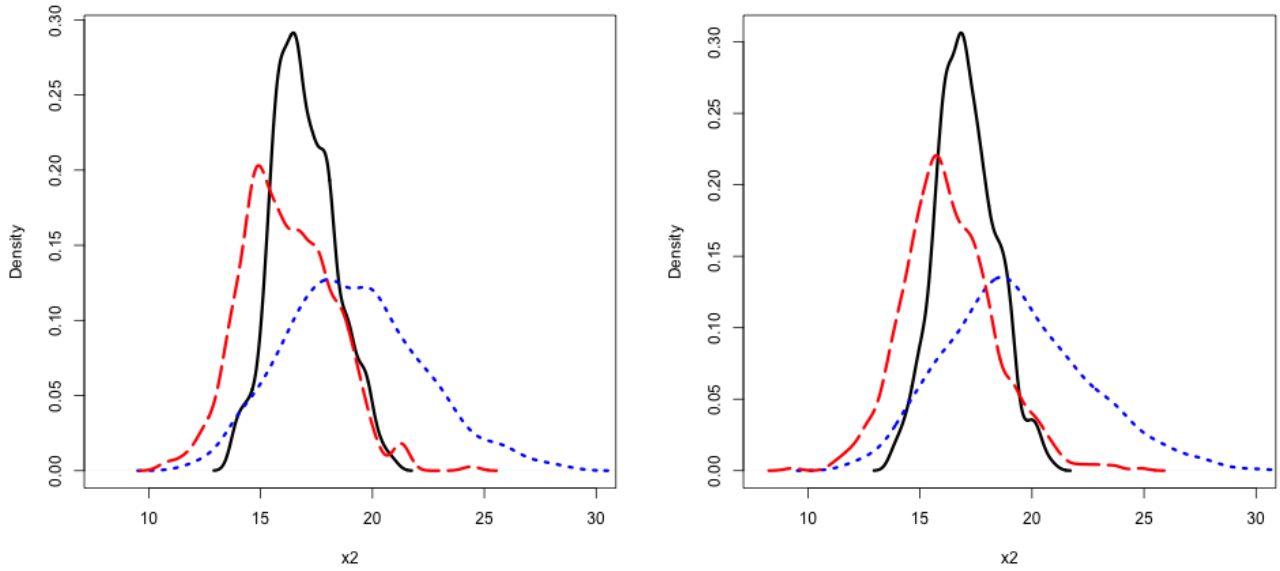
We proceeded as described in Section 4.2.1. To illustrate the performance of the deconvolution kernel estimator, we show the estimated density curves corresponding to different sample

sizes and two error distributions in Figure 4.2 and Figure 4.3. In each case, the average (over 5 simulation replicates) target curve is represented by a solid black line. Figure 4.2 and Figure 4.3 compare, for various sample sizes, the results obtained for estimating densities with respect to the two error distributions. The average deconvolution estimators appear to be more skewed to the right relative to the real values. In our study, violating the normal error assumption appears to significantly affect the performance of the kernel deconvolution estimator as suggested by the density estimators shown in Figure 4.3. However, this is not an issue that we explored in depth and findings are tentative. When the measurement error is normal, we would expect to have better deconvolution estimators when seven (rather than four) independent replicates are available for each sample person, even though the number of individuals in the sample is half as large. This is because the accuracy with which we can estimate the variance of the measurement error depends more directly on the number of replicates within subject than on the number of subjects. Yet Figure 4.2 indicates that there is little advantage – at least in these particular simulation scenarios – in increasing the number of replicates per subject from four to seven.

Table 4.1 contains the mean, variance, and skewness coefficient of the distributions of the true x_{2i}, the x^*_{2i} drawn from the deconvolution estimator of $f(x_2)$ and the distribution of the contaminated sample. Note that, in all cases, the standard deviation of the contaminated sample is larger than that of the sample from the deconvolution estimator. This, in turn, tends to be larger than the standard deviation of the true values. This suggests that the deconvolution estimator of $f(x_2)$ has succeeded in at least partially removing the within-subject variability in the measurements. The mean of the contaminated values X_{2ij} tends to be larger than the means of x_{2i}, x^*_{2i}. This is unexpected at first glance, given that errors are drawn from distributions with zero mean. The reason for the difference in means is that contamination is multiplicative rather than additive (see expression 4.19).

Because the deconvolution estimator of $f(x_2)$ appears to deteriorate significantly when the errors are drawn from a heavy-tailed distribution such as the t_3 distribution, we did not consider these cases further in the simulation study. In the remainder, we present results for the bivariate case, but only when the measurement errors in X_2 are normally distributed. As

