

Abstract

Title of dissertation: Out-of-Sample Fusion

Wen Zhou, Doctor of Philosophy, 2013

Dissertation directed by: Benjamin Kedem, Department of Mathematics

A novel method, called “out-of-sample fusion”, is proposed in this dissertation. This method utilizes artificial samples along with a real data sample of interest to draw statistical inference assuming a semiparametric density ratio model. These artificial samples do not relate directly to the sample of interest, which differentiates the method from the traditional bootstrap approach which is a “within-sample” method. Out-of-sample fusion has been elaborated on through the estimation of threshold probabilities and their confidence intervals. A comparison has been made with the Agresti-Coull and the standard Wald methods in terms of confidence interval estimation. The out-of-sample fusion generates sharper and shorter confidence intervals while the nominal coverage is maintained. The out-of-sample method has been applied to cancer and microarray data. An R package has been developed to facilitate the implementation of the out-of-sample fusion method.

Keywords: Density ratio model; Semiparametric; Bootstrap; Biased Sampling; Agresti-Coull; Fusing; Empirical; Expected length; Coverage Probability

Out-of-Sample Fusion

Wen Zhou

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

2013

Advisory Committee:

Professor Benjamin Kedem, Chair/Advisor

Professor Paul Smith

Professor Doron Levy

Professor Myron Katzoff

Professor Francis Alt

© Copyright by
Wen Zhou
2013

Dedication

To: All my family ——

Huanyu Chen, James Zhou, and Thomas Zhou

Acknowledgments

I owe my gratitude to all the people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I'd like to thank my advisor, Professor Benjamin Kedem for giving me an invaluable opportunity to work on challenging and interesting projects over the past four years. He is not just an academic advisor for me, he is a guide for my career and life. It has been a pleasure to work with and learn from such an extraordinary individual.

I would also like to thank all my professors in the department of mathematics and Public Health Center of University of Maryland, where I learn the statistics and mathematics. Here I mention some of them: Dr. Paul Smith, Dr. Eric Slud, Dr. Grace Yang and Dr. Meiling Ting Lee. Special thanks should go to Dr. Hector Bravo, who let me use his microarray data for colon cancer. I also thank USDA/NIFA grant under SCRI project for supporting my research.

I also want to thank my family, without their support, patience, understanding and love, I could not finish my dissertation. At last, thank all my committee members for devoting their precious time to review my dissertation.

Table of Contents

List of Tables	vii
List of Figures	viii
1 Motivation and Overview	1
1.1 Motivation	1
1.2 Overview	3
2 Out-of-Sample Fusion	6
3 Semiparametric Density Ratio Models	8
3.1 Semiparametric Density Ratio Models	8
3.2 Distribution Estimation for Density Ratio Models	12
3.3 Some asymptotic results of density ratio models	15
4 Threshold Probabilities and Confidence Intervals	19
4.1 Threshold Probabilities	19
4.2 Confidence Intervals	20
4.2.1 CI based on maximum likelihood methods	20
4.2.2 Empirical Wald Interval	21
4.2.3 Clopper-Pearson Interval	22
4.2.4 Wilson score Interval	23
4.2.5 Agresti-Coull Interval	24
5 Density Ratio Models with Out-of-Sample Fusion	28

5.1	OSF DR Model Procedure	29
5.2	Comparing CI_{OSF} to CI_{MLE}	31
5.3	Comparing CI_{OSF} to CI_{EP}, CI_{AC}	33
5.4	Comparing CI_{OSF} to CI_{AC}	38
5.4.1	Diagnostic plots for the goodness-of-fit test	39
5.4.2	Lengths of CI_{OSF}, CI_{AC} and CI_{EP}	40
5.4.3	Coverage Probabilities of CI_{OSF}, CI_{AC} and CI_{EP}	43
5.5	Misspecified OSF DR models	48
5.6	Guideline for CI_{OSF}	49
6	Density Ratio Models with Repeated Out-of-Sample Fusion	53
6.1	Introduction	53
6.2	Procedure to obtain CI_{ROSF} and CI_{HB}	56
6.3	Coverage and length of CI_{HB}	57
6.3.1	Guidelines for the choice of \mathbf{W}	64
7	Analysis of the Testicular Germ Cell Tumor (TGCT) Data	67
7.1	Introduction	67
7.2	Descriptive Analysis of the TGCT data	68
7.3	Analysis of TGCT data	71
7.3.1	Bivariate OSF DR Model for TGCT Data	72
7.3.2	Trivariate OSF DR models for TGCT data	73
7.4	Summary	75

8	Analysis of Microarray Data of Colon Cancers	82
8.1	Introduction	82
8.2	Description of Microarray Data of Colon Cancer	83
8.3	Microarray Data Analysis	84
8.4	Results and Discussion	85
8.5	Summary	87
A	R package: Density Ratio	92
A.1	Summary of functions in the <i>DensityRatio</i> package	92

List of Tables

5.1	Coverages and lengths of 100 CIs obtained from OSF, AC and EP methods	45
5.2	Mean coverage and widths resulting from 100 runs.	46
6.1	95% confidence intervals for $R(T) = 0.05$ using three methods. The experiment was repeated 100 times of which five typical cases are listed here. Efficiency is the ratio of lengths relative to that of CI_{AC} , $\mathbf{x}_0 \sim \mathcal{N}(0, 1)$, and all samples are of size 50.	57
6.2	Coverage and average length from 100 runs for nominal 95% confidence intervals for $R(T) = 0.05$, In (*) the fusion samples were changed to $\mathbf{x}_1 \sim \text{Unif}(0, 3 + i)$ from $\mathbf{x}_1 \sim \text{Unif}(-2 - i, 3 + i)$, $1, \dots, 49$. All sample sizes are 50.	58
7.1	TGCT data set	68
7.2	TGCT case-control summary statistics.	69
7.3	joint distribution of $P(\mathbf{Height} \leq Height, \mathbf{Weight} \leq Weight)$	73
7.4	The fixed variables and their values in contour plots	75
8.1	Part of colon cancer microarray data : 5339 genes and 68 subjects: 30 in control, 38 in case	84
8.2	Ranking of gene or gene groups according to their significantly differential expressions between case and control of colon cancer microarray data.	87

List of Figures

5.1	Goodness-of-fit test: $\hat{G}(t)$ versus $\tilde{G}(t)$	39
5.2	Comparison of nominal 95% CI_{OSF} , CI_{AC} and CI_{EP}	41
5.3	Comparison of the standard deviations of nominal 95% CI_{OSF} , CI_{AC} and CI_{EP} (corresponding to Figure 5.2).	41
5.4	Standard deviation plot for 95% CI_{OSFs} which are obtained by using different number of artificial samples.	42
5.5	Comparison of coverage and length distribution of CI_{OSF} , CI_{EP} and CI_{AC} , Note: SP is OSF here.	44
5.6	Mean coverage and lengths of CI_{OSF} , CI_{AC} and CI_{EP} when the sam- ple sizes are 50, 100,500,2000; each point averages results from 100 runs.	47
5.7	Mean coverage and widths of 100 CIs obtained from OSF, AC and EP methods,sample size:100, artificial samples involved are from ex- ponential, normal and t(5) distribution, the reference sample is from $\mathcal{N}(0, 1)$	50
5.8	Mean coverage and widths of 100 CIs obtained from OSF, AC and EP methods, sample size:100, artificial samples involved are from exponential, binomial, poisson and t(5) distribution, the reference sample is from $\mathcal{N}(0, 1)$	51

6.1 CI Coverages and length distribution from 100 runs. **Running condition:** $\mathbf{x}_0 \sim \mathcal{N}(0, 1)$, $\mathbf{w} = (0.40, 0.40)$; **Summarized results:** CI_{ROSF} 75%, CI_{AC} 97%, CI_{HB} 97%; $R(\bar{T})_{ROSF} = 0.059$, $R(\bar{T})_{AC} = 0.084$; SP intervals here are CI_{ROSF} s. 59

6.2 CI Coverage and length distribution from 100 runs. **Running condition:** $\mathbf{x}_0 \sim \text{Logistic}(0, 1)$, $\mathbf{w} = (0.4, 0.4)$; **Summarized results:** CI_{ROSF} 58%, CI_{AC} 97%, CI_{HB} 96%; $R(\bar{T})_{ROSF} = 0.050$, $R(\bar{T})_{AC} = 0.080$; SP intervals are CI_{OSF} here. 60

6.3 CI Coverages and length distribution from 100 runs. **Running condition:** $\mathbf{x}_0 \sim \text{Uniform}(0, 50)$, $\mathbf{w} = (0.40, 0.40)$; **Summarized results:** CI_{ROSF} 36%, CI_{AC} 97%, CI_{HB} 95%; $R(\bar{T})_{ROSF} = 0.046$, $R(\bar{T})_{AC} = 0.083$; SP intervals are CI_{ROSF} 61

6.4 CI Coverages and length distribution from 100 runs. **Running condition:** $\mathbf{x}_0 \sim \Gamma(5, 0.5)$, $\mathbf{w} = (0.40, 0.40)$; **Summarized results:** CI_{ROSF} 31%, CI_{AC} 96%, CI_{HB} 95%; $R(\bar{T})_{ROSF} = 0.047$, $R(\bar{T})_{AC} = 0.084$; SP intervals are CI_{OSF} here. 62

6.5 CI Coverages and length distribution from 100 runs. **Running condition:** $\mathbf{x}_0 \sim \text{Exp}(1)$, $\mathbf{w} = (0.40, 0.40)$; **Summarized results:** CI_{ROSF} 31%, CI_{AC} 97%, CI_{HB} 93%; $R(\bar{T})_{SP} = 0.046$, $R(\bar{T})_{AC} = 0.077$; SP intervals are CI_{OSF} here. 63

6.6 Coverage of CI_{HB} as a function of $\mathbf{w} = (w_1, w_2)$, $\mathbf{x}_0 : \mathcal{N}(0, 1)$, x axis is w_1 and y axis is w_2 65

6.7	Coverage of CI_{HB} as a function of $\mathbf{W} = (w_1, w_2)$, top left: \mathbf{x}_0 : logistic(0, 1); top right: \mathbf{x}_0 : Uniform(0, 50); bottom left: \mathbf{x}_0 : $\Gamma(0, 1)$; bottom right: \mathbf{x}_0 : $\exp(0, 1)$	66
7.1	Histograms for case (right) and control (left)	70
7.2	Scatterplot matrix for case (bottom) and control (top)	70
7.3	3D plot for kernel density estimates of Height and Weight	74
7.4	Contour plots for the control pdf for fixed weights	76
7.5	Contour plots for the case pdf for fixed weights	77
7.6	Pdf plots conditional on Age and Height. Control: solid line; Case: dashed line.	78
7.7	Contour plots for the control pdf for fixed Heights	79
7.8	Contour plots for the case pdf for fixed Heights	80
7.9	Pdf plots conditional on Age and Weight. Control: solid line; Case: dashed line.	81
8.1	The scatterplots of 4 gene pairs which are most significantly differ- entially expressed. The linear regression line and GAM line are also plotted to illustrate the trend.	88
8.2	The contour plots of joint <i>pdf</i> for No.1 and No.2 gene pairs which are most significantly differentially expressed. Left: Control; Right: Case. Top: group 1; Bottom: group 2.	89

8.3	The contour plots of joint <i>pdf</i> for No. 3 and No. 4 gene pairs which are most significantly differentially expressed. Left: Control; Right: Case. Top: group 3; Bottom: group 4.	90
8.4	The scatterplots of first 4 three-gene groups which are most significantly differentially expressed. The right are the observation from the control	91

Chapter 1

Motivation and Overview

1.1 Motivation

This study was motivated by two different problems. First the problem of designing metrics for food safety where the probability of contamination is very small and needs precise interval estimation. Second is the need to estimate small probabilities for bio-surveillance and public health policy with precision. For example, estimating the probability that a certain cold medication sales exceed a high threshold. In this scenario there is a great difference between probabilities 0.01 and 0.001 for large populations.

Another impetus for the present work is the desire to "spring out of the sample" as it were for the purpose of estimating small probabilities. We note that bootstrap methods "cannot go to the tail" as they are confined by the available data. However, out-of-sample fusion (OSF) gives information about the tails since all distributions are estimated from all available data supported over a wider range than that of the reference sample.

Lastly, as shall demonstrate, misspecified fusion may still lead to good estimates. That is, often despite of fusion with samples which do not conform to density ratio models (DRM) the end product is strikingly similar to that obtained from samples which do satisfy DRM requirements.

1.2 Overview

In this dissertation, a new statistical method, an out-of-sample density ratio model (OSF DR model), is proposed to obtain shorter yet reliable confidence intervals (CI_{OSF}) for threshold probabilities by fusing artificial data with a reference real data sample. The resulting CI_{OSF} is compared extensively with the alternative CI's maximum likelihood confidence interval (CI_{MLE}), empirical Wald confidence interval (CI_{EP}), and Agresti-Coull confidence interval (CI_{AC}), in terms of length and coverage of the true probability values.

This dissertation is organized as follows: The first chapter explains the notion of out-of-sample fusion and differentiates it from the traditional “within-sample” methods. The second chapter briefly reviews semiparametric density ratio methodology and focuses primarily on certain asymptotic results which are essential for this dissertation. Some available approaches of constructing confidence intervals, especially the Wald interval and the CI_{AC} are reviewed in Chapter 4.

Chapter 5 compares the out-of-sample fusion with several available methods in terms of confidence interval estimation. Theorem 5.3.1 states that the asymptotic confidence intervals of threshold probabilities resulting from density ratio models are always shorter than those obtained from the empirical approach (i.e. Wald) and Agresti-Coull methods. The ensuing simulation studies show that although the intervals resulting from out-of-sample fusion are considerably shorter than those from the Agresti-Coull (AC) method, the nominal confidence levels specified are still maintained. This chapter is concluded with a guideline for the practical use of

out-of-sample fusion.

In chapter 6, repeated out-of-sample fusion (ROSF) is introduced. It is considered a derivative and ramification of out-of-sample fusion. It repeatedly fuses a given sample with different sets of artificial samples, which leads to a collection of probability estimates after applying the density ratio model separately for each artificial sample. Confidence intervals can be constructed from these estimates. The resulting confidence intervals are even shorter than those obtained from a typical out-of-sample fusion method, but their coverage of the true probability values are often significantly below the nominal coverage. In order to overcome this problem, new confidence intervals are proposed by hybridizing these short intervals with the intervals computed by the AC method. Although the notion of the hybridization has not been established theoretically, some empirical guidelines are suggested from numerical studies.

In chapter 7, a multivariate out-of-sample fusion density ratio model has been applied to case-control cancer data. The results not only confirm the perviously published results , but also illustrate a graphical approach to compare conditional pdfs from case-control data. The difference between the case and control can be appreciated graphically.

Chapter 8 discusses the application of the density ratio model and out-of-sample fusion to microarray data related to colon cancer. It not only identifies the most significantly differentially expressed gene groups which are made of 2 or 3 member genes, but also shows that different association may exist among member genes in the same gene group. The appendix includes a manual for a new R package

called DensityRatio which implements the density ratio model in general with real or artificial data. All results, tables and figures in this dissertation were generated by this package, which is also useful for traditional density ratio model analysis.

Chapter 2

Out-of-Sample Fusion

Utilizing artificial data is common in statistics. Monte Carlo methods (Metropolis and Ulam, 1949), such as importance sampling, and Markov Chain Monte Carlo, use artificial data. In this spirit, this dissertation proposes a novel method, called “out-of-sample fusion”, to utilize external artificial samples assuming a density ratio model. Unlike a typical bootstrap approach (Efron, 1979), which involves repeated sample-generating strictly “within the original sample”, out-of-sample fusion takes advantage of externally generated artificial samples which can be performed by the computer independently. The artificial samples may resemble the original sample in terms of underlying distribution, moments and certain statistical features. However, they may differ from the original sample dramatically, as will be demonstrated in the next chapters.

The out-of-sample fusion method assumes a density ratio model. In this dissertation, a density ratio model in conjunction with out-of-sample fusion is called an OSF DR model. In a traditional density ratio model, inference is drawn from many samples, which are the various sets of real observations or measurements in the same study. They are often the measurements from multiple instruments. In contrast, in the OSF DR model, the problem is to draw inference about a single

population represented by the reference sample. The rest of the samples required by the OSF DR model are generated by the computer. Since more data are used to draw inference, the resulting estimates are considered more efficient than those obtained from traditional methods only using the reference sample alone, provided that the density ratio model assumption holds.

It is worth mentioning here, although certain distributions are employed to generate the artificial samples, they are not pursued during OSF DR model implementations. The only underlying assumption is a specified tilt function described in the coming chapter.

The OSF DR model can be extended as a repeated out-of-sample fusion (ROSF DR model) by repeatedly fusing the reference sample with different sets of artificial samples to obtain the desired estimates. It is just like viewing the reference sample from the different angles. Thus, it leads to a more thorough interpretation for the reference sample. However, confidence intervals for the threshold probabilities obtained with this method do not have satisfactory coverage which is often less than nominal. We propose a strategy to maintain the optimal coverage with the aid of the Agresti-Coull method. It turns out that the resulting hybrid confidence intervals are about 15-30 % shorter than those derived from the AC method alone. A full examination of these points is given later in Chapter 4.

Implementation of OSF DR models using multivariate cancer data and high dimensional genomic data are described in Chapter 6 and 7, respectively.

Chapter 3

Semiparametric Density Ratio Models

This chapter briefly reviews the underpinnings of the traditional density ratio model which can be considered as a special case of biased sampling models. Our OSF DR model idea is an offshoot of the traditional density ratio model.

3.1 Semiparametric Density Ratio Models

The origin of density ratio models can be traced back to any one of Cox's linear models of the log density ratio (Cox, 1969), Anderson's generalized logistic models for multiple population mixtures (Anderson, 1972) and Vardi's length biased models (Vardi, 1982, Vardi, 1985). However it is more intuitive to link them to biased sampling models. Vardi studied a length-biased sampling model (Vardi, 1982). If the selection probability for any particular object is proportional to its length, then the following model gives the distribution of the length of sampled objects:

$$F(y) = \frac{1}{\mu} \int_0^y x dG(x), \quad y \geq 0 \tag{3.1}$$

where

$$\mu = \int_0^{\infty} x dG(x).$$

Here the cdf G is unknown and is to be estimated. The cdf F , the length-biased distribution corresponding to G , is a weighted version of G in terms of the weight function x . In Vardi's original treatment (Vardi, 1982), the weight functions were assumed completely known. But there are many practical situations in which a complete specification of the weight function is too restrictive and mostly unrealistic. One way to relax the assumption on the weight functions is to assume the weight functions belong to a parametric family. These models are called semiparametric biased sampling models. (Gill et al., 1988; Gilbert, 2000):

$$F_i(y) = W_i(G)^{-1} \int_{-\infty}^y w_i(x) dG(x), \quad i = 1, \dots, s \quad (3.2)$$

$$W_i(G)^{-1} = \int_{-\infty}^{\infty} w_i(x) dG(x), \quad i = 1, \dots, s \quad (3.3)$$

A density ratio model is often considered as a special case of a biased sampling model with a parametrized weight function:

$$w_i(x) = \exp\{\alpha + \beta \cdot h(x)\} \quad (3.4)$$

It can also be motivated from a case-control study in which the case sample is assumed to be a weighted control sample with a weight function $\exp\{\alpha + \beta \cdot$

x }. For example, Prentice and Pyke, 1979; Qin and Zhang, 1997, and Qin, 1998, assumed

$$g_1\{x|D = 1\} = e^{\alpha+\beta \cdot x} \cdot g_0\{x|D = 0\}, \quad (3.5)$$

where $D = 0$ stands for the control and $D = 1$ for the case. It is easy to see that the biased sampling model here is actually a logistic regression model. It means that the logistic regression model for a case-control study is equivalent to the biased sampling model with a weight function $\exp\{\alpha + \beta \cdot x\}$.

Motivated by either biased sampling models or case-control studies, density ratio models were developed and elaborated (Qin, 1993; Qin and Lawless, 1994; Qin and Zhang, 1997; Qin, 1999; Zhang, 2000; Fokianos et al., 2001; Fokianos, 2004; Kedem et al., 2008; Kedem et al., 2009, and Voulgaraki et al., 2012). For the two-sample case:

$$\mathbf{x}_1 = (x_{11}, \dots, x_{1n_1})' \sim g_1(x),$$

$$\mathbf{x}_0 = (x_{00}, \dots, x_{0n_0})' \sim g_0(x),$$

the density ratio model is:

$$\frac{g_1(x)}{g_0(x)} = e^{\alpha+\beta \cdot h(x)}, \quad (3.6)$$

and $h(x)$ is called as a tilt function, which can be regarded as distortion of sample \mathbf{x}_1 's *pdf* from the reference sample \mathbf{x}_0 's *pdf*. Actually this model is quite intuitive since the density ratio of two pdf's has the form (2.6) if both of them come from the same exponential family. Thus, if both $g_1(x, \boldsymbol{\theta})$ and $g_0(x, \boldsymbol{\theta})$ are from an exponential

family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subset \mathbb{R}^k$:

$$P_{\boldsymbol{\theta}} = p(x, \boldsymbol{\theta}) = d(\boldsymbol{\theta})S(x) \exp \left[\sum_{j=1}^k C_j(\boldsymbol{\theta})T_j(x) \right], \quad x \in \chi \subset \mathbb{R}^q,$$

where C_1, \dots, C_k and d are real-valued functions of $\boldsymbol{\theta}$, and real-valued functions T_1, \dots, T_k and S have their supports on \mathbb{R}^q . Then,

$$\begin{aligned} \frac{g_1(x)}{g_0(x)} &= \frac{d(\boldsymbol{\theta}_1)}{d(\boldsymbol{\theta}_0)} \cdot \exp \left\{ \sum_{j=1}^k [C_j(\boldsymbol{\theta}_1) - C_j(\boldsymbol{\theta}_0)] \cdot T_j(x) \right\} \\ &= \exp \left\{ \sum_{j=1}^k [C_j(\boldsymbol{\theta}_1) - C_j(\boldsymbol{\theta}_0)] \cdot T_j(x) + \log \left\{ \frac{d(\boldsymbol{\theta}_1)}{d(\boldsymbol{\theta}_0)} \right\} \right\} \\ &= \exp \{ \alpha + \boldsymbol{\beta} \cdot \mathbf{h}(x) \} \end{aligned}$$

Where,

$$\begin{aligned} \alpha &= \log \left\{ \frac{d(\boldsymbol{\theta}_1)}{d(\boldsymbol{\theta}_0)} \right\}, \\ \boldsymbol{\beta} &= \{C_1(\boldsymbol{\theta}_1) - C_1(\boldsymbol{\theta}_0), \dots, C_k(\boldsymbol{\theta}_1) - C_k(\boldsymbol{\theta}_0)\}, \\ \mathbf{h}(x) &= \{T_1(x), \dots, T_k(x)\} \end{aligned}$$

The one-to-one correspondence between $h(t)$ and pdfs is shown below:

$$\begin{array}{ll} h(t) = t & g(x) \sim \text{Exp}\{\lambda\} \\ h(t) = \{t, t^2\} & g(x) \sim \mathcal{N}(\mu, \sigma^2) \\ h(t) = \{t, \log(t)\} & g(x) \sim \Gamma(k, \lambda) \\ h(t) = \{\log(t), \log(1-t)\} & g(x) \sim \text{Beta}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \end{array}$$

However, if $g_1(x, \boldsymbol{\theta}), g_0(x, \boldsymbol{\theta})$ come from different exponential families with the same support then:

$$\begin{aligned} \frac{g_1(x)}{g_0(x)} &= \frac{d_1(\boldsymbol{\theta}_1)S_1(x)}{d_0(\boldsymbol{\theta}_0)S_0(x)} \cdot \exp \left\{ \sum_{j=1}^k [C_{1j}(\boldsymbol{\theta}_1) \cdot T_{1j}(x) - C_{0j}(\boldsymbol{\theta}_0) \cdot T_{0j}(x)] \right\} \\ &= \exp \left\{ \log \frac{d_1(\boldsymbol{\theta}_1)}{d_0(\boldsymbol{\theta}_0)} + \log \frac{S_1(x)}{S_0(x)} + \sum_{j=1}^k [C_{1j}(\boldsymbol{\theta}_1) \cdot T_{1j}(x) - C_{0j}(\boldsymbol{\theta}_0) \cdot T_{0j}(x)] \right\} \end{aligned} \quad (3.7)$$

Denote

$$\alpha = \log \left\{ \frac{d(\boldsymbol{\theta}_1)}{d(\boldsymbol{\theta}_0)} \right\}, \quad \phi(x, \boldsymbol{\beta}) = \log \frac{S_1(x)}{S_0(x)} + \sum_{j=1}^k [C_{1j}(\boldsymbol{\theta}_1) \cdot T_{1j}(x) - C_{0j}(\boldsymbol{\theta}_0) \cdot T_{0j}(x)]$$

Then

$$\frac{g_1(x)}{g_0(x)} = \exp \{ \alpha + \phi(x, \boldsymbol{\beta}) \} \quad (3.8)$$

Model (3.6) is a special case of model (3.8), which was proposed by Zhang (2000).

3.2 Distribution Estimation for Density Ratio Models

The approach estimating parameters for density ratio models can be demonstrated best by the two-sample univariate case (Qin and Zhang, 1997). The same strategy can be applied to the multiple-sample case (Lu, 2007) or the multivariate case (Voulgaraki et al., 2012). Consider the two independent random samples \mathbf{x}_1 and \mathbf{x}_0 in Model (3.6) with $h(x) = x$ and denote \mathbf{x}_0 as the reference sample :

$$\frac{g_1(x)}{g(x)} = e^{\alpha + \beta x}, \quad g_0(x) = g(x). \quad (3.9)$$

Let \mathbf{t} be the concatenated or fused data from both \mathbf{x}_0 and \mathbf{x}_1 :

$$\mathbf{t} = (\mathbf{x}'_1, \mathbf{x}'_0) = (t_1, \dots, t_n), \quad n = n_1 + n_0.$$

Denote $p_i = dG(t_i)$ as the mass at t_i , $i = 1, \dots, n$. The empirical likelihood function becomes (Owen, 2001):

$$\mathcal{L}(\alpha, \beta, G) = \prod_{i=1}^{n_0} p_i \cdot \prod_{j=1}^{n_1} p_j \cdot e^{\alpha + \beta \cdot x_{1j}} = \prod_{i=1}^n p_i \cdot \prod_{j=1}^{n_1} e^{\alpha + \beta \cdot x_{1j}}. \quad (3.10)$$

It is subject to two constraints:

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i [w_1(t_i) - 1] = 0, \quad \text{where } w_1(t) = e^{\alpha + \beta \cdot t}.$$

The estimates $\hat{\alpha}$ and $\hat{\beta}$ can be computed by maximizing the likelihood function (3.10), and \hat{p}_i and $\hat{G}(t)$ are obtained as functions of $\hat{\alpha}$ and $\hat{\beta}$:

$$\hat{p}_i = \frac{1}{n_0} \cdot \frac{1}{1 + \rho \exp\{\hat{\alpha} + \hat{\beta} t\}},$$

$$\begin{aligned} \hat{G}(t) &= \sum_{i=1}^n \hat{p}_i \mathbf{I}\{t_i \leq t\} \\ &= \frac{1}{n_0} \sum_{i=1}^n \frac{\mathbf{I}\{t_i \leq t\}}{1 + \rho \exp\{\hat{\alpha} + \hat{\beta} \cdot t\}}. \end{aligned}$$

where $\rho = n_1/n_0$ and $\mathbf{I}\{t_i \leq t\}$ is an indicator function which equals one for $t_i \leq t$ and zero otherwise. The estimate $\hat{G}_1(t)$ is estimated by accumulating $\hat{p}_i \cdot e^{\hat{\alpha} + \hat{\beta} \cdot t_i}$. More generally, for multiple samples $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m, \mathbf{x}_0$:

$$\mathbf{x}_i = \{x_{i1}, \dots, x_{in_i}\} \sim g_i(x), \quad i = 1, \dots, m,$$

we still consider \mathbf{x}_0 as the reference sample: $g_0(x) = g(x)$, $G_0(x) = G(x)$. Assuming a multiple sample density ratio model:

$$\frac{g_i(x)}{g(x)} = e^{\alpha_i + \beta_i h(x)}, \quad (3.11)$$

then

$$\hat{p}_i = \frac{1}{n_0} \cdot \frac{1}{1 + \rho_1 \exp(\hat{\alpha}_1 + \hat{\beta}_1 h(t_i)) + \dots + \rho_m \exp(\hat{\alpha}_m + \hat{\beta}_m h(t_i))}. \quad (3.12)$$

Therefore,

$$\hat{G}(t) = \frac{1}{n_0} \cdot \sum_{i=1}^n \frac{\mathbf{I}(t_i \leq t)}{1 + \rho_1 \exp(\hat{\alpha}_1 + \hat{\beta}_1 h(t_i)) + \dots + \rho_m \exp(\hat{\alpha}_m + \hat{\beta}_m h(t_i))}, \quad (3.13)$$

where $\rho_i = n_i/n_0$ and $h(t)$ is the tilt function which is assumed known.

When the data are assumed continuous, the \hat{p}_i can be smoothed by a kernel to produce a more precise kernel density estimate than the traditional kernel density estimate since more data are used due to a sample concatenation (Fokianos, 2004, Cheng and Chu, 2004 and Qin and Zhang, 1997). A multivariate extension is

discussed in Voulgaraki et al. (2012), where it is also shown how to obtain optimal bandwidths.

3.3 Some asymptotic results of density ratio models

The asymptotic behavior of $\hat{G}(t)$ is described in Qin and Zhang (1997) and Zhang (2000) for the two-sample case. Moreover, Lu (2007) discussed the multiple-sample case using the same strategy. The efficiency of \hat{G} is elaborated in Gilbert (2000). Although the formula do not look straightforward, computing the parameters and $\hat{G}(t)$ is surprisingly fast for moderate sample sizes and $1 \leq m \leq 5$.

Qing and Zhang (1997) and Zhang (1998) derived some asymptotic results for the two-sample case. Let (α_0, β_0) be the true value of (α, β) . Then under the density ratio model (3.6), as $n \rightarrow \infty$,

$$n^{-1/2} \begin{pmatrix} \tilde{\alpha} - \alpha_0 \\ \tilde{\beta} - \beta_0 \end{pmatrix} \rightarrow N(0, S^{-1}VS^{-1}) = N(0, \Sigma), \quad (3.14)$$

$$\Sigma = \frac{1 + \rho}{\rho} \left[A^{-1} - \begin{pmatrix} 1 + \rho & 0 \\ 0 & 0 \end{pmatrix} \right],$$

where the matrix A can be obtained from:

$$A_k(t) = \int_{-\infty}^t \frac{\exp(\alpha_0 + \beta_0 y)}{1 + \rho \cdot \exp(\alpha_0 + \beta_0 y)} \cdot y^k \cdot dG(y),$$

$$A_0 = A_0(\infty), \quad A_1 = A_1(\infty), \quad A_2 = A_2(\infty), \quad A = \begin{pmatrix} A_0 & A_1 \\ A_1 & A_2 \end{pmatrix}.$$

They also showed that under the model and suitable regularity conditions, as $n \rightarrow \infty$, $\sqrt{n}\{\hat{G}(t) - \tilde{G}(t)\}$ converges weakly,

$$\sqrt{n}\{\hat{G}(t) - \tilde{G}(t)\} \xrightarrow{\mathcal{D}} W(t), \quad (3.15)$$

where $W(t)$ is a Gaussian process with mean 0 and covariance function specified by

$$E\{W(s)W(t)\} = \rho(1 + \rho)(A_0(s), A_1(s)) \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} - A^{-1} \begin{pmatrix} A_0(t) \\ A_1(t) \end{pmatrix} \right\},$$

and

$$\rho = \frac{n_1}{n_0}, \quad G(t) = \frac{1}{n_0} \sum_{i=1}^n \frac{\mathbf{I}\{t_i \leq t\}}{1 + \rho \cdot \exp\{\alpha_0 + \beta_0 \cdot t\}}.$$

Here $\tilde{G}(t)$ is the empirical cdf of $G(t)$,

$$\tilde{G}(t) = \frac{1}{n_0} \sum \mathbf{I}\{t_i \leq t\}.$$

In a later work, Qin and Zhang (1998) obtained the results:

$$\sqrt{n}(\hat{G}(t) - G(t)) \xrightarrow{\mathcal{D}} U(t), \quad (3.16)$$

where $U(t)$ is a Gaussian process with mean 0 and covariance function:

$$E\{U(s)U(t)\} = (1 + \rho)\{G(t) - G^2(t)\} - \left[\rho(1 + \rho)(A_0(s), A_1(s)) \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} - A^{-1} \begin{pmatrix} A_0(t) \\ A_1(t) \end{pmatrix} \right\} \right].$$

Lu (2007) obtained the multisample version of (3.14), (3.15) and (3.16) (as shown in Theorem 2.1, Theorem 3.8 and Theorem 3.9 (Lu, 2007)). The following quantities must first be defined to describe the asymptotic behavior of $\hat{G}(t)$,

$$A_j(t) = \int \frac{w_j(y)I(y \leq t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y),$$

$$B_j(t) = \int \frac{w_j(y)h(y)I(y \leq t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y),$$

$$\bar{A}(t) = (A_1(t), \dots, A_m(t))', \quad \bar{B}(t) = (B_1'(t), \dots, B_m'(t))'.$$

Here $\boldsymbol{\rho} = \text{diag}(\rho_1, \dots, \rho_m)_{m \times m}$ where I_p is the $p \times p$ identity matrix, \otimes denotes Kronecker product, and $\rho_i = n_i/n_0, i = 1, \dots, m$.

The main result of Lu (2007) is the following thorem: **Theorem** (Lu 2007, Thoerem 3.9): The process $\sqrt{n}(\hat{G}(t) - G(t))$ converges to a zero-mean Gaussian process in the space of real right continuous functions that have left limits with

covariance matrix given by

$$\begin{aligned}
& \text{Cov}\{\sqrt{n}(\hat{G}(t) - G(t)), \sqrt{n}(\hat{G}(s) - G(s))\} = \\
& \left(\sum_{k=0}^m \rho_k \right) \left(G(t \wedge s) - G(t)G(s) - \sum_{j=1}^m \rho_j A_j(t \wedge s) \right) \\
& + (\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes I_p)) S^{-1} \begin{pmatrix} \boldsymbol{\rho}\bar{A}(t) \\ (\boldsymbol{\rho} \otimes I_p)\bar{B}(t) \end{pmatrix}. \tag{3.17}
\end{aligned}$$

The immediate application of this theorem is the construction of pointwise symmetric confidence intervals for $G(t)$ for each given t . The following chapter is based on this result to obtain the confidence interval for the probability of passing the threshold t , $R(t) = 1 - G(t)$.

Chapter 4

Threshold Probabilities and Confidence Intervals

The purpose of this chapter is to review threshold probabilities and the widely available confidence intervals resulting from various methods, especially the interval from maximum likelihood method, the Wald interval and the Agresti-Coull interval. These intervals will be compared with the interval obtained from OSF DR models.

4.1 Threshold Probabilities

This dissertation deals with an out-of-sample fusion approach through confidence interval estimation for threshold probabilities. By threshold probability we mean a probability that a variable falls in a designated domain, which usually is a specific interval. As an example in engineering, the threshold probability can be interpreted as a failure probability of vehicles after the velocity or weight exceeds a certain value. In time-to-event data from clinical trials, it is the probability of surviving longer than a designated time. We say it follows a general risk definition if the designated domain is considered a risk domain where there is an undesirable, adverse or failure consequence, such as the risk to have certain cancer or disease

once the blood pressure or weight exceeds a certain threshold. In this study, we use both risk and threshold probability interchangeably. This chapter briefly reviews the commonly available approaches to construct a confidence interval for threshold probabilities.

In this dissertation by threshold probability we mean $1 - G(t)$, t being the threshold.

4.2 Confidence Intervals

4.2.1 CI based on maximum likelihood methods

In parametric scenarios, it is straightforward to construct confidence intervals for threshold probabilities using the maximum likelihood method and the delta method. Assume $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$. According to asymptotic theory, under regularity conditions:

$$\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \sim \mathcal{N}(0, I^{-1}(\boldsymbol{\theta})) \quad (4.1)$$

where $I(\boldsymbol{\theta})$ denotes Fisher information. Since the threshold probability is a function of a parameter $\boldsymbol{\theta}$, $R(t, \boldsymbol{\theta}) = 1 - G(t, \boldsymbol{\theta})$ can be obtained by maximum likelihood.

This interval, called CI_{MLE} , is often the most efficient CI estimate asymptotically. However it requires too many assumptions and sometime it is not easy to compute and often no explicit formi available.

4.2.2 Empirical Wald Interval

The empirical interval, called CI_{EP} , is one of earliest confidence intervals constructed (Laplace, 1812). It is always taught in standard statistical texts. It results from inverting a Wald test. Therefore it is always called the standard Wald interval. It is obtained as follows: In the nonparametric scenario, let $\mathbf{x} = \{x_1, \dots, x_n\}$, $i = 1, \dots, n$, be an *iid* sample from a distribution $G(t)$. Then the empirical distribution function is defined as:

$$\tilde{G}(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{x_i \leq t\}. \quad (4.2)$$

For a fixed t , the indicator $\mathbf{1}\{x_i \leq t\}$ is a Bernoulli random variable with parameter $p = G(t)$, hence $\tilde{G}(t)$ is a binomial random variable with mean $G(t)$ and variance $nG(t)(1 - G(t))$. The empirical distribution converges weakly to the true *cdf* for every t :

$$\sqrt{n}(\tilde{G}(t) - G(t)) \xrightarrow{\mathcal{D}} \mathcal{N}\{0, G(t)(1 - G(t))\} = \mathcal{N}\{0, \sigma_{EP}^2(t)\} \quad (4.3)$$

Thus, once the confidence level $1 - \alpha$ is specified, a confidence interval for $G(t)$ can be constructed :

$$\{\tilde{G}(t) - z_{1-\alpha/2} \cdot \sigma_{EP}(t)/\sqrt{n}, \tilde{G}(t) + z_{1-\alpha/2} \cdot \sigma_{EP}(t)/\sqrt{n}\} \quad (4.4)$$

Confidence intervals of threshold probabilities $R(t) = 1 - G(t)$ follow from (4.4).

This method is a robust nonparametric method that does not assume any distribution for the sample. It is used widely especially in survival analysis. Without censoring, the well known Kaplan-Meier is just the empirical distribution.

However, this confidence interval, which is based on a Wald test, may not have the coverage claimed by the confidence level. This is true not only in small samples, but also even for some relatively large samples its coverage can be well below the specified nominal coverage, especially when the threshold t goes to either boundary, which means that $G(t)$ is close to 0 or 1 (Brown et al., 2001 and Brown et al., 2002). Some alternative methods have been proposed, such as the Clopper-Pearson, Wilson score, and the Agresti-Coull methods. We shall proceed to describe them briefly.

4.2.3 Clopper-Pearson Interval

The Clopper-Pearson interval is an early and widespread method for calculating binomial confidence intervals (Clopper and Pearson, 1934). This is often referred to as an exact method since it is based on the cumulative probabilities of the binomial distribution. However, the intervals are not exact because of the discreteness of the binomial distribution. The Clopper-Pearson interval is given by

$$\{\theta | P[Bin(n, \theta) \leq X] \geq \alpha/2\} \cap \{\theta | P[Bin(n, \theta) \geq X] \geq \alpha/2\}, \quad (4.5)$$

where X is the number of successes observed in the sample and $Bin(n; \theta)$ is a binomial random variable with n trials and probability of success θ . The relationship between the cumulative binomial distribution and the beta distribution leads to an alternate format of the Clopper-Pearson interval. Note that the endpoints are what would be expected from a Bayesian analysis which yields these posterior distribu-

tion limits under the assumption of a Beta(1,1) or, equivalently, uniform prior on (0,1).

$$B(\alpha/2; x, n - x + 1) < \theta < B(1 - \alpha/2; x + 1, n - x) \quad (4.6)$$

Where $B(\cdot, \cdot, \cdot)$ is the incomplete beta function. This interval is conservative since its coverage is greater than or equal to the nominal coverage for any population proportion.

4.2.4 Wilson score Interval

The Wilson interval is given by (Wilson, 1927):

$$\frac{\hat{p} + \frac{1}{2n} z_{1-\alpha/2}^2 \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{1}{n} z_{1-\alpha/2}^2} \quad (4.7)$$

It improves the normal approximation interval in terms of coverage for small probabilities. The Wilson interval is derived from

$$\left\{ \theta \mid z_{\alpha/2} \leq \frac{\hat{p} - \theta}{\sqrt{\theta(1-\theta)/n}} \leq z_{1-\alpha/2} \right\}$$

by solving for θ . Its actual coverage probability is closer to the nominal value than the Wald interval. The center of the Wilson interval is a weighted average of $\hat{p} = X/n$ and $1/2$:

$$\hat{p} \left(\frac{n}{n + z_{1-\alpha/2}^2} \right) + \frac{1}{2} \left(\frac{z_{1-\alpha/2}^2}{n + z_{1-\alpha/2}^2} \right),$$

with \hat{p} receiving greater weight as the sample size increases.