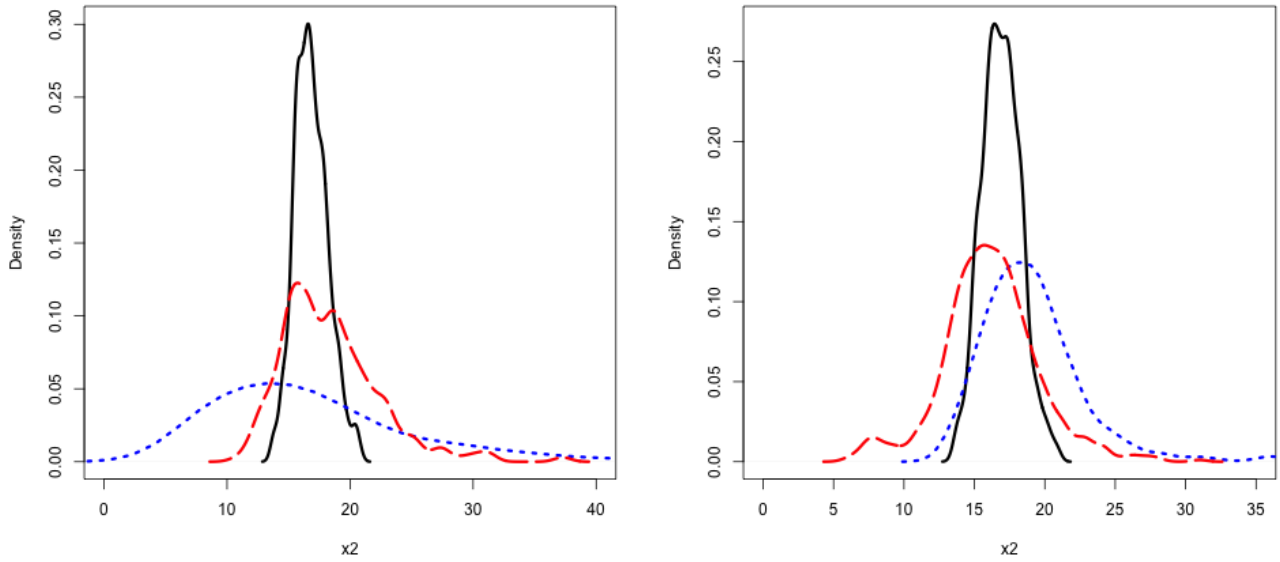
Figure 4.2 Errors are $\epsilon \sim N(0, 0.5)$. Black solid curve is the average (over 15 reps) of the true density of x_2; blue dotted curve is the average of the naive density estimator, ignoring measurement error; red dashed curve is the average of the deconvolution estimator. The left panel corresponds to the case where $n = 200$ and $r = 7$ and the right panel corresponds to the case where $n = 350$ and $r = 4$.



discussed earlier, the distribution of the ratio of two variables is of interest in some practical applications. For example, estimating the population distribution of the usual intake of a nutrient in energy consumption units requires determination of individual-level ratios, i.e the percent of all calories consumed that are attributable to dietary fat, or the usual dietary density of vitamin C consumption per 1000 calories in the diet. We therefore continued with the simulation study and computed the joint distribution of x_1, x_2 for the case where the measurement error is normal, but the strength of the correlation between the two random variables varies from strong, to moderate to weak and for the two sample size scenarios. We then used our estimated joint distribution to obtain the density of the ratio x_2/x_1 to explore how well the estimated ratio density compares to the true ratio density.

Figures 4.4, 4.5 and 4.6 below show the true ratio density (black curve) and the two estimated densities. The red dashed curves are obtained using a deconvolution estimate of $f(x_2)$ and a Gaussian copula estimate of the joint distribution of x_1, x_2. The blue dotted curves are naive estimates of the ratio density, computed as the empirical distribution of the observed

Figure 4.3 Errors are $\epsilon \sim t(3)$. Black solid curve is the average (over 15 reps) of the true density of x_2; blue dotted curve is the average of the naive density estimator, ignoring measurement error; red dashed curve is the average of the deconvolution estimator. The left panel corresponds to the case where $n = 200$ and $r = 7$ and the right panel corresponds to the case where $n = 350$ and $r = 4$.



mean ratios. In the three figures, the left panel corresponds to the case where 7 replicates are available for 200 subjects; the right panel corresponds to the case where 4 replicates are available for 350 subjects.

We note from the figures, that the estimator we propose approximates the true ratio density quite well when the correlation between the two variables is high. The performance of the method, however, deteriorates as the correlation decreases. Tables 4.2 and 4.3 display estimated percentiles of the distribution of the ratio under different simulation scenarios. The mean percentiles and estimated standard deviations were computed over the 5 replicated simulation samples. Overall, our approach performs better than the naive approach, at least when the two random variables are highly or moderately correlated. When the correlation between x_1, x_2 is high, the performance of our approach improves as we approach the upper tail of the ratio distribution; in this case, only the lower tail percentiles of the estimated ratio distribution are significantly different from the true ratio percentiles. Even when the correlation between x_1, x_2 is only moderate or even low and the estimated percentiles are significantly different

Table 4.1 Moments of the distributions of the target values x_{2i}, deconvolution estimates x^*_{2i} and contaminated observations X_{2ij} for different sample sizes and error distributions.

	Mean	Standard Deviation	Skewness
$\epsilon_{2ij} \sim N(0, 0.5), n = 200, r = 7$			
x_{2i}	16.95	1.40	0.30
x^*_{2i}	16.19	2.03	0.30
X_{2ij}	19.12	3.06	0.33
$\epsilon_{2ij} \sim N(0, 0.5), n = 350, r = 4$			
x_{2i}	17.05	1.32	0.20
x^*_{2i}	16.30	2.04	0.45
X_{2ij}	19.27	3.17	0.40
$\epsilon_{2ij} \sim t(3), n = 200, r = 7$			
x_{2i}	16.85	1.35	0.39
x^*_{2i}	18.35	4.00	1.22
X_{2ij}	22.80	53.22	11.71
$\epsilon_{2ij} \sim t(3), n = 350, r = 4$			
x_{2i}	16.90	1.36	0.19
x^*_{2i}	16.14	3.44	0.21
X_{2ij}	21.94	20.41	8.80