4.2.5 Agresti-Coull Interval

The Agresti-Coull Interval (Agresti and Coull, 1998), CI_{AC} , is called an interval resulting from an “adjusted Wald test” initially which improves the coverage of the CI_{EP} , especially when the sample size is small and the threshold probabilities are close to 0 or 1. However it is a very conservative confidence interval since it is usually wider than the alternative CIs. It can be considered as a derivative of Wilson score intervals regarding coverage and interval length. Like the CI resulting from inverting a Wald test, the Agresti-Coull interval still uses the normal approximation, however with a modified sample size and a modified point estimate to replace the true sample size and the usual point estimation $\hat{p} = X/n$. Given X successes in n trials, define

$$\tilde{n} = n + z_{1-\alpha/2}^2, \quad \tilde{p} = \frac{X + z_{1-\alpha/2}^2}{\tilde{n}}$$

Then, a confidence interval for p is given by

$$\tilde{p} \pm z_{1-\alpha/2} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}. \quad (4.8)$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ percentile of a standard normal distribution, as before. For example, for a 95% confidence interval, let $\alpha = 0.05$, so $z_{1-\alpha/2} = 1.96$ and $z_{1-\alpha/2}^2 = 3.84$. If we use 2 instead of 1.96 for $z_{1-\alpha/2}$, this is the so called “add 2 successes and 2 failures” interval.

Additional methods other than those described above can be found in Brown and DasGupta (2001).

In this dissertation, a new statistical method, an out-of-sample density ratio model (OSF DR model), is proposed to obtain shorter yet reliable confidence intervals (CI_{OSF}) for threshold probabilities by fusing artificial data with a reference real data sample. The resulting CI_{OSF} is evaluated extensively with the alternative CIs including CI_{MLE} , CI_{EP} and CI_{AC} in terms of their lengths and coverages to the true value of actual parameters.

This dissertation is organized as follows: The first chapter explains the notion of out-of-sample fusion and differentiates it from the traditional “within-sample” methods. The second chapter briefly reviews semiparametric density ratio models and focuses primary on some asymptotic results which are essential for this dissertation. Some available approaches of constructing confidence intervals, especially the Wald interval and the Agresti-Coull interval are reviewed in chapter 3.

Chapter 5 compares the out-of-sample fusion with several available methods in terms of confidence interval estimation. Theorem 5.3.1 states that confidence intervals of threshold probabilities resulting from density ratio models are always shorter than those obtained from the empirical approach. The ensuing simulation studies show that although the intervals resulting from out-of-sample fusion are con-

siderably shorter than those from the Agresti-Coull method, the nominal confidence levels specified are still maintained. This chapter is concluded with a guideline for the practical use of out-of-sample fusion.

In chapter 6, “repeated out-of-sample fusion” is introduced. It is considered a derivative and ramification of out-of-sample fusion. It repeatedly fuses a given sample with different sets of artificial samples, which leads to a collection of estimates after applying the density ratio model to each scenarios. Confidence intervals can be constructed from these estimates. Although the resulting confidence intervals are even shorter than those obtained from a typical out-of-sample fusion method, their coverages of the true value are often significantly below the nominal coverage. In order to achieve optimal confidence intervals, new confidence intervals are proposed by hybridizing these intervals with the intervals computed by the AC method. Although the notion of the hybridization has not been established theoretically, some empirical guidelines are summarized from numerical examples.

In chapter 7, a multivariate out-of-sample fusion density ratio model has been applied to case-control cancer data. The results not only confirm the perviously published results , but also illustrate a graphical approach to compare conditional pdfs from case-control data. The difference between the case and control can be better appreciated graphically.

Chapter 8 discusses the application of the density ratio model and out-of-sample fusion to microarray data related to colon cancer. It not only identifies the most significantly differentially expressed gene groups which are made of 2 or 3 member genes, but also shows the different association may exist among member

genes in the same gene group.

The appendix includes a manual for a new R package *DensityRatio* which implements the density ratio model including the out-of-sample fusion strategy. All results, tables and figures in this dissertation have been generated by this package, which is also useful for traditional density ratio model analysis.

Chapter 5

Density Ratio Models with Out-of-Sample Fusion

This chapter provides the implementation procedure for the density ratio model with out-of-sample fusion (OSF DR model) through CI estimation for threshold probabilities $1 - G(t)$. The resulting CIs are compared to those obtained from the Agresti-Coull method (AC) and the empirical approach (EP) in terms of their expected lengths and coverage probabilities. Finally a guideline is proposed for its practical implementation.

This chapter is organized as follows: Section 5.1 gives the procedure to implement the OSF DR model. Section 5.2 deals with comparison of OSF DR model with MLE approach in a parametric setting in terms of CI estimation. In Section 5.3, we state that and prove a theorem that shows that CI_{OSF} is shorter than the standard Wald CI, which is called empirical CI in this dissertation (CI_{EP}). One will notice that the theorem is also applicable under DR model with samples from different real sources instead of artificial data.

Section 5.4 compares CI_{OSF} to CI_{AC} in terms of their expected lengths and coverage probabilities. CI_{AC} is a threshold probability CI produced by Agresti-Coull method (1998). It improves the coverages of CI_{EP} when the threshold probabilities

are close to 0 or 1. In section 4.5 one observes that even in some misspecified scenarios, that is, where the density ratio model assumption does not hold, the OSF DR model still gives reasonable CIs for threshold probabilities. Section 5.6 summarizes the results and provides a guideline to employ the OSF DR model for CI estimation in practical uses.

5.1 OSF DR Model Procedure

This section presents a typical method utilizing an OSF DR model to produce estimates of threshold probabilities and their CIs (CI_{OSF}) with one artificial sample. The same strategy can also be applied in multiple-artificial-sample scenarios.

Let \mathbf{x}_0 be an *iid* sample from a given population:

$$\mathbf{x}_0 = \{x_{01}, x_{02}, \dots, x_{0n}\}.$$

1. \mathbf{x}_1 is a sample from a normal distribution with the same mean and variance of \mathbf{x}_0 , $n_1 = n_0$.
2. Assume a density ratio model:

$$\frac{g_1(x)}{g(x)} = \exp\{\alpha + \beta \cdot x + \gamma \cdot x^2\},$$

where $g(x), g_1(x)$ are the pdfs of $\mathbf{x}_0, \mathbf{x}_1$, respectively. For convenience, set

$$h(x) = (x, x^2).$$

3. Estimate the threshold probability at t : $R(t) = 1 - G(t)$ according to the estimation method presented in Chapter 2:

$$\hat{R}(t) = 1 - \hat{G}(t) = 1 - \frac{1}{n} \sum_{i=1}^n \frac{\mathbf{I}\{t_i \leq t\}}{1 + \rho \cdot \exp\{\hat{\alpha} + \hat{\beta} \cdot t + \hat{\gamma} \cdot t^2\}},$$

where

t_i is a component from: $\mathbf{t} = \{\mathbf{x}'_1, \mathbf{x}'_0\}', i = 1, \dots, n,$

$\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ are estimates of α, β, γ , respectively,

$\rho = \frac{n_1}{n_0} = 1$, since $n_1 = n_0 = n$, and

$\mathbf{I}\{t_i \leq t\}$ is an indicator function which equals one for $t_i \leq t$ and zero otherwise.

4. According to (3.16), the confidence interval for $R(t)$ at the $100(1 - \alpha)\%$ level is:

$$\{\hat{R}(t) - z_{1-\alpha/2} \cdot \hat{\sigma}(t)/\sqrt{n}, \hat{R}(t) + z_{1-\alpha/2} \cdot \hat{\sigma}(t)/\sqrt{n}\}, \quad (5.1)$$

$\hat{\sigma}(t)$ is the estimate for $\sigma(t)$ from formula (3.16) when $t = s$:

$$\sqrt{n}(\hat{G}(t) - G(t)) \longrightarrow W(t) = N(0, \sigma^2(t)), \quad (5.2)$$

$$\begin{aligned}
\sigma^2(t) &= E\{U(t)^2\} \\
&= (1 + \rho)\{G(t) - G^2(t)\} \\
&\quad - \left[\rho(1 + \rho)(A_0(t), A_1(t)) \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} - A^{-1} \begin{pmatrix} A_0(t) \\ A_1(t) \end{pmatrix} \right\} \right].
\end{aligned}$$

5.2 Comparing CI_{OSF} to CI_{MLE}

In standard statistical texts, given an *iid* sample \mathbf{x}_0 from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with a sample size n , the procedure to obtain the confidence interval for a threshold probability $R(t)$ with the nominal confidence level $100(1 - \alpha)\%$ is stated as below, where

$$R(t) = P\{\mathbf{X} \geq t\} = 1 - \Phi\left(\frac{t - \mu}{\sigma}\right) \equiv \Psi(\boldsymbol{\theta}), \quad \boldsymbol{\theta} = (\mu, \sigma)'$$

To construct a $100(1 - \alpha)\%$ CI for $R(t)$, we resort to asymptotic results:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim \mathcal{N}(0, I^{-1}(\boldsymbol{\theta})),$$

$$\sqrt{n}(\hat{R}(t) - R(t)) \sim \mathcal{N}(0, [\Psi'(\boldsymbol{\theta})] \cdot I^{-1}(\boldsymbol{\theta}) \cdot [\Psi'(\boldsymbol{\theta})]^T) \equiv \mathcal{N}(0, \sigma_{R(t)}^2),$$

$$\Psi'(\boldsymbol{\theta}) = \left\{ \frac{\partial \Psi}{\partial \mu}, \frac{\partial \Psi}{\partial \sigma} \right\} = \left\{ \phi\left(\frac{t - \mu}{\sigma}\right), \frac{t - \mu}{\sigma^2} \phi\left(\frac{t - \mu}{\sigma}\right) \right\}.$$

The CI, called a CI_{MLE} in this dissertation, is obtained:

$$\{R(t) - z_{1-\alpha/2} \cdot \sigma_{R(t)}/\sqrt{n}, R(t) + z_{1-\alpha/2} \cdot \sigma_{R(t)}/\sqrt{n}\},$$

where $\Phi(t)$ and $\phi(t)$ are the standard normal cdf and pdf at t , respectively and $I(\boldsymbol{\theta})$ is the Fisher information. Once the asymptotic variance is acquired, the CI_{MLE} can be constructed easily by plugging in their estimates.

However, generally obtaining MLEs is not always straightforward. Simple explicit forms of MLEs are not often available. Numerical methods are frequently needed. Usually iterative steps are involved and sometimes slow convergence may occur and extra caution needs to be exercised to differentiate a global maximizer from local maximizers. An example is a CI_{MLE} from a sample drawn from a beta distribution:

$$g(x, \boldsymbol{\theta}) = \frac{\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)} \cdot x^{\theta_1-1}(1-x)^{\theta_2-1}, \quad 0 \leq x \leq 1, \quad \boldsymbol{\theta} = (\theta_1, \theta_2).$$

In general, CI_{OSF} can be computed regardless of the complexity of the reference population pdf. However, the CI_{OSF} is not as efficient as the CI_{MLE} , obviously, since it does not take into account all the information available in the reference distribution since the distribution of the population is not utilized during the calculation for CI_{OSF} .

Fokianos and Qin used a density ratio method to draw inference for a beta distributed sample by combining it with one uniformly distributed artificial sample

(Fokianos and Qin, 2008). The variance of the semiparametric estimate of θ is larger compared to its MLE counterpart. They believe that the decreased efficiency is due to the introduction of extra uncertainty from the uniformly distributed artificial sample.

In general the OSF DR model provides a convenient alternative to obtain statistical inference even in some parametric settings, a slight loss of efficiency is not a paramount issue.

5.3 Comparing CI_{OSF} to CI_{EP}, CI_{AC}

A typical DR model requires more than one sample. An OSF DR model makes it possible to apply a density ratio model to only the reference sample itself. Other samples required by the model can be generated by the computer. The resulting CI_{OSF} has a shorter length than that of CI_{EP} which utilizes the reference sample only. Theorem 5.3.1 proves this assertion.

Theorem 5.3.1 *Under regularity conditions, a confidence interval for a threshold probability obtained from an OSF DR model is always shorter than that computed by either an empirical method or AC method:*

$$CI_{OSF} \leq CI_{EP}; \quad CI_{OSF} \leq CI_{AC} \quad \text{in terms of their lengths.} \quad (5.3)$$

Proof

One only needs to compare their standard deviations in order to compare lengths of two confidence intervals. The theorem can be proved by following Theorems 3.8 and 3.9 in Lu (2007).

According to Theorem 3.9 in Lu (2007):

$$\sqrt{n}\{\hat{G}(t) - G(t)\} \sim \mathcal{N}(0, \sigma^2(t)), \quad (5.4)$$

where $\sigma^2(t)$ is given by

$$\sum_{k=0}^m \rho_k G(t)(1-G(t)) - \left\{ \sum_{k=0}^m \sum_{j=1}^m \rho_j A_j(t) - (\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho} \otimes I_p))S^{-1} \begin{pmatrix} \boldsymbol{\rho}\bar{A}(t) \\ (\boldsymbol{\rho} \otimes I_p)\bar{B}(t) \end{pmatrix} \right\}.$$

Here n is the combined number of observations from $m + 1$ samples, $\rho_0 = 1$, $n = n_1 + \cdots + n_m + n_0$. The symbols \otimes , $\boldsymbol{\rho}$, $A_j(t)$, $B_j(t)$, $\bar{A}(t)$ and \bar{B} have already been defined in Chapter 2.

From Theorem 3.8 in Lu (2007):

$$\sqrt{n}\{\hat{G}(t) - \tilde{G}(t)\} \sim \mathcal{N}(0, \sigma_E^2(t)), \quad (5.5)$$

where $\sigma_E^2(t)$ is given by

$$\sum_{k=0}^m \sum_{j=1}^m \rho_j A_j(t) - (\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho} \otimes I_p))S^{-1} \begin{pmatrix} \boldsymbol{\rho}\bar{A}(t) \\ (\boldsymbol{\rho} \otimes I_p)\bar{B}(t) \end{pmatrix}.$$

Also for the nonparametric estimate $\tilde{G}(t)$:

$$\sqrt{n_0}\{\tilde{G}(t) - G(t)\} \sim \mathcal{N}(0, \sigma_{EP}^2) = \mathcal{N}(0, G(t)(1 - G(t))), \quad (5.6)$$

where n_0 is the number of observations from the reference sample only. Consider $\rho_i = 1, i = 0, \dots, m$. Then from (5.4) and (5.6),

$$\begin{aligned} \sigma^2(t) &= \sum_{k=0}^m \rho_k G(t)(1 - G(t)) - \sigma_E^2(t) \\ &= (m + 1)\sigma_{EP}^2(t) - \sigma_E^2(t). \end{aligned}$$

Denote standard errors (*SEs*) of σ^2, σ_E^2 and σ_{EP}^2 as:

$$SE(t) = \frac{\sigma(t)}{\sqrt{(m+1) \cdot n_0}}, \quad SE_E(t) = \frac{\sigma_E(t)}{\sqrt{(m+1) \cdot n_0}}, \quad SE_{EP}(t) = \frac{\sigma_{EP}(t)}{\sqrt{n_0}}.$$

It follows that

$$SE^2(t) = SE_{EP}^2(t) - SE_E^2(t) \leq SE_{EP}^2(t). \quad (5.7)$$

Since SE s are nonnegative, then,

$$SE(t) \leq SE_{EP}(t).$$

It follows:

$$CI_{OSF} \leq CI_{EP} \quad \text{in terms of their length;}$$

$CI_{OSF} \leq CI_{AC}$ results naturally since CI_{EP} and CI_{AC} are equivalent asymptotically.

For example, $m = 1$, i.e. the two-sample case, Theorems 3.8 and 3.9 in Lu (2007) for multiple samples reduce to the results in Qin and Zhang's papers (Qin and Zhang, 1997 and Zhang, 2000). We have

$$\sqrt{n}\{\hat{G}(t) - \tilde{G}(t)\} \sim \mathcal{N}(0, \sigma_E^2(t)), \quad (5.8)$$

$$\sqrt{n}(\hat{G}(t) - G(t)) \longrightarrow W(t) = N(0, \sigma^2(t)), \quad (5.9)$$

$$\sigma_E^2(t) = \rho(1 + \rho)(A_0(t), A_1(t)) \left\{ \left(\begin{array}{c} 1 \\ 0 \end{array} \right) - A^{-1} \left(\begin{array}{c} A_0(t) \\ A_1(t) \end{array} \right) \right\},$$

$$\begin{aligned}
\sigma^2(t) &= (1 + \rho)\{G(t) - G^2(t)\} \\
&\quad - \left[\rho(1 + \rho)(A_0(t), A_1(t)) \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} - A^{-1} \begin{pmatrix} A_0(t) \\ A_1(t) \end{pmatrix} \right\} \right] \\
&= (1 + \rho)\{G(t) - G^2(t)\} - \sigma_E^2(t).
\end{aligned}$$

Compare $\sigma^2(t)$ and $\sigma_E^2(t)$:

$$\begin{aligned}
\sigma^2(t) &= (1 + \rho)\sigma_{EP}^2(t) - \sigma_E^2(t) \\
&= \frac{n_0 + n_1}{n_0}\sigma_{EP}^2(t) - \sigma_E^2(t) \\
&\leq \frac{n}{n_0}\sigma_{EP}^2(t).
\end{aligned}$$

Consider

$$\begin{aligned}
\sigma^2(t) &= n \cdot SE^2(t) = (n_0 + n_1) \cdot SE^2(t), \\
\sigma_{EP}^2(t) &= n_0 \cdot SE_{EP}^2(t),
\end{aligned}$$

it is easy to obtain:

$$SE(t) \leq SE_{EP}(t).$$

This theorem shows OSF DR models provide a shorter CI than the empirical method if the model assumption is justified. One can easily observe that the theorem applies to a typical density ratio model with real multisource samples instead of

artificial samples. However, the theorem does not address the coverage of these CIs. The common practice to evaluate CIs is to ensure they provide the desirable coverage for the true parameter values. The next section addresses their coverage and provides numerical evidence to support this theorem.

5.4 Comparing CI_{OSF} to CI_{AC}

We have shown in Section 4.3 theoretically that CI_{OSF} is shorter than either CI_{AC} or CI_{EP} . In this section, Theorem 5.3.1 is confirmed numerically CI coverage of the actual value of the threshold probability. Moreover, a comparison of CI_{OSF} with the CI_{AC} is made in terms of both their expected length and coverage. The effect of sample size and model misspecification on CI_{OSF} length and coverage is investigated subsequently. Based on these results, a guideline to use DR OSF models for threshold probability CI construction is formulated in the following section.

The reference sample used in this section is from a population with a standard normal distribution: $\mathcal{N}(0, 1)$, $n_0 = 100$, fused with some of $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$. These samples, having the same size 100 also, are from the populations:

$$\begin{aligned} \mathbf{x}_1 &\sim \mathcal{N}(1, 1), & \mathbf{x}_2 &\sim \mathcal{N}(1, 2), \\ \mathbf{x}_3 &\sim \mathcal{N}(0, 1), & \mathbf{x}_4 &\sim \mathcal{N}(0, 1). \end{aligned}$$

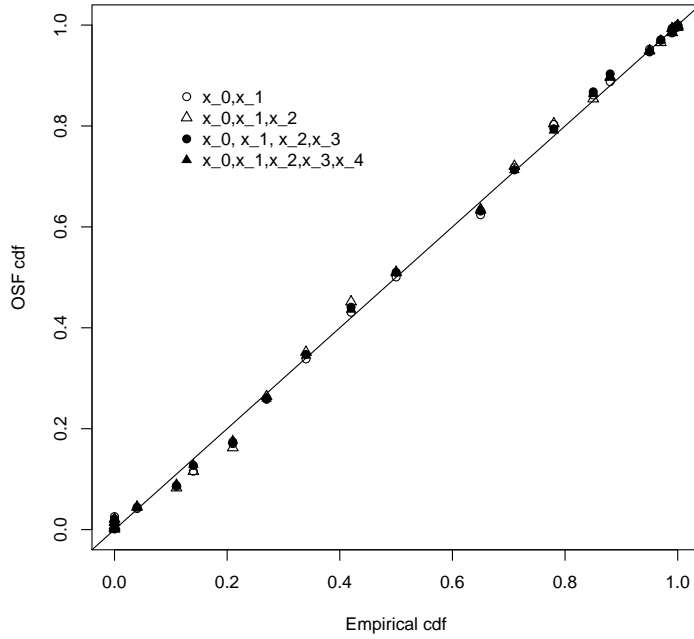


Figure 5.1: Goodness-of-fit test: $\hat{G}(t)$ versus $\tilde{G}(t)$

5.4.1 Diagnostic plots for the goodness-of-fit test

Goodness-of-fit tests are needed to justify the density ratio model applicability. If $\hat{G}(t)$ is the reference cdf acquired by OSF and $\tilde{G}(t)$ is the corresponding empirical cdf, most goodness-of-fit methods measure the discrepancy between $\hat{G}(t)$ and $\tilde{G}(t)$ (Qin and Zhang, 1997, Gilbert, 2004 and Voulgaraki et al., 2012). A simple graphical method is to plot $\hat{G}(t)$ versus $\tilde{G}(t)$.