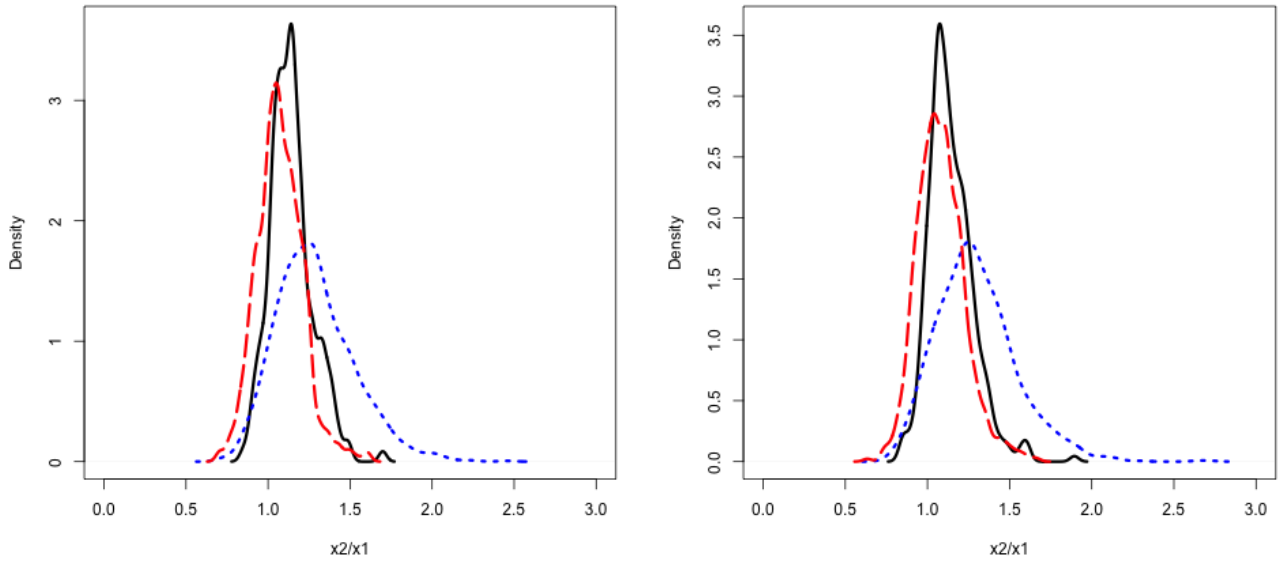
from the true percentiles, the naive estimated distribution has percentiles that are even further away from the true values.

4.4 Discussion

We have proposed an approach to estimate the joint distribution of two non-normal variables when one is contaminated with normal measurement error. The approach consists of two steps. First, we use a deconvolution method to estimate the marginal distribution of the unobservable variable that is observed with error. Next we use a Gaussian copula to estimate the joint distribution of x_1, x_2 using information about the marginals. Copulas are used to model the correlation structure among variables and requires few assumptions about the form of the multivariate distribution to be estimated. Therefore, this approach is applicable more broadly.

Estimation of the marginal distribution of the contaminated random variable is difficult if we wish to minimize assumptions about the form of the unobservable density. Here we have assumed that the errors are normally distributed, but it would be possible, given the inde-

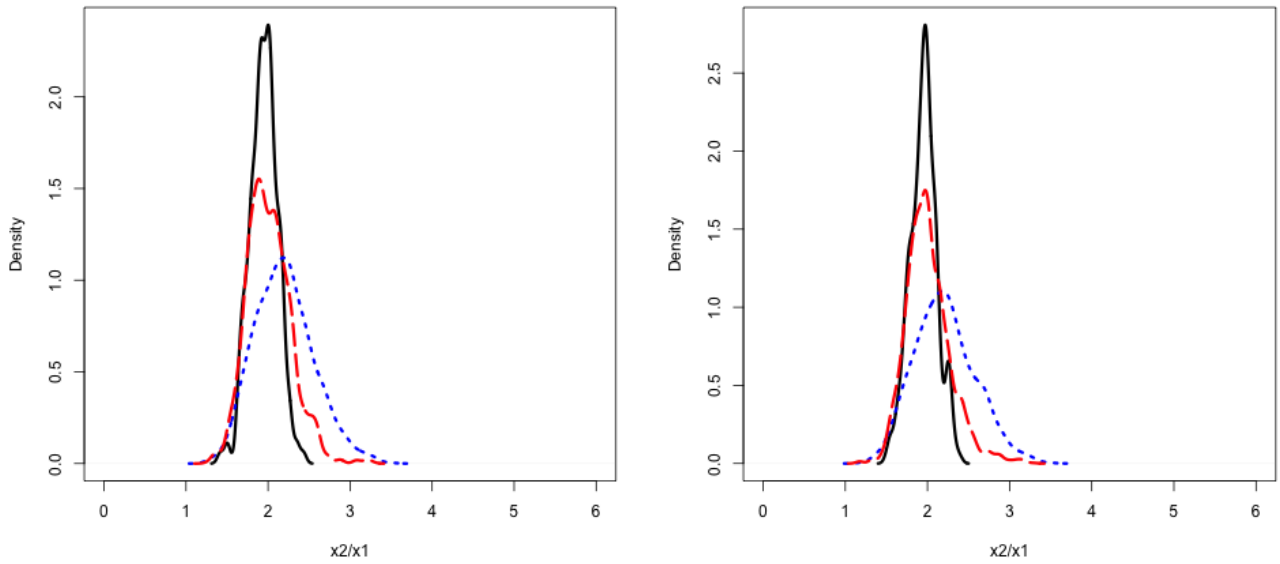
Figure 4.4 Density of the ratio x_2/x_1 with $x_{1i}|x_{2i} \sim \chi^2(x_{2i})$. Left panel corresponds to $n = 200$ subjects with $r = 7$ independent replicates each; right panel corresponds to $n = 350$ subjects with $r = 4$ replicates. The black solid curve is the true density; the blue dotted curve is the density of the observed ratio (ignoring measurement error); the red dashed curve is obtained using the deconvolution estimate of $f(x_2)$.



pendent replicates available for each person, to estimate the distribution of the measurement error empirically. The choice of deconvoluting kernel and of the bandwidth parameter is not straightforward and here we have made choices of convenience. It may be possible to improve the accuracy with which we estimate the marginal distribution of the contaminated random variable. On the other hand, the fact that even choices of convenience greatly improved over the naive estimator of the density suggests that the method we developed might be applicable in a wide range of problems.

The performance of the methods we implement is affected by the degree of association between the two random variables. When the correlation between them is high, the copula approach performs well and the distribution of the ratio of the two variables is closely approximated by the estimated density. When the two random variables are weakly correlated, however, the copula fails, because there is no association to model. In this case, while the estimated ratio density is still a better approximation to the true density that the observed

Figure 4.5 Density of the ratio x_2/x_1 with $x_{1i}|x_{2i} \sim \Gamma(5,1) + \sqrt{x_{2i}}$. Left panel corresponds to $n = 200$ subjects with $r = 7$ independent replicates each; right panel corresponds to $n = 350$ subjects with $r = 4$ replicates. The black solid curve is the true density; the blue dotted curve is the density of the observed ratio (ignoring measurement error); the red dashed curve is obtained using the deconvolution estimate of $f(x_2)$.



empirical density, the performance of the estimator is poor, particular in the tails of the distribution.

Before settling on the deconvolution copula methodology, we investigated an approach that uses a piecewise normal linear approximation to estimate the bivariate density. The method was proposed by Dimitris and Efthymia (2010) and an algorithm to implement the method was presented by Kugiumtzis and Bora-Senta (2010). We found that this approach required tuning a large number of model parameters and that it was difficult to account for the contamination in one margin.

Figure 4.6 Density of the ratio x_2/x_1 with $x_{1i}|x_{2i} \sim e^{\Gamma(3,5)} + sin(x_{2i})$. . Left panel corresponds to $n = 200$ subjects with $r = 7$ independent replicates each; right panel corresponds to $n = 350$ subjects with $r = 4$ replicates. The black solid curve is the true density; the blue dotted curve is the density of the observed ratio (ignoring measurement error); the red dashed curve is obtained using the deconvolution estimate of $f(x_2)$.

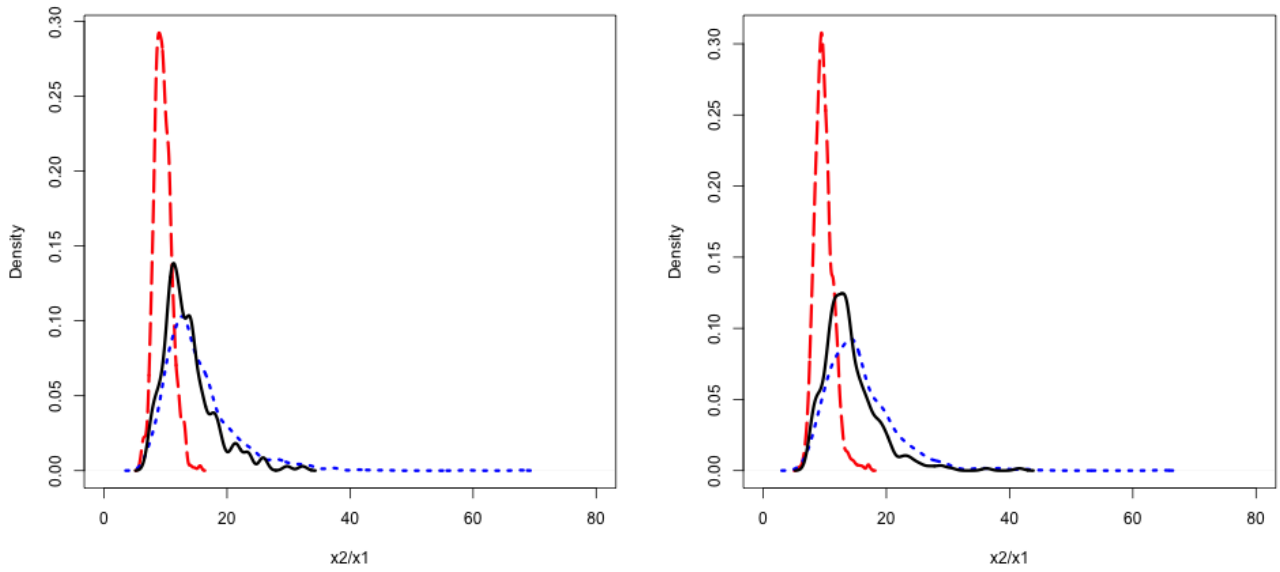


Table 4.2 Percentiles of the ratio $\frac{x_2}{x_1}$ under the three correlation structures. The measurement error distribution is N(0,0.5) and the size is 200 subjects with 7 replicates each. $\hat{r}_k$ is estimated ratio; r_k is the true ratio; r_k^o is the observed ratio with measurement error, and k indicates the corresponding correlation structure.

Quantile	$\hat{r}_1$	r_1	r_1^o	$\hat{r}_2$	r_2	r_2^o	$\hat{r}_3$	r_3	r_3^o
1%	0.3 (0.016)	0.47	0.25	0.55 (0.027)	0.75	0.44	2.35 (0.106)	2.33	1.56
5%	0.44 (0.012)	0.6	0.38	0.81 (0.014)	1.00	0.67	3.57 (0.054)	3.4	2.59
10%	0.53 (0.010)	0.67	0.47	0.96 (0.012)	1.17	0.84	4.28 (0.057)	4.31	3.37
25%	0.7 (0.007)	0.82	0.69	1.28 (0.010)	1.48	1.22	5.88 (0.078)	6.03	5.36
50%	0.96 (0.012)	1.03	1.04	1.78 (0.017)	1.88	1.84	8.32 (0.144)	8.91	9.23
75%	1.31 (0.019)	1.32	1.6	2.5 (0.032)	2.34	2.76	12.02 (0.257)	14.93	17.14
90%	1.72 (0.026)	1.73	2.36	3.28 (0.045)	2.8	3.99	16.2 (0.310)	27.4	32.1
95%	2.05 (0.030)	1.98	2.95	3.89 (0.072)	3.11	4.91	19.5 (0.481)	38.41	47.52
99%	2.88 (0.105)	2.73	4.59	5.59 (0.290)	3.54	7.18	28.17 (1.235)	71.97	101.77

NOTE: Values in parentheses are estimated standard errors for the Monte Carlo mean percentiles.

Table 4.3 Percentiles of the ratio $\frac{x_2}{x_1}$ under three correlation structures. The measurement error distribution is N(0,0.5) and the size is 350 subjects with 4 replicates each. $\hat{r}_k$ is estimated ratio; r_k is the true ratio; r_k^o is the observed ratio with measurement error.