Figure 5.1 suggests that regardless of the number of artificial samples used to obtain $\hat{G}(t)$, $\tilde{G}(t)$ and $\tilde{G}(t)$ are close enough to justify the density ratio model. It is obvious due to the normality of both the reference sample and the artificial samples which conform to the density ratio model with tilt function: (x, x^2) .

5.4.2 Lengths of CI_{OSF} , CI_{AC} and CI_{EP}

Figure 5.2 shows that CI_{OSF} , CI_{AC} and CI_{EP} are functions of the threshold T . The CI_{OSF} is obtained by concatenating or fusing the reference sample \mathbf{x}_0 with \mathbf{x}_1 and \mathbf{x}_2 . In order to better appreciate the difference of their CI lengths, the correspondent standard deviations of Figure 5.2 are plotted in Figure 5.3, which shows all SEs reach their maximum around 0 and diminish gradually to either end. Without considering either end of the support, the CI_{OSF} is always shorter than either CI_{AC} or CI_{EP} , which confirms Theorem 5.3.1. The striking chaotic behaviors in either end of the support deserve special attention. The CI_{EP} turns out to have 0 length where there are no sample observations available. CI_{AC} , considered an improved CI_{EP} , agreeing well with CI_{EP} at the middle of its support, however, is exceptionally wide at either end of the support. The whole purpose of introducing CI_{AC} is to let the confidence interval resulting from the Wald test have an overall coverage which agrees with the nominal confidence level specified over the whole support (Agresti and Coull, 1998). In contrast, the length CI_{OSF} diminishes more reasonably and naturally as the threshold goes to the boundaries. Under the scenario that CI_{EP} is nonzero, the relation below is always observed:

$$CI_{OSF} < CI_{EP} < CI_{AC}.$$

Figure 5.4 shows CI_{OSF} only. The four lines are obtained when different numbers of artificial samples are concatenated with the reference sample. The gap in CI_{OSF} becomes wider in the middle when more artificial samples are used, however

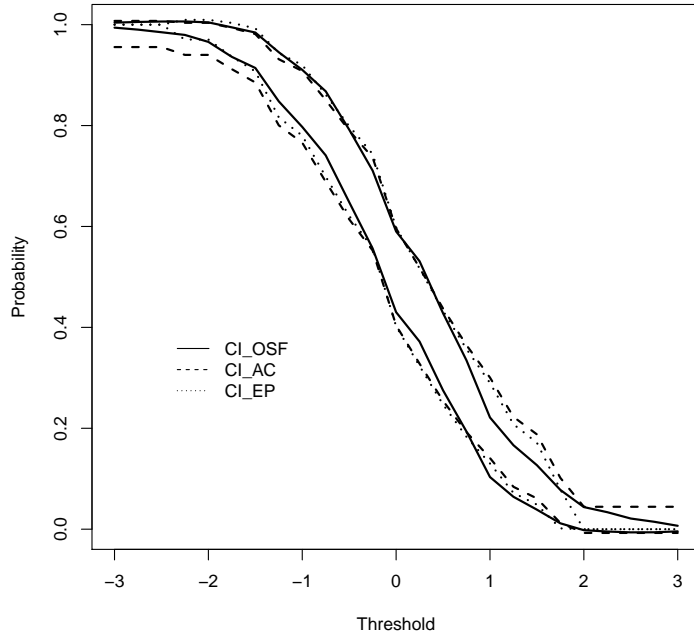


Figure 5.2: Comparison of nominal 95% CI_{OSF} , CI_{AC} and CI_{EP} .

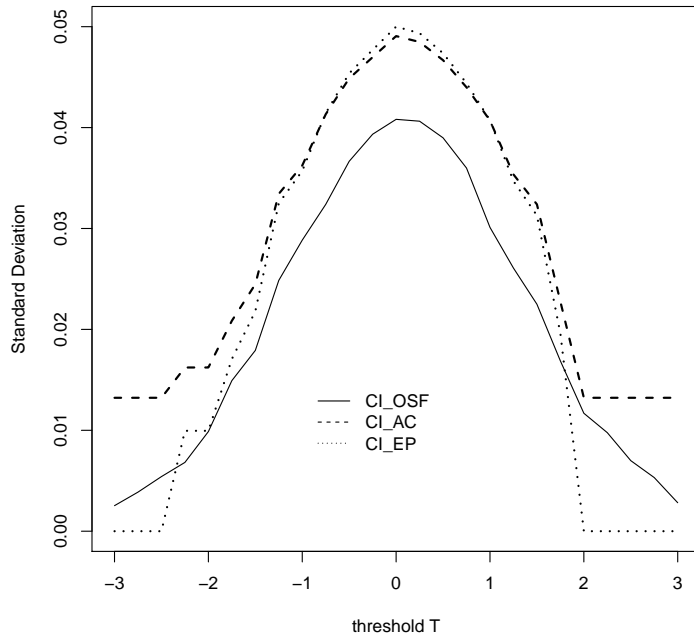


Figure 5.3: Comparison of the standard deviations of nominal 95% CI_{OSF} , CI_{AC} and CI_{EP} (corresponding to Figure 5.2).

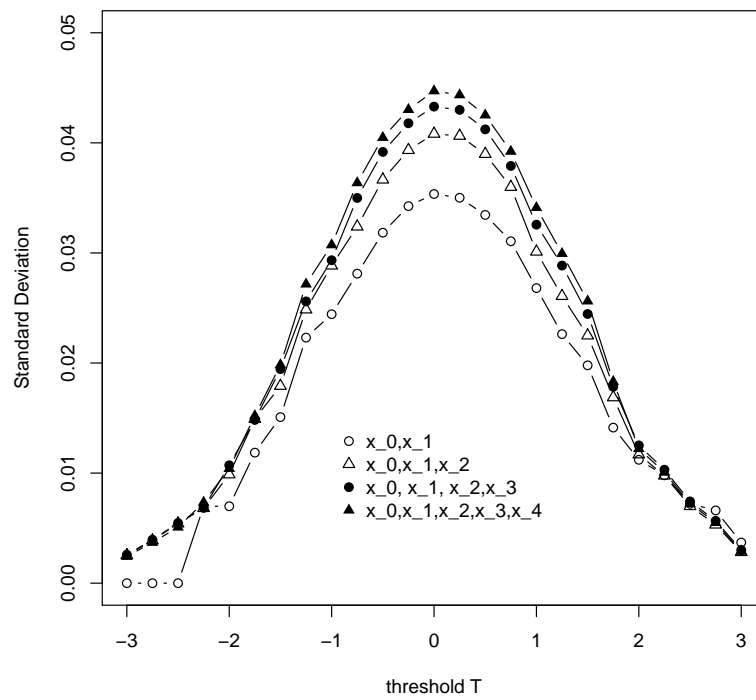


Figure 5.4: Standard deviation plot for 95% CI_{OSFS} which are obtained by using different number of artificial samples.

in either end of the threshold T support ($R(t) = 1 - P(X > t) = 0.05$), its length does not change much with an increase of the number of artificial samples relative to the case when only one artificial sample is fused. So including more artificial sample does not have an obvious advantage in terms of its CI length for very large or small threshold probabilities, in which case it seems to be reasonable to fuse only 2 or 3 artificial samples with the reference sample for CI_{OSF} estimation when we deal with small threshold probabilities.

5.4.3 Coverage Probabilities of CI_{OSF} , CI_{AC} and CI_{EP}

A short length does not secure a superiority of a CI over other intervals as mentioned before. It is required to take the coverage of the true value of the parameter into consideration. The coverage probability of a confidence interval is the probability of covering the actual value of the parameter of interest. Here the parameter is the threshold probability $1 - G(t)$.

Suppose the population of interest is $\mathcal{N}(0, 1)$. Given the threshold $T = 1.645$, the theoretical threshold probability is $R(T) = 1 - P(X > T) = 0.05$. Figure 5.5 shows the 100 intervals generated by 100 different reference samples from the same population with OSF, AC and EP methods separately. i.e. each interval is obtained by a different reference sample, however all reference samples are from the same population specified. In this case, it is $\mathcal{N}(0, 1)$. The sample size is 100 for either the artificial sample or the reference sample. The CI_{OSF} is calculated using \mathbf{x}_1 and \mathbf{x}_2 along with the reference sample \mathbf{x}_0 . At the left panel of Figure 5.5, 100 vertical lines

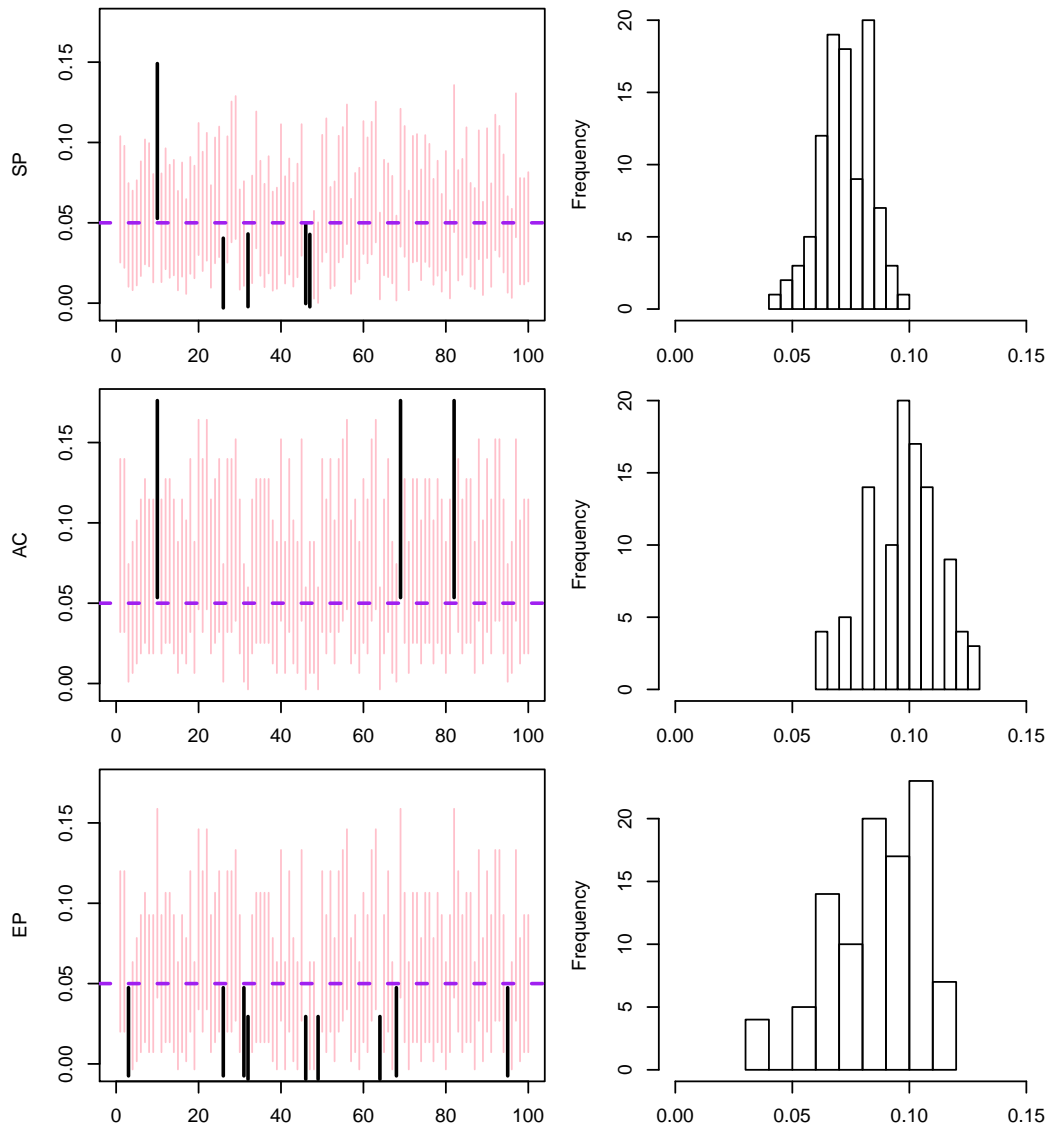


Figure 5.5: Comparison of coverage and length distribution of CI_{OSF} , CI_{EP} and CI_{AC} , Note: SP is OSF here.

Table 5.1: Coverages and lengths of 100 CIs obtained from OSF, AC and EP methods

	OSF	AC	EP
Coverage %	95	97	91
Mean length	0.0715	0.0968	0.0858
length/AC length %	74	100	89

are 100 CIs while the horizontal dash line indicates the true value of the threshold probability, 0.05. The bold intervals which do not intersect with the 0.05 dash line are those that fail to cover the true value. Thus the percentages of the non-bold intervals are the coverage probabilities of the confidence intervals. The right panel of Figure 5.5 shows the histogram of length distribution of 100 CIs produced from all three methods. The CI expected lengths and CI coverages for all three methods have been calculated and stated in Table 5.1. It shows the expected length of CI_{OSF} is much shorter than those of CI_{AC} and CI_{EP} , however, its coverage is more than that of CI_{EP} while lower than that of CI_{AC} which is a very conservative CI since it is always higher than the nominal level (95%) with an exceptional large length.

It is obvious that

$$\mathbf{Coverage:} \quad CI_{AC} > CI_{OSF} > CI_{EP}$$

$$\mathbf{Width:} \quad CI_{AC} > CI_{EP} > CI_{OSF}$$

However, the results shown in Table 5.1 results from one run only. In order to

Table 5.2: Mean coverage and widths resulting from 100 runs.

	OSF	AC	EP
Coverage %	94.78	96.94	91.51
Mean length	0.0595	0.0944	0.0827
length/AC length %	63	100	88

make the above assertion more convincible, 100 runs are performed. The results are summarized at Table 5.2 which shows the expected coverages and expected lengths obtained. Table 5.2 confirms that in the pervious single run scenario CI_{OSF} achieves significant reductions in terms of the CI length. It is about 30% shorter than either CI_{AC} or CI_{EP} . The superiority of CI_{OSF} over CI_{EP} is fairly convincing. This trend has been confirmed by varying the sample sizes for either artificial samples or the reference sample. Figure 5.6 shows the results of CI comparisons when varying the sample sizes: 50, 100, 500, 2000. Obviously all three methods are asymptotically consistent since in considerable big sample sizes, all of these CIs converge to the nominal coverage while their expected lengths converge to 0. CI_{AC} always achieves the nominal coverage: 95%. CI_{OSF} coverage probability is very close to nominal and both methods are superior to CI_{EP} in terms of coverage. The trend tends to more obvious as the sample size shrinks. The notable difference in small sample size makes the superiority of CI_{OSF} fairly obvious.

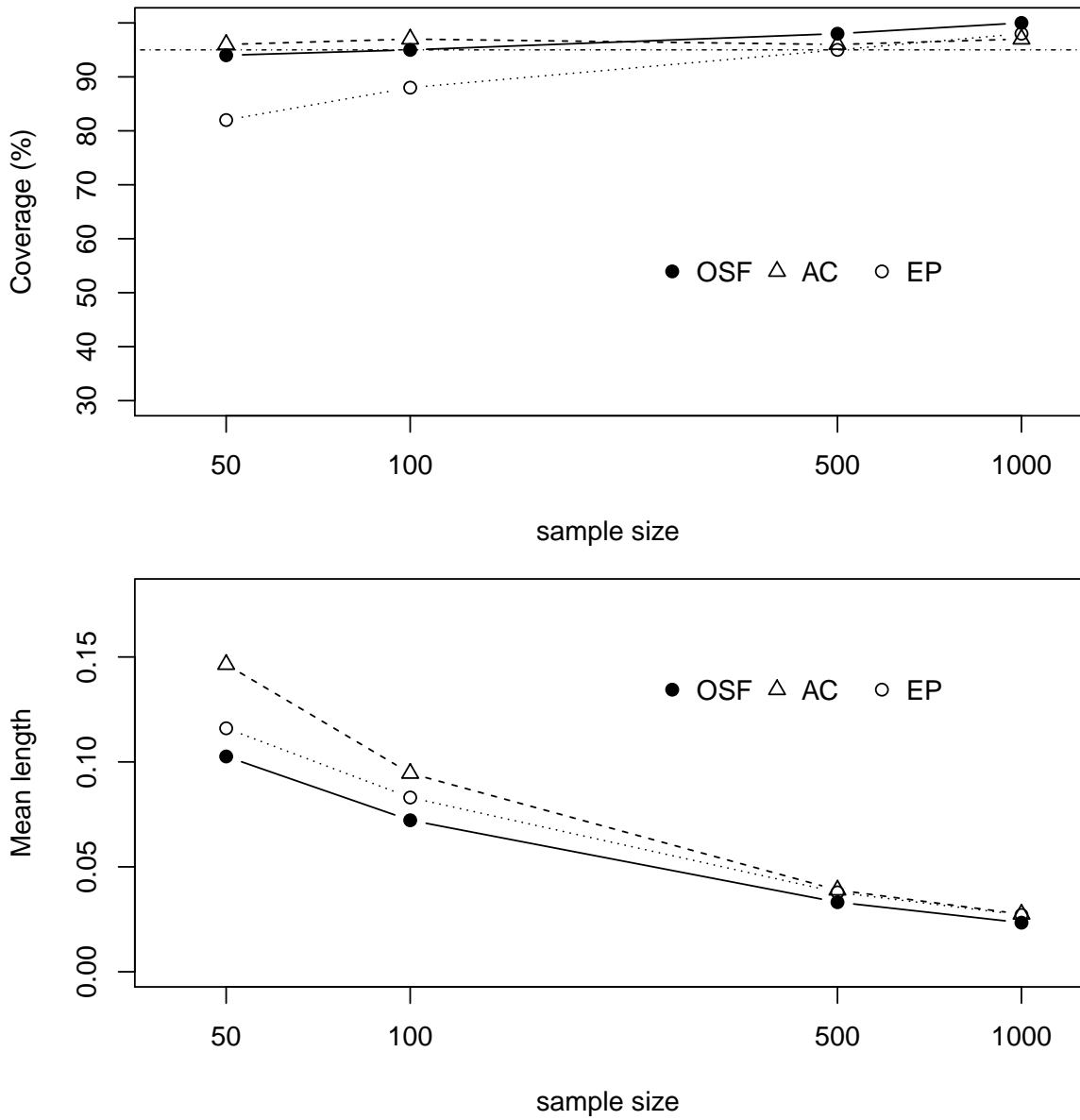


Figure 5.6: Mean coverage and lengths of CI_{OSF} , CI_{AC} and CI_{EP} when the sample sizes are 50, 100, 500, 2000; each point averages results from 100 runs.

5.5 Misspecified OSF DR models

We revisit the choice of artificial samples. An educated choice on purpose for the artificial samples leads to a near perfect goodness-of-fit test, which may incur criticisms of subjective sample manipulation. In a typical density ratio model, all of the samples come from real multiple sources, so the density model may not hold as wished. However, for a density ratio model with artificial samples (OSF DR model), there is flexibility in making some reasonable preference during the artificial sample selection. It is possible to let the artificial samples resemble the reference sample to some extent by carefully examining its statistical structures. The “optimal” artificial samples can be screened by goodness-of-fit tests. This leads to a better satisfaction of the assumption of the density ratio model. Still we are interested in the outcome in case of choosing artificial samples unwisely. That is, the artificial samples do not follow the density ratio model completely. We therefore study the robustness of the density ratio model with OSF for misspecified cases. We focus on small sample size since the above study suggests the CI_{OSF} has more convincing advantage over the other CIs in this scenario. Suppose the artificial samples chosen are:

$$\mathbf{x}_1 \sim \text{Expon}(1), \quad \mathbf{x}_2 \sim \mathcal{N}(1, 1), \quad \mathbf{x}_3 \sim t(5).$$

The reference sample is still from a standard normal $\mathcal{N}(0, 1)$ and the tilt function used in the density model is (x, x^2) . The difference from the reference is striking: \mathbf{x}_1 has support $(0, +\infty)$, while \mathbf{x}_0 has all of \mathbb{R} as its support. Suppose we totally

ignore model conformation and insist on using the OSF DR model. Figure 5.7 is obtained and we see it has the same format as Figure 5.5. The similarity to Figure 5.5 is striking. CI_{OSF} has the same coverage as CI_{AC} but its intervals are much shorter, and both methods are superior to CI_{EP} .