Quantile	$\hat{r}_1$	r_1	r_1^o	$\hat{r}_2$	r_2	r_2^o	$\hat{r}_3$	r_3	r_3^o
1%	0.24 (0.016)	0.48	0.25	0.48 (0.034)	0.74	0.42	2.15 (0.135)	2.32	1.47
5%	0.42 (0.009)	0.60	0.39	0.83 (0.014)	1.00	0.67	3.61 (0.067)	3.48	2.57
10%	0.51 (0.009)	0.68	0.47	0.98 (0.015)	1.16	0.84	4.39 (0.064)	4.30	3.41
25%	0.68 (0.011)	0.82	0.69	1.3 (0.017)	1.47	1.23	5.99 (0.083)	6.09	5.46
50%	0.96 (0.017)	1.03	1.05	1.81 (0.027)	1.88	1.85	8.56 (0.118)	9.15	9.66
75%	1.34 (0.022)	1.33	1.61	2.48 (0.031)	2.36	2.80	12.28 (0.167)	15.59	17.59
90%	1.80 (0.027)	1.71	2.38	3.30 (0.029)	2.81	3.98	16.84 (0.180)	27.34	32.65
95%	2.16 (0.045)	2.00	3.02	3.87 (0.043)	3.12	4.94	20.32 (0.297)	38.79	47.90
99%	3.03 (0.074)	2.91	4.78	5.16 (0.144)	3.76	7.25	29.07 (0.699)	87.75	102.93

NOTE: Values in parentheses are estimated standard errors for the Monto Carlo mean percentiles.

CHAPTER 5. GENERAL CONCLUSIONS

In this dissertation, we develop statistical tools to explore health related problems. We achieved the aim through three papers that discuss disease risk scoring systems, statistical tests for proportion differences in one-to-two matched binary data and bivariate measurement error model for nutrition epidemiology. While we consider specific areas of application, the methods we propose can be applied in multiple areas.

In chapter 2 , we propose to utilize the group lasso algorithm for logistic regression to construct a risk scoring systems for predicting disease in swine (i.e. PRRS). Our proposed method significantly improves upon the current risk scoring system based on expert opinion for predicting whether a swine breeding site experienced a PRRS outbreak. Choice of penalty parameter λ is determined by leave-one-out cross validation with criterion of AUC. The analysis and results demonstrate how a program like PADRAP, that is supported by a professional association and used by a community of veterinarians, can generate valuable data that contributes to our understanding of the relative importance of risk factors and areas of risk factors for clinical outcomes. The results may also be used to decrease the reliance upon expert opinion to identify questions that should remain in the survey and those that may be eliminated to iteratively increase the value of the program and the data.

In chapter 3, we propose exact and asymptotic tests for one-to-two matched binary data. Our methods fit more general situation without assuming observations from the same subject are mutually independent. It can be applied to all kinds of diagnostic studies with one-to-two matched data structure other than dual sample pooling, such as one-to-two case control study etc. The results of simulations study show that Miettinen's test performs poorly when the multiple observations from the same matched set are dependent. Except for very small numbers of matched sets, in general, the results support that both exact and asymptotic test

have good power and control type one error well. Asymptotic test out performs the exact test by effectiveness and computational speed. The estimated power for asymptotic test based on 2000 simulated data sets is very close to the calculated results from the power function. The tests proposed in the present work have rather wide applicability in medical and other research. We have theoretically generalized our tests to one-to-N situation but a related question arise: does the exact and asymptotic test remain accurate for $N > 2$? It is a question worthy of future investigation.

We have proposed an approach in chapter 3 to estimate the joint distribution of two non-normal variables when one is contaminated with normal measurement error. The approach consists of two steps. First, we use a deconvolution method to estimate the marginal distribution of the unobservable variable that is observed with error. Next we use a Gaussian copula to estimate the joint distribution of x_1, x_2 using information about the marginals. Copulas are used to model the correlation structure among variables and requires few assumptions about the form of the multivariate distribution to be estimated. Therefore, this approach is applicable more broadly. Estimation of the marginal distribution of the contaminated random variable is difficult if we wish to minimize assumptions about the form of the unobservable density. Here we have assumed that the errors are normally distributed, but it would be possible, given the independent replicates available for each person, to estimate the distribution of the measurement error empirically. The choice of deconvoluting kernel and of the bandwidth parameter is not straightforward and here we have made choices of convenience. It may be possible to improve the accuracy with which we estimate the marginal distribution of the contaminated random variable. On the other hand, the fact that even choices of convenience greatly improved over the naive estimator of the density suggests that the method we developed might be applicable in a wide range of problems.

The performance of the methods we implement is affected by the degree of association between the two random variables. When the correlation between them is high, the copula approach performs well and the distribution of the ratio of the two variables is closely approximated by the estimated density. When the two random variables are weakly correlated, however, the copula fails, because there is no association to model. In this case, while the

estimated ratio density is still a better approximation to the true density that the observed empirical density, the performance of the estimator is poor, particular in the tails of the distribution.

APPENDIX A. ADDITIONAL MATERIAL

A.1 Tables for difference parameterization

Table A.1 Table for Setting 1, $\delta = 0$

	Test 2		
Test 1	0.01	0.01	0.28
	0.08	0.41	0.21

Table A.2 Table for Setting 2, $\delta = 0$

	Test 2		
Test 1	0.08	0.15	0.07
	0.01	0.27	0.42

Table A.3 Table for Setting 3, $\delta = 0$

	Test 2		
Test 1	0.15	0.24	0.01
	0.01	0.24	0.35

Table A.4 Table for Setting 4, $\delta = 0$

	Test 2		
Test 1	0.01	0.03	0.26
	0.05	0.45	0.10

A.2 Generalize to One to More Matched Test

Both tests can be generalized to one to more matched test, say one to L.

Table A.5 Outcome for Subject j for 1 to L Matched Test

	Test 2			
Test 1	$Z_{1L}^{(j)}$	$Z_{1,L-1}^{(j)}$	...	$Z_{10}^{(j)}$
	$Z_{0L}^{(j)}$	$Z_{0,L-1}^{(j)}$	...	$Z_{00}^{(j)}$

Table A.6 Counting Table for N Sets of Observations for 1 to L Matched Test

		Test 2				
		L	L-1	...	0	Total
Test 1	1	Z_{1L}	$Z_{1,L-1}$	...	Z_{10}	$N_{1.}$
	0	Z_{0L}	$Z_{0,L-1}$	...	Z_{00}	$N_{0.}$
	Total	$N_{.L}$	$N_{.L-1}$	...	$N_{.0}$	N

$$p_1 = \sum_{l=0}^{L} \frac{l}{L}(p_{0l} + p_{1l})$$

$$p_2 = \sum_{l=0}^{L} p_{1l}$$

$$\delta = p_1 - p_2 = \sum_{l=0}^{L} \left\{ \frac{l}{L} p_{0l} - (1 - \frac{l}{L}) p_{1l} \right\}$$

A.2.1 Exact Binomial Test

In this section, we extended the exact binomial test to more general situation. Let $R_{0l}^{(j)} \mid Z_{0l}^{(j)} \sim Bin(Z_{0l}^{(j)}, \frac{l}{L})$ and $R_{1l}^{(j)} \mid Z_{1l}^{(j)} \sim Bin(Z_{1l}^{(j)}, 1 - \frac{l}{L})$.

Claim: $\sum_{l=1}^{L} R_{0l}^{(j)} \sim Ber(\sum_{l=1}^{L} \frac{l}{L} p_{0l})$ and $\sum_{l=1}^{L-1} R_{1l}^{(j)} \sim Ber(\sum_{l=1}^{L-1} (1 - \frac{l}{L}) p_{1l})$

Proof: Since only one cell in Table A.2 is 1 and all the others are 0's. $\sum_{l=1}^{L} R_{0l}^{(j)}$ can only be 0 or 1.

$$\begin{aligned}
Pr\{\textstyle\sum_{l=1}^{L} R_{0l}^{(j)} = 1\} &= \textstyle\sum_{l=1}^{L} Pr\{R_{0l}^{(j)} = 1,\ R_{0l'}^{(j)} = 0\ \forall l' \neq l\} \\
&= \textstyle\sum_{l=1}^{L} Pr\{R_{0l'}^{(j)} = 0,\ \forall l' \neq l \mid R_{0l}^{(j)} = 1\} Pr\{R_{0l}^{(j)} = 1\} \\
&= \textstyle\sum_{l=1}^{L} 1 \cdot Pr\{R_{0l}^{(j)} = 1\} \\
&= \textstyle\sum_{l=1}^{L} \Big[Pr\{R_{0l}^{(j)} = 1,\ Z_{0l}^{(j)} = 1\} + Pr\{R_{0l}^{(j)} = 1,\ Z_{0l}^{(j)} = 0\} \Big] \\
&= \textstyle\sum_{l=1}^{L} \Big[Pr\{R_{0l}^{(j)} = 1,\ Z_{0l}^{(j)} = 1\} + 0 \Big] \\
&= \textstyle\sum_{l=1}^{L} Pr\{R_{0l}^{(j)} = 1 \mid Z_{0l}^{(j)} = 1\} Pr\{Z_{0l}^{(j)} = 1\} \\
&= \textstyle\sum_{l=1}^{L} \frac{l}{L} p_{0l}
\end{aligned}$$

Similarly we can show that $Pr\{\sum_{l=1}^{L-1} R_{1l}^{(j)} = 1\} = \sum_{l=1}^{L-1}(1 - \frac{l}{L}) p_{1l}$.

Under H_0, $\sum_{l=1}^{L-1}(1 - \frac{l}{L}) p_{1l} = \sum_{l=1}^{L} \frac{l}{L} p_{0l}$. Denote $S = \sum_{l=1}^{L} R_{0l}^{(j)} + \sum_{l=1}^{L-1} R_{1l}^{(j)}$, then under H_0 we have $\sum_{l=1}^{L} R_{0l}^{(j)} \mid S \sim Bin(S, \frac{1}{2})$ and exact test can be constructed.

A.2.2 Asymptotic Test

The test statistics

$$T = \sum_{j=1}^{N} T^{(j)} = \sum_{j=1}^{N} \frac{\sum_{l=0}^{L} \left\{ l Z_{0l}^{(j)} + l Z_{1l}^{(j)} - L Z_{1l}^{(j)} \right\}}{L} = \sum_{j=1}^{N} \sum_{l=0}^{L} \left\{ \frac{l}{L} Z_{0l}^{(j)} - (1 - \frac{l}{L}) Z_{1l}^{(j)} \right\}$$

with $T^{(j)} = \sum_{l=0}^{L} \left\{ \frac{l}{L} Z_{0l}^{(j)} - (1 - \frac{l}{L}) Z_{1l}^{(j)} \right\}$ and $E[T^{(j)}] = \delta$.

We can similarly build Asymptotic Test by deriving mean and variance of test statistics.

$$\begin{aligned}
E[T^{(j)2}] &= E[\left\{ \textstyle\sum_{l=0}^{L} \frac{l}{L} Z_{0l}^{(j)} \right\}^2 + \left\{ \sum_{l=0}^{L} (1 - \frac{l}{L}) Z_{1l}^{(j)} \right\}^2 - 2 \left\{ \sum_{l=0}^{L} \frac{l}{L} Z_{0l}^{(j)} \right\} \left\{ \sum_{l=0}^{L} (1 - \frac{l}{L}) Z_{1l}^{(j)} \right\}] \\
&= E[\textstyle\sum_{l=0}^{L} \left\{ \frac{l}{L^2} Z_{0l}^{(j)2} + (1 - \frac{l}{L})^2 Z_{1l}^{(j)2} \right\}] \\
&= \textstyle\sum_{l=0}^{L} \left\{ \frac{l}{L^2} p_{0l} + (1 - \frac{l}{L})^2 p_{1l} \right\}
\end{aligned}$$

$$Var[T^{(j)}] = E[T^{\{j\}2}] - E[T^{(j)}]^2 = \sum_{l=0}^{L} \left\{ \frac{l}{L^2} p_{0l} + (1 - \frac{l}{L})^2 p_{1l} \right\} - \delta$$

$$\mu(\delta) = E[T] = E[\sum_{j=1}^{N} T^{(j)}] = N\delta$$

$$\sigma^2(\delta) = Var[T] = Var[\sum_{j=1}^{N} T^{(j)}] = N \sum_{l=0}^{L} \left\{ \frac{l}{L^2} p_{0l} + (1 - \frac{l}{L})^2 p_{1l} \right\} - N\delta$$

The asymptotic test statistics is: $\frac{|T|}{\sigma(0)}$ which is compared to standard normal distribution.