Another extreme example is shown in Figure 5.8. The artificial samples used are:

$$\mathbf{x}_1 \sim \text{Expon}(1), \quad \mathbf{x}_2 \sim \text{Binom}(5, 0.6), \quad \mathbf{x}_3 \sim \text{Poisson}(1), \quad \mathbf{x}_4 \sim t(5).$$

Note these two examples are representatives of many arbitrarily chosen artificial samples that we experimented with. We see that misspecified OSF DR models often give coverage identical to that from correctly specified models. However if any one of the artificial samples comes from a Cauchy distribution, the OSF DR method fails, see Voulgaraki et al. (2012).

5.6 Guideline for CI_{OSF}

In the previous sections, we suggest that the CI_{OSF} has an advantage in both large or small sample size scenarios, compared to either CI_{AC} or CI_{EP} . Since CI_{AC} has excellent coverage probability of the true value of $1 - G(t)$ over the whole support, it has been suggested as an alternative to the CI_{EP} , i.e. the Wald CI which is regarded as the “standard confidence interval” taught in most of standard statistical texts (Agresti and Coull, 1998 and Brown et al., 2001). However, its rather

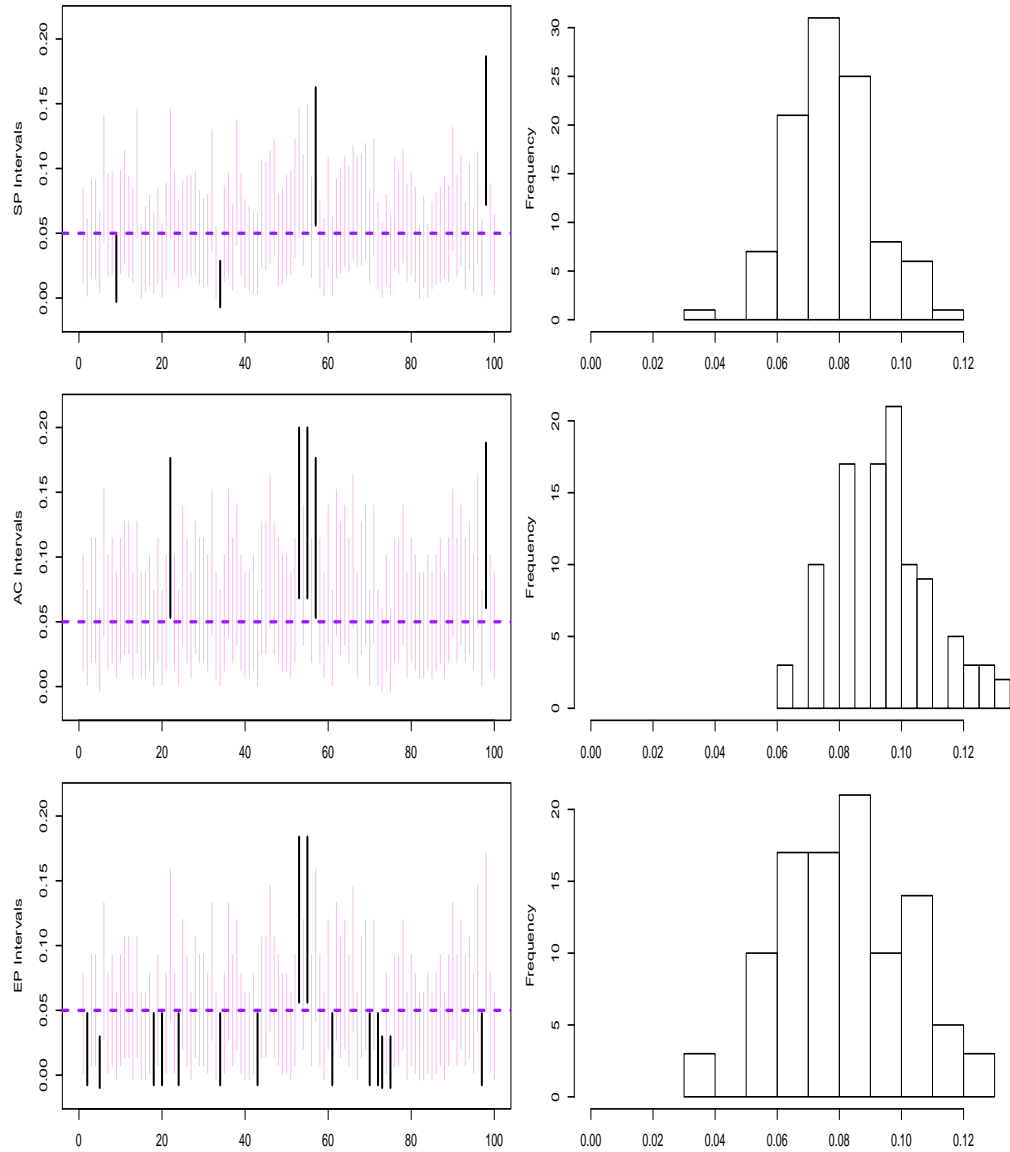


Figure 5.7: Mean coverage and widths of 100 CIs obtained from OSF, AC and EP methods, sample size:100, artificial samples involved are from exponential, normal and $t(5)$ distribution, the reference sample is from $\mathcal{N}(0, 1)$.

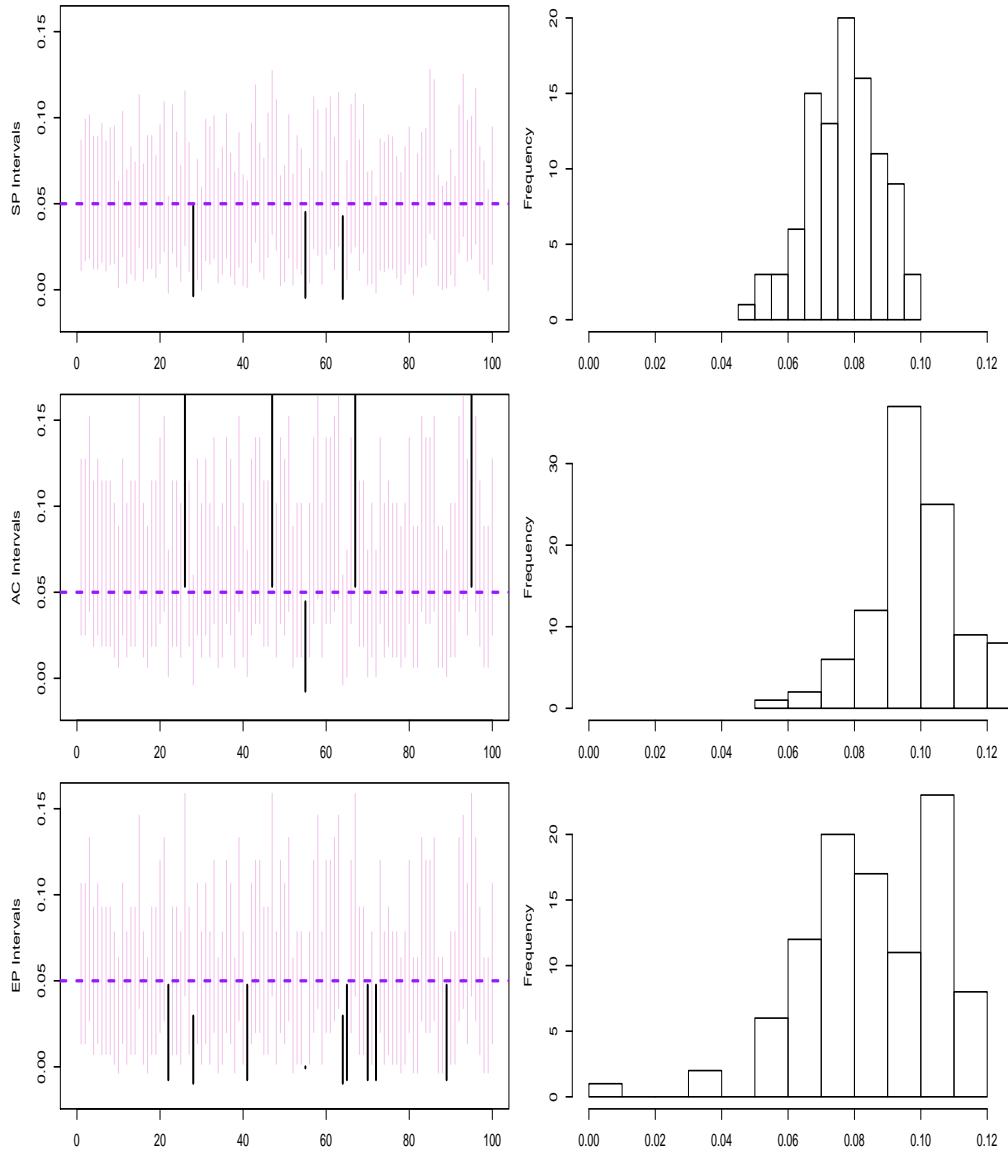


Figure 5.8: Mean coverage and widths of 100 CIs obtained from OSF, AC and EP methods, sample size:100, artificial samples involved are from exponential, binomial, poisson and $t(5)$ distribution, the reference sample is from $\mathcal{N}(0, 1)$

wide interval length is at odds with practitioners who prefer sharp intervals with a reasonable coverage, which may be somewhat lower than the nominal confidence level. Obtaining a CI which is up to 30% shorter than the best CI at the price of a slight coverage loss is often desirable. In relatively large sample size scenarios, the CI_{OSF} can often be more than 20% shorter than CI_{AC} without sacrificing any coverage probability. We can acquire even better CIs by taking advantage of CI_{OSF} and CI_{AC} using hybridization, which will be discussed in the next chapter.

Chapter 6

Density Ratio Models with Repeated Out-of-Sample Fusion

6.1 Introduction

In the previous chapter, the OSF DR model has been illustrated through the CI estimation for a threshold probability $R(t) = 1 - G(t)$ at t . The resulting CI_{OSF} appears to have considerable advantages over the alternative CIs. However, in some scenarios, its coverage is somewhat lower than that of CI_{AC} and sometimes it is even slightly lower than the nominal coverage specified in small samples. An enhanced OSF DR model is introduced in this chapter, which is called DR model with repeated out-of-sample fusion (ROSF DR model). The OSF DR model computes ONE estimate of $R(t)$ and then constructs its CI by acquiring its asymptotic variance. In contrast, ROSF DR model calculates many estimates of $R(t)$ by concatenating different sets of artificial samples repeatedly. Based on these estimates, a CI is constructed (CI_{ROSF}). Computation of the CI_{ROSF} involves no asymptotic theory.

We illustrate the ROSF DR model by CI estimation of threshold probabilities

also in a two-sample scenario. Suppose

$$\mathbf{x}_0 \sim g(x) \quad \text{with support } [a, b],$$

$$\mathbf{x}_1 \sim g_1(x) \equiv \text{Uniform}[a, b].$$

Then

$$\frac{g_1(x)}{g(x)} = [(b-a)g(x)]^{-1} = \exp\{ \text{constant} - \log(g(x)) \}.$$

This suggests that if the second derivative of $\log(g(x))$ does not vary much, then the density model (3.6) with tilt function as $h(x) = (x, x^2)$ approximately holds. This includes the normal, exponential distribution, and all the distributions which are not far from normal, such as gamma with a moderately large shape parameter and t distribution with $df > 5$. However, the choice of $h(x) = (x, x^2)$ may still produce reasonable results even for some misspecified cases which seem to violate the model assumption because the second derivative of the quantity $\log(g(x))$ is far from constant. This discussion provides a rationale to choose uniformly distributed artificial samples under the density model with $h(x) = (x, x^2)$. Usually we choose subsets of the reference support as the supports of artificial samples although it is not necessary. A procedure, based on this strategy, is presented in Section 5.2.

The expected length and coverage probability of $CI_{ROS F}$ have been investigated. A comparison of $CI_{ROS F}$ to CI_{AC} has been made. Aiming for a CI having a coverage no lower than the nominal coverage anywhere, a hybrid CI (CI_{HB}) is

proposed by combining CI_{ROSF} and CI_{AC} :

$$CI_{HB} = (w_1 L_{ROSF} + (1 - w_1) L_{AC}, w_2 U_{ROSF} + (1 - w_2) U_{AC}), \quad (6.1)$$

where

$\mathbf{W} = (w_1, w_2)$: the relative weight of CI_{ROSF} and CI_{AC} ,

L_{ROSF}, U_{ROSF} : lower bounds and upper bounds of CI_{ROSF} ,

L_{AC}, U_{AC} : lower bounds and upper bounds of CI_{AC} .

The HB confidence interval is intended to take advantage of both shortness of CI_{ROSF} and superior coverage of CI_{AC} . A detailed study on CI_{HB} has been performed when the reference sample is from various populations: such as normal, gamma, logistic, uniform and exponential.

Even though a theoretical treatment is not yet available a guideline to choose an optimal weight vector is provided based on numerical evaluation. A carefully formulated Hybrid CI is useful in terms of both expected length and coverage of the true value.

6.2 Procedure to obtain CI_{ROSF} and CI_{HB}

A possible procedure to obtain a CI_{ROSF} and CI_{HB} is described next. Suppose that \mathbf{x}_0 is a sample of size $n_0 = 50$ from an unknown population. It is regarded as the reference or original sample when the density ratio model is applied.

1. \mathbf{x}_0 is fused with any one of 49 uniform samples from $\text{Unif}(-2 - i, 3 + i)$, $i = 1, \dots, 49$, respectively, each of size 50, to yield 49 DR estimates \hat{p}_i ($\hat{p}_i = \hat{R}_i(t)$) in total: $\hat{p}_1, \dots, \hat{p}_{49}$.
2. Construct a CI based on percentiles: Order the \hat{p}_i according to their value and take the 2.5 and the 97.5 percentiles as the limits of a nominal 95% confidence interval for p .
3. Repeat 50 times steps 1 and 2 to produce 50 CIs and average them to yield a CI_{ROSF} .
4. Compute CI_{AC} of \mathbf{x}_0 according to the AC method described in Chapter 3.
5. Obtain CI_{HB} based on formula (6.1).

To illustrate the procedure, consider the case where $\mathbf{x}_0 \sim \mathcal{N}(0, 1)$, $\mathbf{w} = (0.25, 1.5)$. The choice of the weight vector \mathbf{W} brings about a shift of the HB confidence intervals relative to AC as can be seen from Table 6.1. The table reports five cases out of 100, from which the coverage probabilities for CI_{ROSF} , CI_{AC} and CI_{HB} were 81%, 95% and 95%, respectively. We do not advocate such an extreme choice as $w_2 = 1.5$. However, we use it here to point out the potential of the hybrid method.

Clearly, the choice of weight vectors determines how much the CI_{HB} moves inside CI_{AC} . In this example, for $w_2 = 0.25$ the left limit of CI_{HB} moves much less, thus deferring more to CI_{AC} . In general, small values of the weights bring about a more cautious movement inside CI_{AC} . We may decide to set one of the weights to zero, in which case the corresponding limit is the same as that from CI_{AC} .

Table 6.1: 95% confidence intervals for $R(T) = 0.05$ using three methods. The experiment was repeated 100 times of which five typical cases are listed here. Efficiency is the ratio of lengths relative to that of CI_{AC} , $\mathbf{x}_0 \sim \mathcal{N}(0, 1)$, and all samples are of size 50.

Method	$\bar{R}(T)$	\bar{L}	\bar{U}	$\bar{U} - \bar{L}$	Efficiency
ROSF	0.062	0.041	0.104	0.063	0.409
AC	0.091	0.014	0.168	0.154	1.000
HB		0.021	0.072	0.051	0.331
ROSF	0.102	0.074	0.153	0.079	0.441
AC	0.129	0.039	0.218	0.179	1.000
HB		0.048	0.121	0.073	0.408
ROSF	0.080	0.052	0.129	0.077	0.461
AC	0.110	0.026	0.194	0.167	1.000
HB		0.032	0.097	0.065	0.389
ROSF	0.048	0.021	0.123	0.103	0.741
AC	0.073	0.003	0.142	0.139	1.000
HB		0.008	0.114	0.106	0.763
ROSF	0.080	0.060	0.099	0.039	0.232
AC	0.110	0.026	0.194	0.168	1.000
HB		0.035	0.052	0.017	0.101

6.3 Coverage and length of CI_{HB}

Aiming to generalize the results in the previous section, an investigation has been made where the reference samples are from distributions frequently encountered. Although only gamma, logistic, uniform and exponential scenarios are pre-

sented here, lognormal and t have also been investigated. In this section we use graphical displays along with tables to illustrate the preceding hybrid HB method in the scenarios other than normal. The graphical displays for CI_{HB} show reduced confidence interval length and at the same time very similar coverage as that from AC. All sample sizes are 50 for the reference and the fusion samples. As noted, there are many ways to fuse the reference data, but we shall follow the procedure specified in the previous section.

Table 6.2: Coverage and average length from 100 runs for nominal 95% confidence intervals for $R(T) = 0.05$, In (*) the fusion samples were changed to $\mathbf{x}_1 \sim \text{Unif}(0, 3 + i)$ from $\mathbf{x}_1 \sim \text{Unif}(-2 - i, 3 + i)$, $1, \dots, 49$. All sample sizes are 50.

\mathbf{x}_0	T	w_1	w_2	Coverage			Ave. Width		
				SP	AC	HB	SP	AC	HB
N(0, 1)	1.645	0.40	0.40	0.75	0.97	0.97	0.046	0.142	0.103
Logistic(0,1)	2.944	0.40	0.40	0.58	0.97	0.96	0.039	0.143	0.101
U(0, 50)	47.500	0.40	0.40	0.28	0.98	0.96	0.026	0.143	0.096
$\Gamma(5, 0.5)$	18.307	0.40	0.40	0.31	0.96	0.95	0.035	0.138	0.097
$\Gamma(5, 0.5)^*$	18.307	0.40	0.40	0.52	0.96	0.95	0.043	0.146	0.105
Exp(1)	2.996	0.25	0.40	0.48	0.97	0.96	0.037	0.139	0.093

All the results reported below were obtained from 100 runs. The coverage and average length of confidence intervals from 100 runs are given in Table 6.2. The reference samples are different in all runs. In the first five cases reported in the table 6.2, with $\mathbf{w} = (0.40, 0.40)$, the CI_{HB} s are about 30% shorter than the corresponding CI_{AC} s. In the exponential case, the choice $\mathbf{w} = (0.25, 0.40)$ gives similar coverage and length reduction as in the other cases.

The scenarios of the reference coming from various distribution have also been shown in the following Figures 6.1, 6.2, 6.3, 6.4 and 6.5. All figures have the same format as Figure 5.5 .