When $\delta \neq 0$, $\frac{T-\mu(\delta)}{\sigma(\delta)}$ is asymptotic standard normal when n is large by CLT. The power with respect to δ is

$$\beta(\delta) = 2\Phi(\frac{\phi_{\alpha/2}\sigma(0) - \mu(\delta)}{\sigma(\delta)})$$

where $\phi_{\alpha/2}$ is the $\alpha/2$ lower quantile of standard normal distribution and $\Phi(\cdot)$ is the cumulative density function of standard normal distribution.

BIBLIOGRAPHY

[1] P. D. Allison, *Convergence problems in logistic regression*, Micah Altman, Jeff Gill, and Michael McDonald (eds.), New York: Wiley-Interscience (2004), pp. 247-262.

[2] E. Barranger, C. Coutant, A. Flahault, Y. Delpech, E. Darai, and S. Uzan, *An axilla scoring system to predict non-sentinel lymph node ststus in breast cancer patients with sentinel lymph node involvement*, Breast Cancer Res. Tr. 91 (2005), pp. 113-119.

[3] E.R. DeLong, D.M. DeLong and D.L. Clarke-Pearson, *Comparing the areas under two or more correlated receiver operating characteristics curves: a nonparametric approach*, Biometrics 44 (1988), pp. 837-845.

[4] Y. Kim, J. Kim and Y. Kim, *Blockwise sparse regression*, Statist. Sin. 16 (2006), pp. 375-390.

[5] Y. Li, X. Wang, K. Bo, X. Wang, B. Tang, B. Yang, W. Jiang and P. Jiang, *Emergence of a highly pathogenic porcine reproductive and respiratory syndrome virus in the Mid-Eastern region China*, Vet. J. 174 (2007), pp. 557-584.

[6] L. Meier, *grplasso: Fitting user specified models with Group Lasso penalty.*, R package version 0.4-2. (2009) http://cran.r-project.org/package=grplasso

[7] L. Meier, S. van de Geer, and P. Buhlmann, *The group lasso for logistic regression*, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (2008), pp. 53-71.

[8] E.J. Neumann, J.B. Kliebenstein, C.D. Johnson, J.W. Mabry, E.J. Bush, A.H. Seitzinger, A.L. Green and J.J. Zimmerman, *Assessment of the economic impact of porcine reproductive and respiratory syndrome on swine production in the United States*, J. Am. Vet. Med. Assoc. 227 (2005), pp. 385-392.

[9] M.S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, 2003.

[10] E. Rhatigan, I. Tyrmpas, G. Murray and J. N. Plevris, (2010) *Scoring system to identify patients at high risk of oesophageal cancer*, Brit. J. Surg. 97 (2010), pp. 1831-1837.

[11] C. Terpstra, G. Wensvoort G and J. Pol, *Experimental reproduction of porcine epidemic abortion and respiratory syndrome (mystery swine disease) by infection with Lelystad virus: Koch's postualtes fulfilled*, Vet. Quart. 13 (1991), pp. 131-136.

[12] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1996), pp. 267-288.

[13] K.J. Van Zee, D. E. Manasseh, J.L.B. Bevilacqua, S.K. Boolbol, J.V. Fey, L.K. Tan, P.I. Borgen, H.S. Cody III, and M.W. Kattan, *A nomogram for predicting the likelihood of additional nodal metastases in breast cancer patients with a positive sentinel node biopsy*, Ann. Surg. Oncol. 10 (2003), pp. 1140-1151.

[14] G. Wensvoort, C. Terpstra, J. Pol, E.A. Ter Laak, M. Bloemraad, E.P. DeKluyver, C. Kragten and L. Van Buiten, *Mystery swine disease in the Netherlands: the isolation of Lelystad virus*, Vet. Quart. 13 (1991), pp. 121-130.

[15] G.W. Yeo and C.B. Burge, *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals*, J. Computnl Biol. 11 (2004), pp. 475-494.

[16] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. R. Stat. Soc. Ser. B Stat. Methodol. 68 (2007), pp. 49-67.

[17] J.J. Zimmerman , D.A. Benfield , S.A. Dee , P. Murtaugh , T. Stadejek , G.W. Stevenson and M. Torremorell, *Porcine reproductive and respiratory virus (Porcine Aterivirus)*, (2012) In: J.J. Zimmerman , L.A. Karriker, A. Ramirez, K.J. Schwartz, G.W. Stevenson (eds). Diseases of Swine, 10th edition. Wiley-Blackwell, Hoboken NJ, pp. 461-486.

[18] Angulo, F.J. and D.L. Swerdlow. Epidemiology of human Salmonella enteric server Enteritidis in the United States. Iowa State University Press, Ames, 1999

[19] Bennett, B. M. and Underwood, R. E. (1970). On McNemar's test for the 2$\times$2 table and its power function. Biometrics 26, 339-343.

[20] Charles J, Geyer and Glen D. Meeden, Fuzzy and Randomized Confidence Intervals and P-values, . Statistical Science, Vol. 20, No. 4, 358-366. 2005

[21] Dorfman, R. 1943, The Detection of Defective Members of large Populations , Annals of Mathematical Statistics. 14, 436-440

[22] Eugene Litvak, Xin M. Tu, Marcello Pagano, 1994, Screening for the Presence of a Disease by Pooling Sera Samples , Journal of the American Statistical Association, Vol. 89, No. 426, pp. 424-434

[23] Hanson TE; Johnson WO; Gastwirth JL , 2006, Bayesian inference for prevalence and diagnostic test accuracy based on dual-pooled screening , BIOSTATISTICS Volume: 7 Issue: 1 Pages: 41-57

[24] McNemar, Q. 1947, Note on the sampling error of the differences between correlated proportions of percentages, Psychometrika 12, 153-157

[25] Olli S. Miettinen, 1969, Individual Matching with Multiple Controls in the Case of All-or-None Responses, Biometrics, Vol. 25, No. 2, pp. 339-355

[26] Patrick, M.E., P.M. Adcock, T.M. Gomez, S.F.Altekruse, B.H.Holland, R.V. Tauxe and D.L. Swerdlow. Salmonella enteritidis infections, United States, 1985-1999. Emerging Infectious Diseases 10:1-7. 2004

[27] STEFANOS A. ZENIOS AND LAWRENCE M. WEIN , 1998, pooled testing for hiv prevalence estimation: exploiting the dilution effect Statist. Med. 17, 1447-1467

[28] Stephen W. Duffy, 1984, Asymptotic and Exact Power for the McNemar Test and Its Analogue with R Controls Per Case, Biometrics, Vol. 40, No. 4, pp. 1005-1015

[29] S. Vansteelandt, E. Goetghebeur, T. Verstraeten, 2000, Regression Models for Disease Prevalence with Diagnostic Tests on Pools of Serum Samples ,Biometrics, vol 56, No. 4,pp. 1126-1133

[30] Tu XM, Litvak E, Pagano M, 1994, Studies of AIDS and HIV surveillance. Screening tests: can we get more by doing less? ,Stat Med. 1994 Oct 15-30;13(19-20):1905-19

[31] Wesley O. Johnsona; Joseph L. Gastwirth , 2000, Dual group screening ,Journal of Statistical Planning and Inference 83 pp. 449-473

[32] A. Delaigle and I. Gijbels (2002) Practical bandwidth selection in deconvolution kernel density estimation,Preprint submitted to Elsevier Science

[33] A. Delaigle and I. Gijbels (2004) Bootstrap bandwidth selection in kernel density estimation from a contaminated sample, Ann. Inst. Statist. Math. Vol. 56, No. 1, 19-47

[34] Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006). Measurement Error in Nonlinear Models, 2nd ed. Boca Raton, FL: Chapman and Hall CRC Press. MR2243417

[35] Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: Properties and pitfalls. In Dempster, M., editor, Risk Management: Value at Risk and Beyond., pages 176-223. Cambridge Univ. Press, Cambridge.

[36] M. Frechet. (1951) Sur les tableaux de correlation dont les marges sont donnes, Ann. Univ. Lyon, Science, 4, 13-84

[37] Frechet, M. R. (1958). Remarques au sujet de la note precedente. C. R. Acad. Sci. Paris Ser. I Math. 246, 2719-2720.

[38] G. Dall, Aglio. (1972) Frechet classes and compatibility of distribution functions, Symposia Math., 9, 131-150

[39] Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semi-parametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 82, 543-552

[40] Genest, C.,(1987). Frank's family of bivariate distributions, Biometrika 74 (3), 549-555.

[41] Genest, C., Rivest, L.-P., 1993. Statistical inference procedures for bivariate Archimedean copulas. Journal of the American Statistical Association, 88 (423), 1034-1043.

[42] Dimitris Kugiumtzis, Efthymia Bora-Senta (2010) Normal correlation coefficient of non-normal variables using piece-wise linear approximation, Comput Stat 25:645-662

[43] Habbema, J.D., Hermans,J. and van den Broek, K. (1974). A stepwise discrimination analysis program using density estimation. Compstat 1974: Proceedings in computational statistics. (G. Bruckman, ed.) 101-110. Vienna: Physica Verlag.

[44] Johnson N, Kotz S (1970) Distributions in statistics, continuous univariate distributions. Houghton Mifflin Company, Boston

[45] Joe, H. (1997): Multivariate Models and Dependence Concepts. Chapman & Hall, London.

[46] M. C. Jones; J. S. Marron; S. J. Sheather (1996) A Brief Survey of Bandwidth Selection for Density Estimation, Journal of the American Statistical Association, Vol. 91, No. 433. , pp. 401-407.

[47] Mendelsohn,J. and Rice, J. (1982) Deconvolution of microfluorometric histograms with B splines. Journal of the American Statistical Association, 77 748-753.

[48] Merritt, D.(1997), Recovering velocity distributions via penalized likelihood. Astronomical J. 114 228-237.

[49] Nusser, S. M., A.L. Carriquiry, W.A.Fuller, and K.W.Dodd. 1996a. A semiparametric approach to estimating usual intake distributions. Jounal of the American Statistical Association.

[50] Oakes,D., (1982). A model for association in bivariate survival data, Journal of the Royal Statistical Society Series B 44, 414-422.

[51] Paul Embrechts, Filip Lindskog and Alexander McNeil,(2001), Modelling Dependence with Copulas and Applications to Risk Management, Working paper, ETH, Zurich, http://www.math.ethz.ch/Finance

[52] Saijuan zhang, Douglas Midthune, Patricia M. Guenther, Susan M. Krebs-Smith, Victor Kipnis, Kevin W. Dodd, Dennis W. Buckman, Janet A. Tooze, Laurence Freedman and Raymond J. Carroll (2011), A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment, The Annals of Applied Statistics, Vol.5,No.2B, 1456-1487

[53] Sklar, A. (1959). Fonctions de repartition a n dimensions e leurs marges. Publications de l'Institut de Statistique de l'Univiversite de Paris 8, 229-231.

[54] Stefanski, L.A. and Carroll.,R.J. (1990) Deconvoluting kernel density estimators. Statistics 2, 169-84

[55] Yu GH, Huang CC (2001) A distribution free plotting position. Stoch Environ Res Risk Assess 15:462-476.