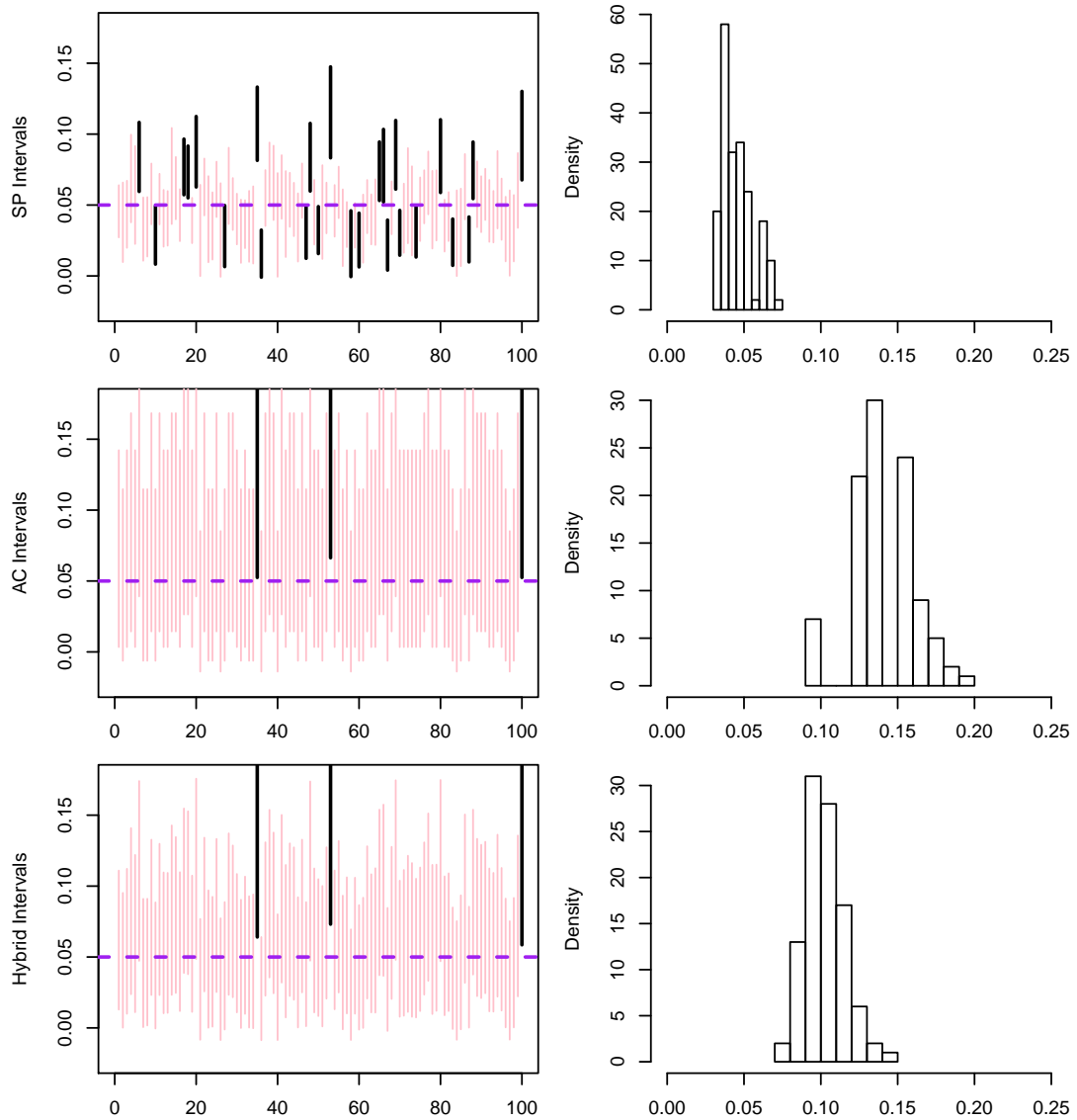


Figure 6.1: CI Coverages and length distribution from 100 runs.

Running condition: $\mathbf{x}_0 \sim \mathcal{N}(0, 1)$, $\mathbf{w} = (0.40, 0.40)$;

Summarized results: $CI_{ROS F}$ 75%, CI_{AC} 97%, CI_{HB} 97%; $R(\bar{T})_{ROS F} = 0.059$, $R(\bar{T})_{AC} = 0.084$; SP intervals here are $CI_{ROS F}$ s.

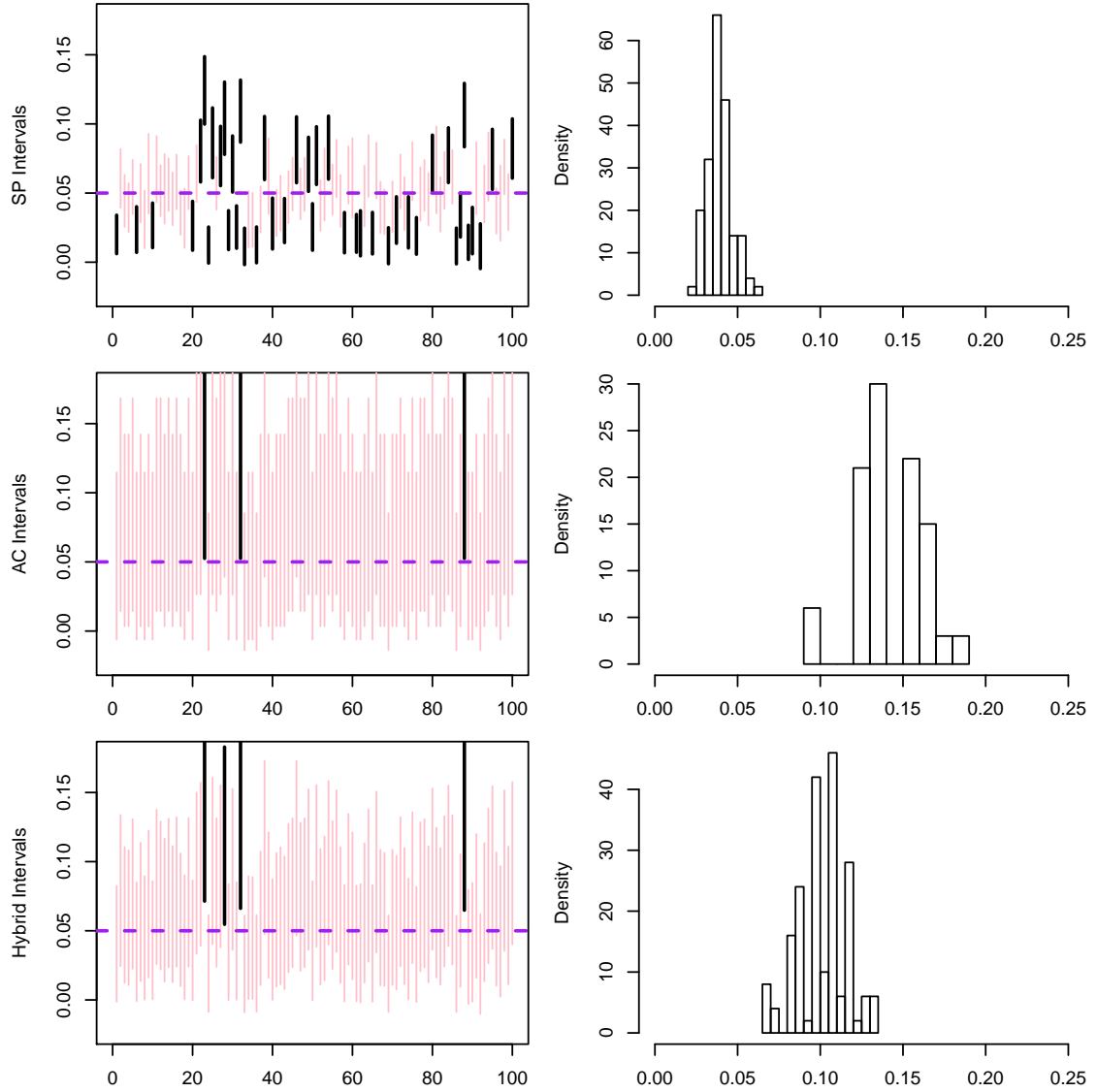


Figure 6.2: CI Coverage and length distribution from 100 runs.

Running condition: $\mathbf{x}_0 \sim \text{Logistic}(0, 1)$, $\mathbf{w} = (0.4, 0.4)$;

Summarized results: CI_{ROSF} 58%, CI_{AC} 97%, CI_{HB} 96%; $R(\bar{T})_{ROSF} = 0.050$,
 $R(\bar{T})_{AC} = 0.080$; SP intervals are CI_{OSF} here.

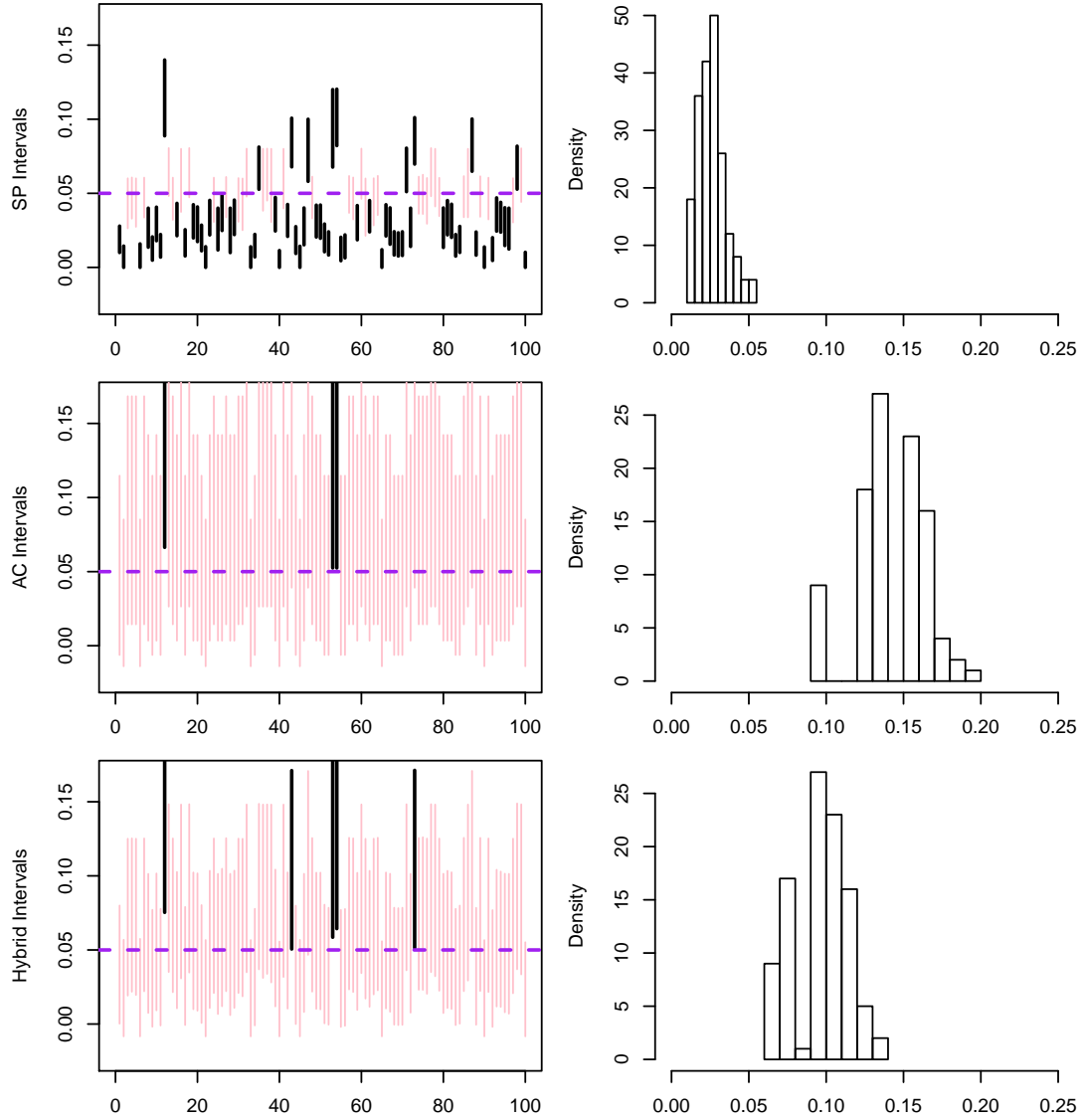


Figure 6.3: CI Coverages and length distribution from 100 runs.

Running condition: $\mathbf{x}_0 \sim \text{Uniform}(0, 50)$, $\mathbf{w} = (0.40, 0.40)$;

Summarized results: CI_{ROSF} 36%, CI_{AC} 97%, CI_{HB} 95%; $R(\bar{T})_{ROSF} = 0.046$,
 $R(\bar{T})_{AC} = 0.083$; SP intervals are CI_{ROSF} .

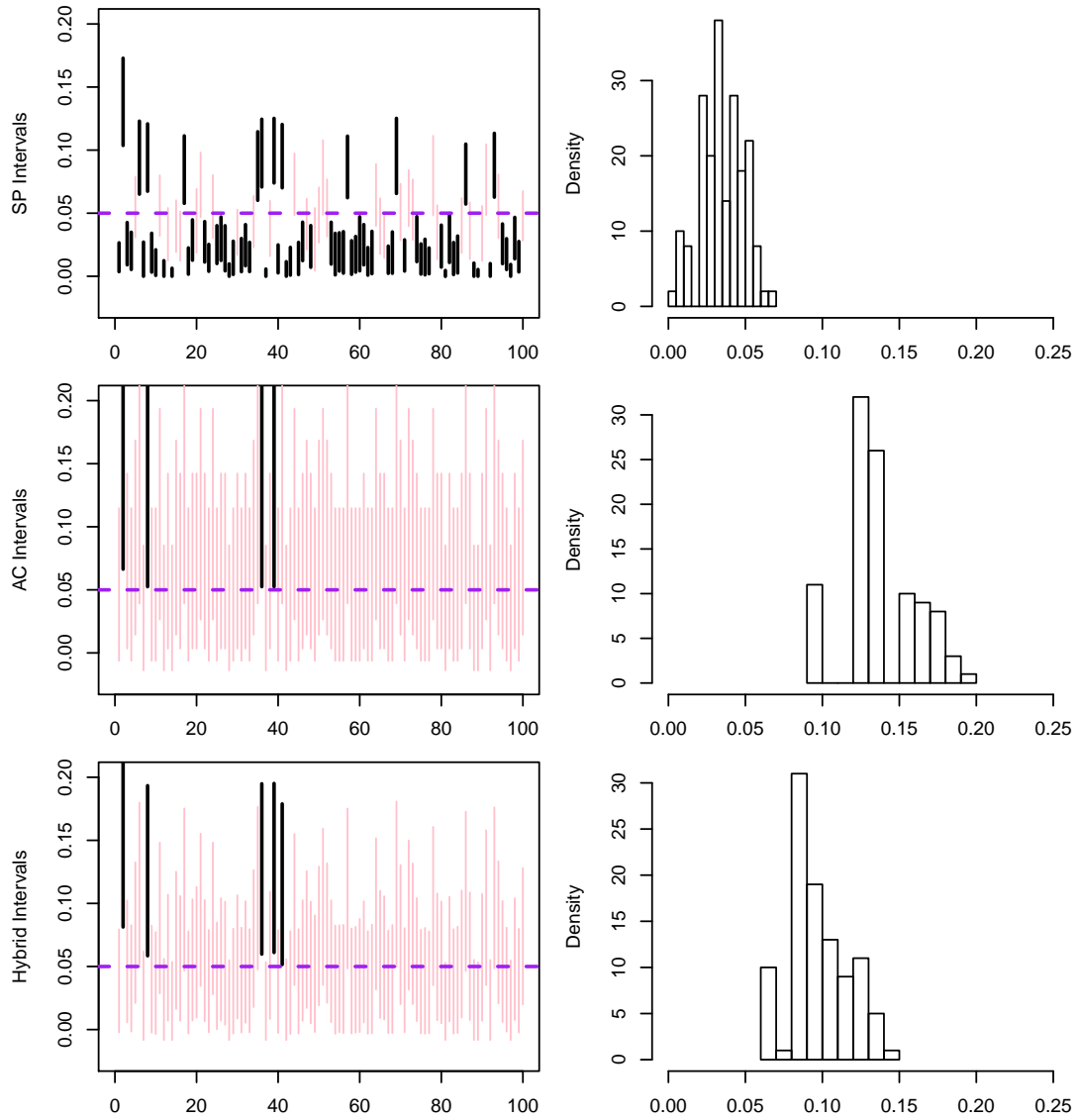


Figure 6.4: CI Coverages and length distribution from 100 runs.

Running condition: $\mathbf{x}_0 \sim \Gamma(5, 0.5)$, $\mathbf{w} = (0.40, 0.40)$;

Summarized results: CI_{ROSF} 31%, CI_{AC} 96%, CI_{HB} 95%; $R(\bar{T})_{ROSF} = 0.047$,
 $R(\bar{T})_{AC} = 0.084$; SP intervals are CI_{OSF} here.

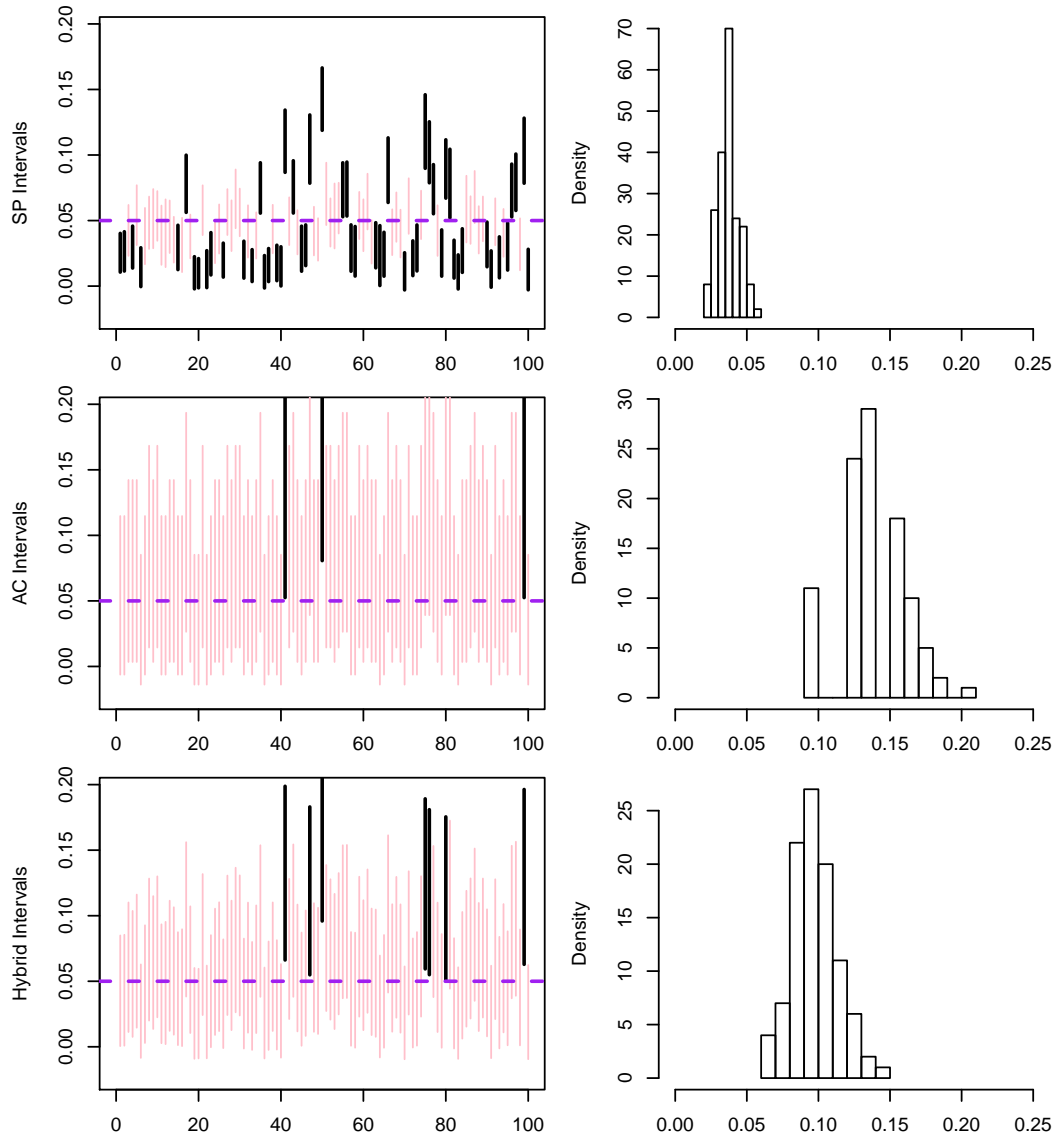


Figure 6.5: CI Coverages and length distribution from 100 runs.

Running condition: $\mathbf{x}_0 \sim \text{Exp}(1)$, $\mathbf{w} = (0.40, 0.40)$;

Summarized results: CI_{ROSF} 31%, CI_{AC} 97%, CI_{HB} 93%; $R(\bar{T})_{SP} = 0.046$,
 $R(\bar{T})_{AC} = 0.077$; SP intervals are CI_{OSF} here.

Figures 6.1 and Figure 6.5 show that choosing $\mathbf{w} = (0.40, 0.40)$ for the reference with any one of normal, logistic, uniform distributions, and $\mathbf{w} = (0.25, 0.40)$ for an exponential distributed reference sample we obtained a nice hybrid confidence interval for $R(t)$ which has a much shorter length while its coverage specified by the nominal confidence level is still maintained. The detailed results are shown in the captions of related figures. The gain of 15% \sim 30% shortening is significant in practical applications. We also notice that the point estimate for $R(t)$ seems more accurate for ROSF method than AC. The validation of this assertion requires a thorough investigation in the future. As mentioned previously, we also observed that a useful CI_{HB} can be obtained when the reference is different from the distributions presented above.

6.3.1 Guidelines for the choice of \mathbf{W}

From the discussion above, one will notice that the choice of the weight vector \mathbf{W} is essential in order to implement the HB methods. Although we do not have a rigorous way obtain the optimal \mathbf{W} has been found yet, we find the choice of \mathbf{W} does not entirely depend on the reference distributions. While a universal \mathbf{W} may exist. A rigorous treatment is not available so far. Clearly the coverage probability is a function of \mathbf{W} . Consider the case of Figure 6.1, treat The $\mathbf{W} = (w_1, w_2)$ as a variable instead of a specific value: $\mathbf{w} = (0.40, 0.40)$. Allow either w_1 or w_2 runs over $(0, 1)$. The result is plotted at Figure 6.6 as a contour picture. The x, y axis are w_1 and w_2 respectively. The numbers on the contour lines are coverage probabilities.

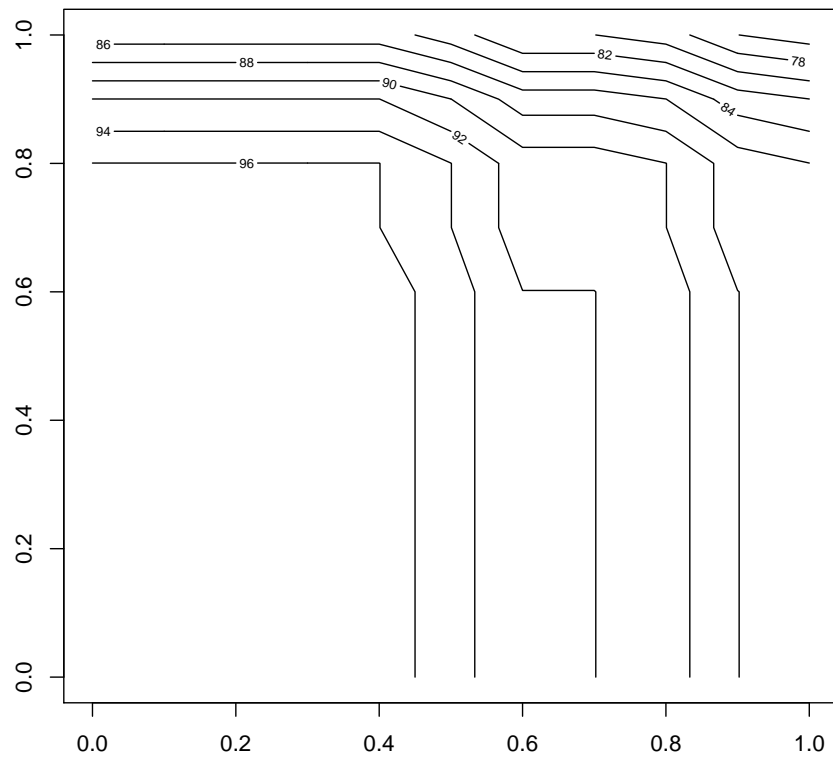


Figure 6.6: Coverage of CI_{HB} as a function of $\mathbf{w} = (w_1, w_2)$, $\mathbf{x}_0 : \mathcal{N}(0, 1)$, x axis is w_1 and y axis is w_2 .

It is obvious that the $\mathbf{w} = (0.0, 0.0)$ is the case CI_{AC} while the $CI_{ROS F}$ has the $\mathbf{w} = (1.0, 1.0)$ according to the formula proposed in 6.1. It seems that the choice $\mathbf{w} = (0.40, 0.40)$ is a feasible choice for a standard normal distributed reference sample to get a CI_{HB} with a 95 % coverage. The other reference scenarios have also been studied. Figure 6.7 has the same format as Figure 6.6, except that the reference distribution is different, and is stated in the captions. If these figures are observed collectively, it seems that $\mathbf{w} = (0.25, 0.4)$ is a good candidate for reference samples from a range of distributions to achieve approximately 95% coverage.

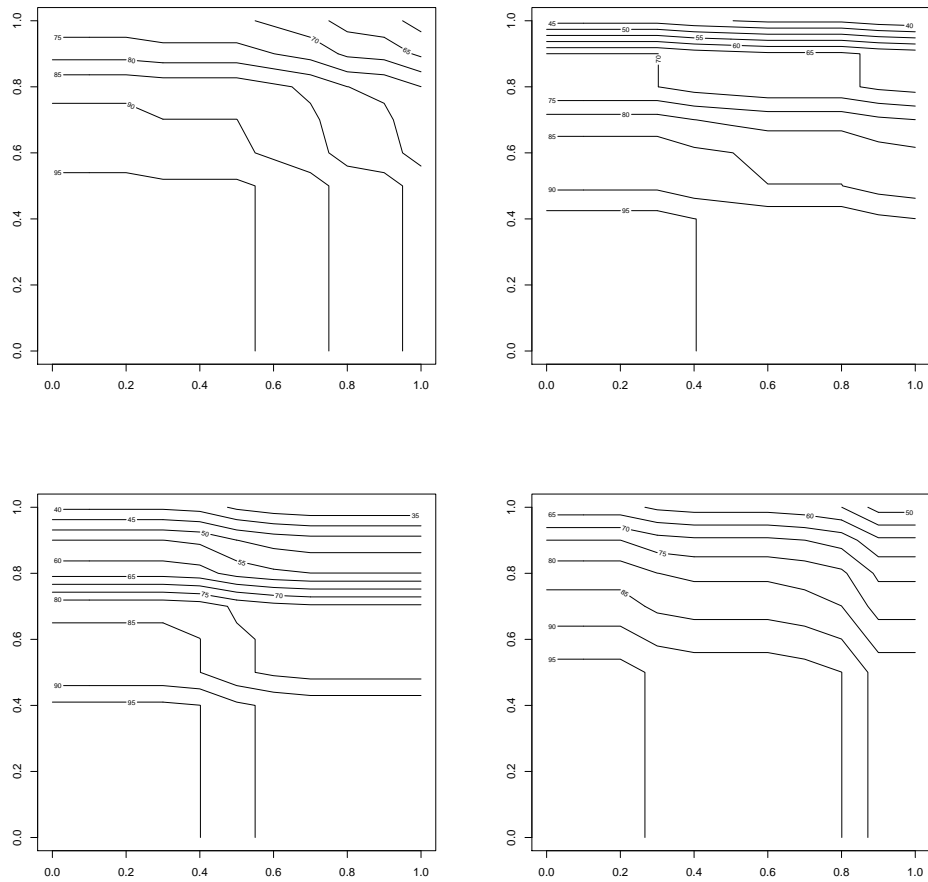


Figure 6.7: Coverage of CI_{HB} as a function of $\mathbf{W} = (w_1, w_2)$, top left: \mathbf{x}_0 : logistic(0, 1); top right: \mathbf{x}_0 : Uniform(0, 50); bottom left: \mathbf{x}_0 : $\Gamma(0, 1)$; bottom right: \mathbf{x}_0 : exp(0, 1)

Chapter 7

Analysis of the Testicular Germ Cell Tumor (TGCT) Data

In this chapter we apply the OSF DR model discussed in the previous chapters to the Testicular Germ Cell Tumor (TGCT) data set. These TGCT data have been analyzed by Voulgaraki et al. (2012) using a multivariate density ratio model. A summary of their work is made after a brief description of the TGCT data. Our analysis can be regarded as an extension of their work. The results show that the analysis with an OSF DR model not only confirms their inference, but also provide further evidence regarding the differences between cases and controls in the TGCT study.

7.1 Introduction

Germ cell tumors account for 90-95% of primary testicular tumors, which are the most common malignant tumors in men ages 20 to 35 years old in the United States. Approximately 9,000 new cases are diagnosed in the United States each year, and about 350 to 400 cases lead to deaths. Germ cell tumors are very sensitive to chemotherapy and cure rates are 90-95 %. It is commonly believed that the primary risk factor for the development of this cancer is undescended testicles, along with other risk factors: family history, physical activity, weight, dairy consumption, and

age at puberty.

7.2 Descriptive Analysis of the TGCT data

The TGCT data from the Servicemen’s Testicular Tumor Environmental and Endocrine Determinants Study (2002-2005) contain 763 case and 928 controls after removing the incomplete observations.

Table 7.1: TGCT data set

	SUJECT ID	Age years	CCTL (0-4)	Height cm	Weight kg	BMI	Race	Family History
1	TC10012SA	19	2	172.72	77.11	25.85	1	0
2	TC10022SA	33	2	177.80	81.65	25.83	1	0
3	TC10041SN	24	1	193.04	99.34	26.66	1	0
4	TC10050SA	23	0	182.88	83.92	25.09	1	0
5	TC10051SA	22	1	187.96	104.33	29.53	1	0
6	TC10060SA	26	0	170.18	64.41	22.24	1	0
7	TC10061SA	26	1	180.34	85.28	26.22	1	0
8	TC10073SA	42	3	172.72	70.31	23.57	1	0
9	TC10080SA	36	0	170.18	72.58	25.06	1	0
10	TC10081SA	36	1	180.34	80.29	24.69	1	0
	⋮		⋮		⋮		⋮	
	⋮		⋮		⋮		⋮	

The TGCT data have eight variables: subject ID, Age, an indicator for case or control (0=case, 1-4=control), Height (*cm*), Weight (*kg*), BMI (*kg/m²*) of participants, family history of testicular cancer (0=no, 1=yes) and race (1=white,2=black, 3=other). In this study, we focus on three variables: Height, Weight and Age (see Table 6.1). Table 7.2 shows the summary statistics for both case and control

groups.

Table 7.2: TGCT case-control summary statistics.

Variables	CCTL	Range	Mean	SD
Age	Control	18.00 ,46.00	27.91	5.93
	Case	18.00 ,45.00	27.82	5.99
Height (cm)	Control	152.4, 215.9	178.3	7.06
	Case	160.0, 203.2	179.6	7.03
Weight (kg)	Control	38.55, 127.01	80.13	11.14
	Case	50.80, 131.54	81.43	11.69

It is obvious that the controls and the cases are very similar in terms of their summary statistics. The similarity can be further demonstrated by histograms of Age, Height and Weight (see Figure 6.1). The patterns of the histograms for either case or control are almost the same. Normal curves have been added to the histograms, which have the same means and variances of the variables given in Table 7.2. The Age distribution is skew since this cancer is associated with younger groups. Scatterplot matrices for both case and control are shown in Figure 6.2. In order to illustrate the correlations, regression lines have been added. The correlation matrices are denoted by ρ_{Control} and ρ_{Case} :

$$\rho_{\text{Control}} = \begin{matrix} & \begin{matrix} \text{A} & \text{H} & \text{W} \end{matrix} \\ \begin{matrix} \text{A} \\ \text{H} \\ \text{W} \end{matrix} & \begin{pmatrix} 1.000 & -0.021 & 0.115 \\ -0.021 & 1.000 & 0.505 \\ 0.115 & 0.505 & 1.000 \end{pmatrix} \end{matrix} \quad \rho_{\text{Case}} = \begin{matrix} & \begin{matrix} \text{A} & \text{H} & \text{W} \end{matrix} \\ \begin{matrix} \text{A} \\ \text{H} \\ \text{W} \end{matrix} & \begin{pmatrix} 1.000 & 0.021 & 0.162 \\ 0.021 & 1.000 & 0.521 \\ 0.162 & 0.521 & 1.000 \end{pmatrix} \end{matrix}$$

Both scatterplot matrices and correlation matrices demonstrate that in either case the correlations among the variables are strikingly similar. Weight and Height

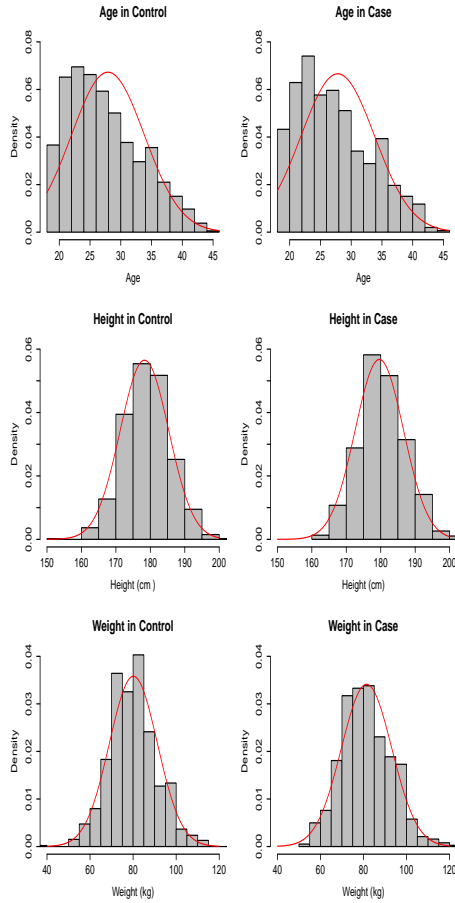


Figure 7.1: Histograms for case (right) and control (left)

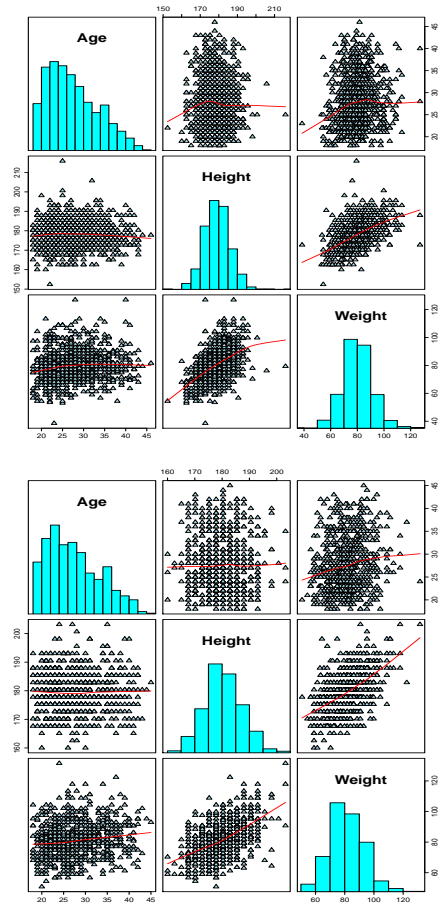


Figure 7.2: Scatterplot matrix for case (bottom) and control (top)

have a significant correlation ($\rho \approx 0.5$), and a relatively weak correlation between Age and Height is observed ($\rho \approx 0.15$). The correlation between Age and Weight is close to zero ($\rho \approx \pm 0.02$).

We see that the TGCT profile of both case and control are very similar, and that the descriptive data analysis does not differentiate the cases from the controls.

7.3 Analysis of TGCT data

Although the descriptive data analysis fails to detect any obvious discrepancy between the case and control, Voulgaraki et al (2012) pointed to a difference after analyzing the TGCT data with a density ratio model using a tilt function $h(\mathbf{t}) = \mathbf{t}$. Note that this tilt function holds when the reference and distortion distributions are normal with the same covariance matrices,

$$\mathbf{x}_0 \sim g_0(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}), \quad \mathbf{x}_1 \sim g_1(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}).$$

Taking \mathbf{x}_0 as the reference sample, the density ratio of the two pdfs is:

$$\frac{g_1(\mathbf{x})}{g_0(\mathbf{x})} = \exp\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_0' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1)\}. \quad (7.1)$$

Denote

$$\alpha = -\frac{1}{2}(\boldsymbol{\mu}_0' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1), \quad \boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0),$$

to obtain

$$\frac{g_1(\mathbf{x})}{g_0(\mathbf{x})} = \exp(\alpha + \boldsymbol{\beta}' \mathbf{x}), \quad (7.2)$$

which is a density ratio model with tilt function: $h(\mathbf{x}) = \mathbf{x}$.

Voulgaraki et al. (2012) show that (7.2) is justified by plotting \hat{G} vs \tilde{G} . They show that compared with generalized additive models (GAM), multiple regression (MR) and Nadaraya-Watson (NW) regression, the density ratio model leads to an

equally good regression of Weight given Height and Age due to efficient multivariate kernel density estimates based on multiple sources.

In this study, we use an OSF DR model. Either the control or the case sample of the TGCT data is combined with the same computer generated artificial sample separately. The artificial sample can be the sample with statistical structure resembling either case or control. The flexibility of choosing artificial samples lets us view the reference sample from different angles. We validated the results of Voulgaraki et al. (2012) by providing a graphical means to the difference between case and control.

7.3.1 Bivariate OSF DR Model for TGCT Data

Since Age has low correlation with Weight or Height and the Weight and Height are moderately correlated, it is reasonable to analyze the TGCT with a bivariate OSF DR model. According to what was previously demonstrated, the assumption of density ratio models will be better satisfied if the artificial samples are similar to the reference sample. We choose a normal sample with mean and variance of the pooled case and control sample as our artificial sample. To generate kernel density estimates of either case and control, the bandwidth matrices have been chosen according to the algorithm suggested by Voulgaraki et al. (2012), and are given below:

$$\mathbf{H}_{\text{Control}} = \begin{matrix} & \text{Height} & \text{Weight} \\ \text{Height} & \begin{pmatrix} 1.1010 & 0 \\ 0 & 1.7566 \end{pmatrix} \\ \text{Weight} & \end{matrix} \quad \mathbf{H}_{\text{Case}} = \begin{matrix} & \text{Height} & \text{Weight} \\ \text{Height} & \begin{pmatrix} 1.1236 & 0 \\ 0 & 2.1079 \end{pmatrix} \\ \text{Weight} & \end{matrix}$$

Figure 7.3 shows plots of joint kernel density estimates of Height and Weight for both case and control in contour and perspective styles. The table 7.3 shows the joint distribution of Height and Weight:

Table 7.3: joint distribution of $P(\mathbf{Height} \leq \text{Height}, \mathbf{Weight} \leq \text{Weight})$

Height	Weight	Control	Case
170.	60.	0.3165	0.3031
175.	60.	0.4033	0.3747
180.	60.	0.5244	0.4987
185.	60.	0.6335	0.6149
170.	70.	0.3165	0.3031
175.	70.	0.4033	0.3747
180.	70.	0.5244	0.4987
185.	70.	0.6335	0.6149
170.	80.	0.3165	0.3031
175.	80.	0.4033	0.3747
180.	80.	0.5244	0.4987
185.	80.	0.6335	0.6149
170.	90.	0.3165	0.3031
175.	90.	0.4033	0.3747
180.	90.	0.5244	0.4987
185.	90.	0.6335	0.6149

7.3.2 Trivariate OSF DR models for TGCT data

The joint distribution of Age, Height and Weight of case and control have also been evaluated. The artificial sample used follows a normal distribution having the mean vector and covariance matrix of the pooled data from both case and

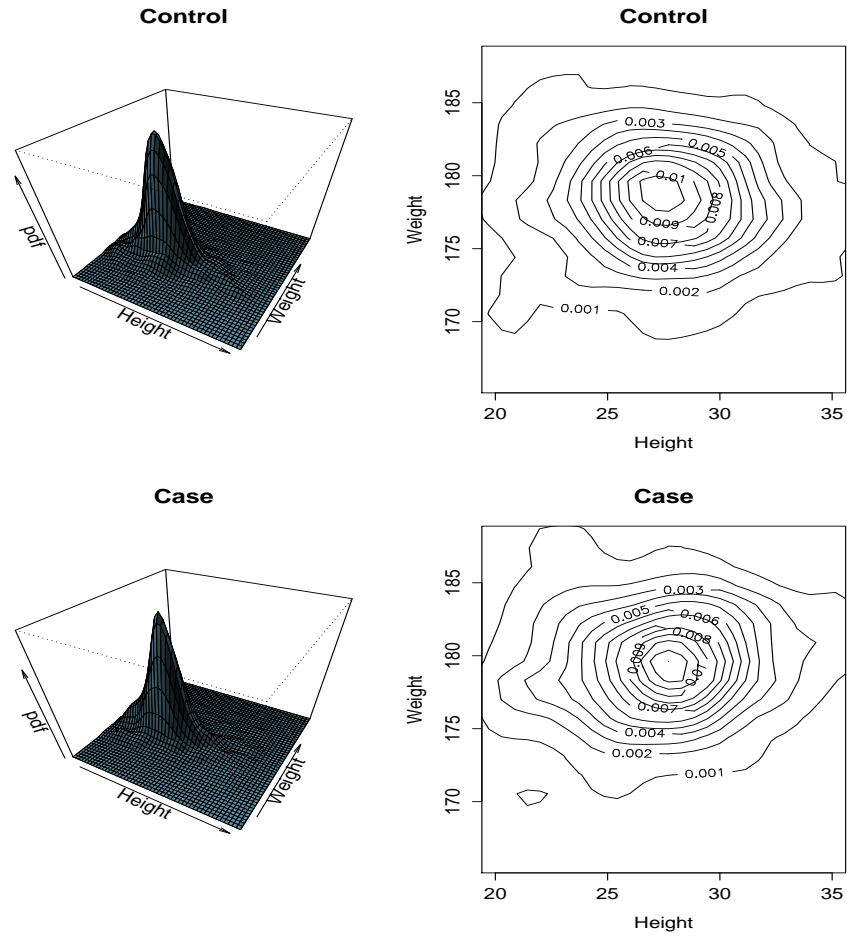


Figure 7.3: 3D plot for kernel density estimates of Height and Weight

control. The resulting pdfs are difficult to visualize since they have four dimensions. For illustration purposes, the third variable is fixed at designated values and the resulting pdfs (for example $f(A, H, W = w_0)$) can be viewed in contour plots. These

Table 7.4: The fixed variables and their values in contour plots

Figures	Variable fixed	Values of the fixed variable fixed
Figures 7.4 , 7.5	Weight (kg):	72,76,80,84,88,92
Figures 7.7 , 7.8	Height (cm):	172.1,174.7,177.3,179.9,182.5,185.1

contour plots are in general similar although some differences in the area away from the contour center can be discerned. However these discrepancies result from the sparseness of the observations.

We next consider conditional pdfs. Figures 7.6, and 7.9 show the estimated $f(W|A, H)$'s and $f(H|A, W)$'s, respectively. For both case and control, these conditional densities are quite similar. However once the conditional values increase, the conditional densities show different behaviors. The conditional densities can be used in the estimation of the corresponding conditional expectations.

It is important to note that the results obtained from different artificial data samples are strikingly similar, but this is not shown here.

7.4 Summary

In this chapter we have implemented a multivariate version of the OSF method. The results confirm perviously published results graphically using plots of joint and conditional pdfs.

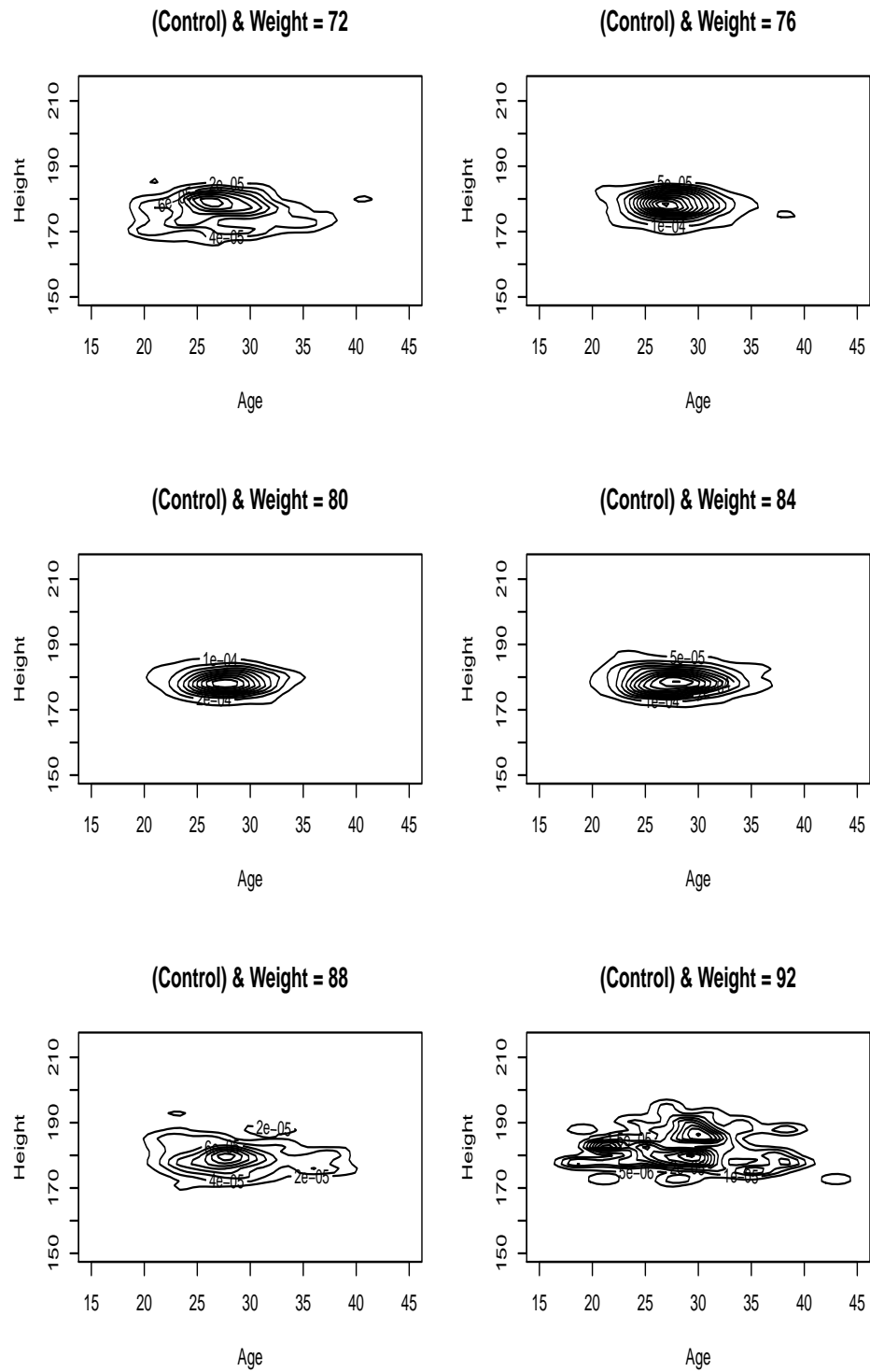


Figure 7.4: Contour plots for the control pdf for fixed weights

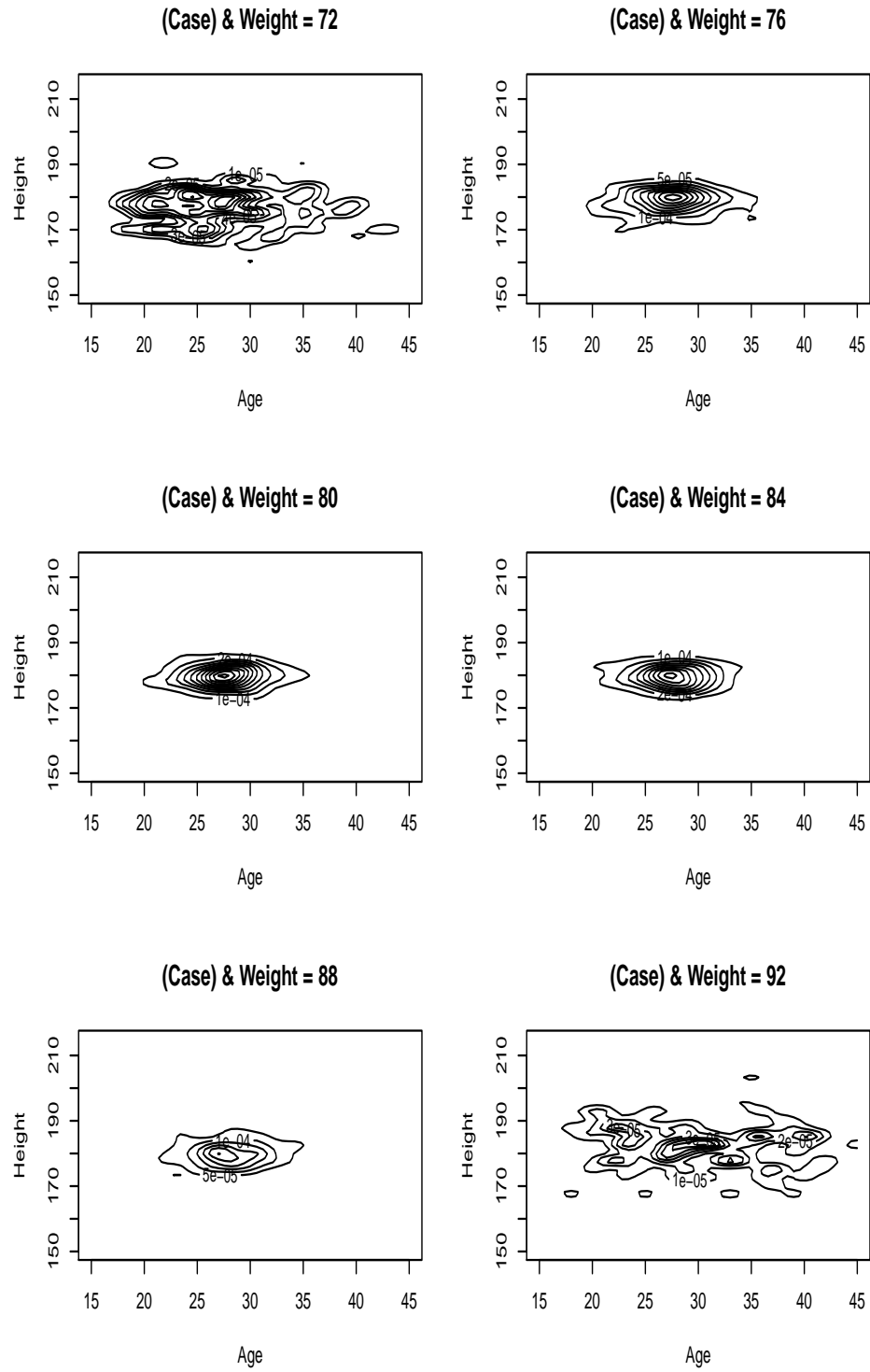


Figure 7.5: Contour plots for the case pdf for fixed weights

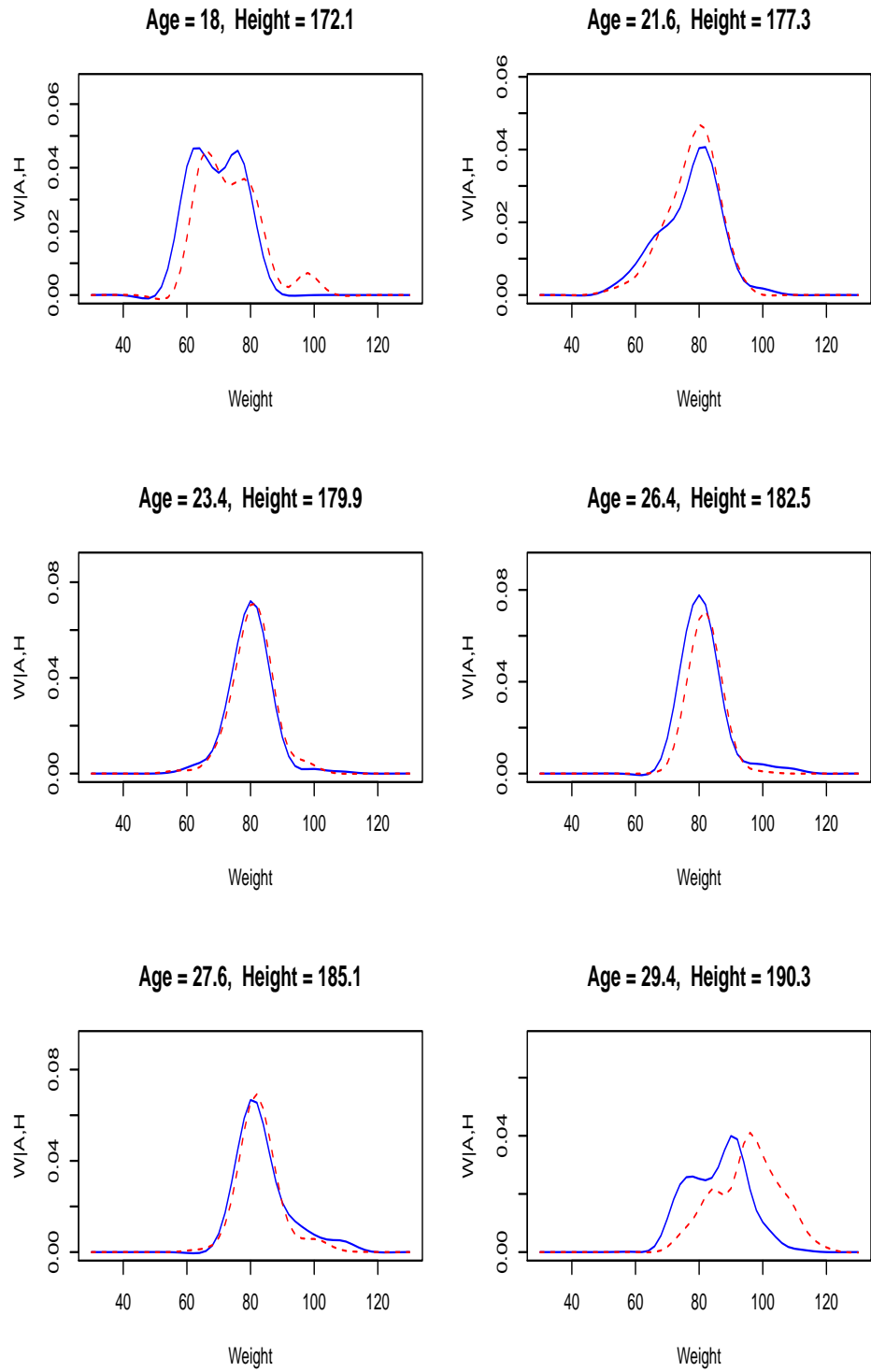


Figure 7.6: Pdf plots conditional on Age and Height. Control: solid line; Case: dashed line.

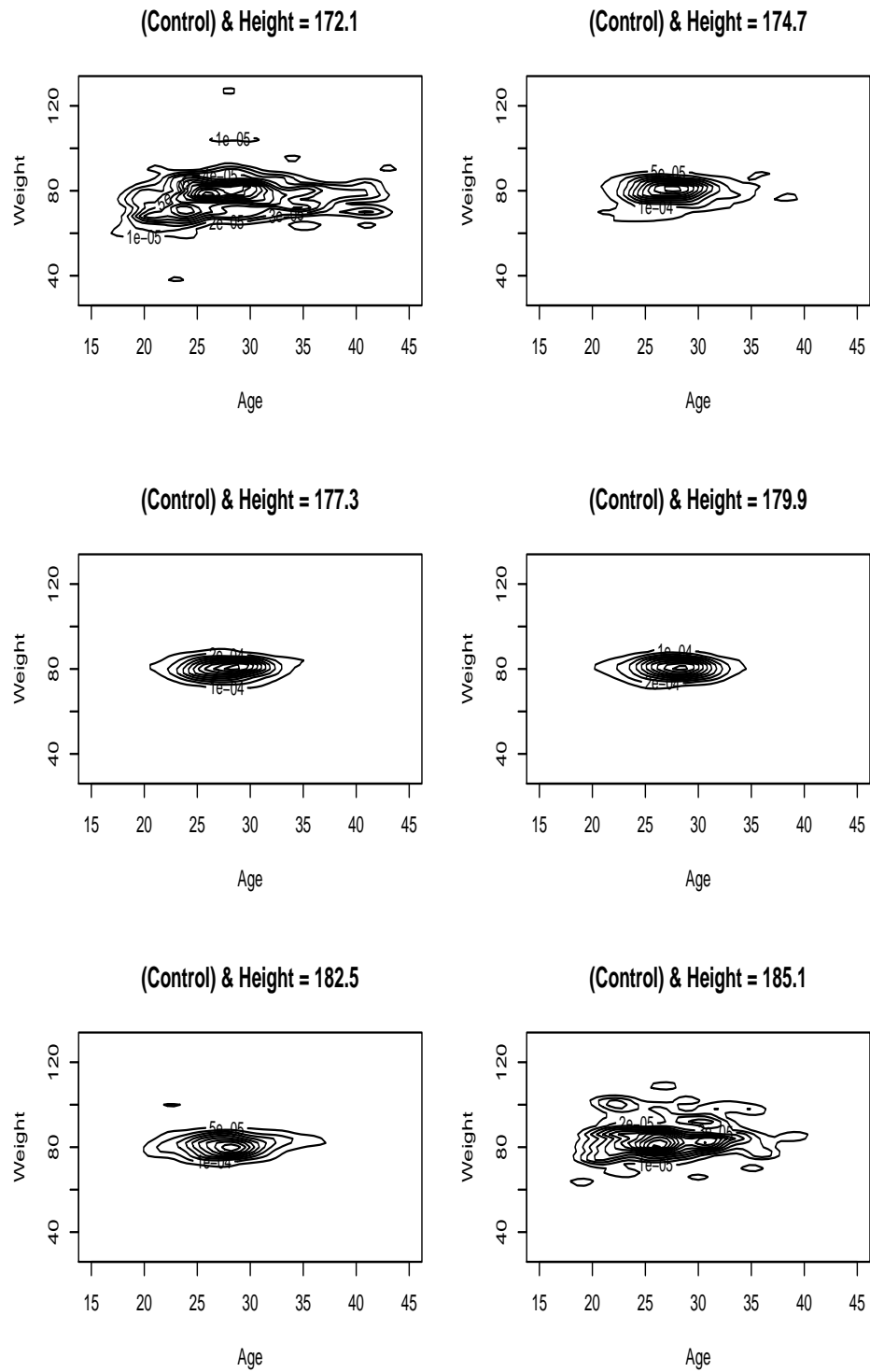


Figure 7.7: Contour plots for the control pdf for fixed Heights

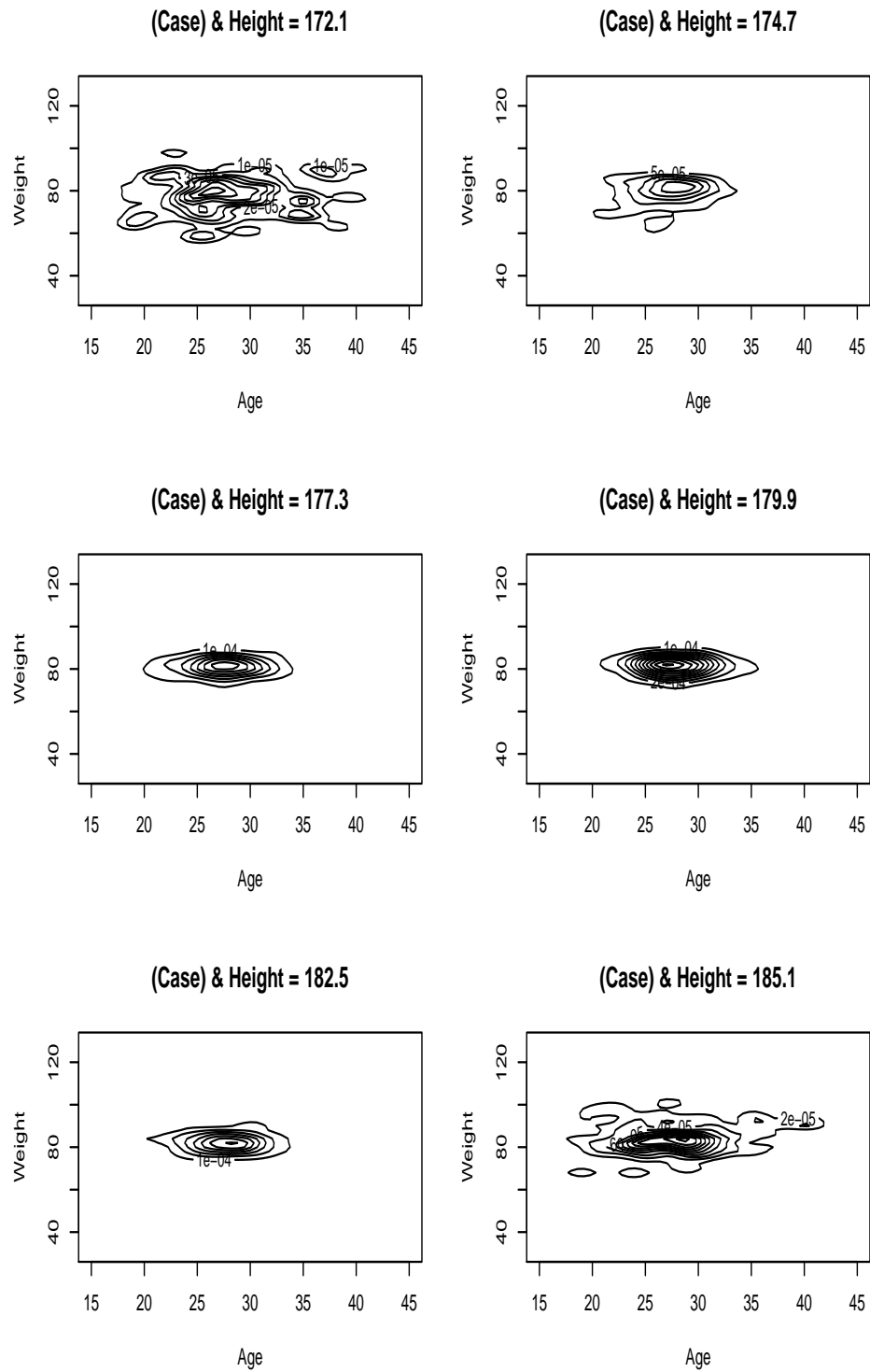


Figure 7.8: Contour plots for the case pdf for fixed Heights

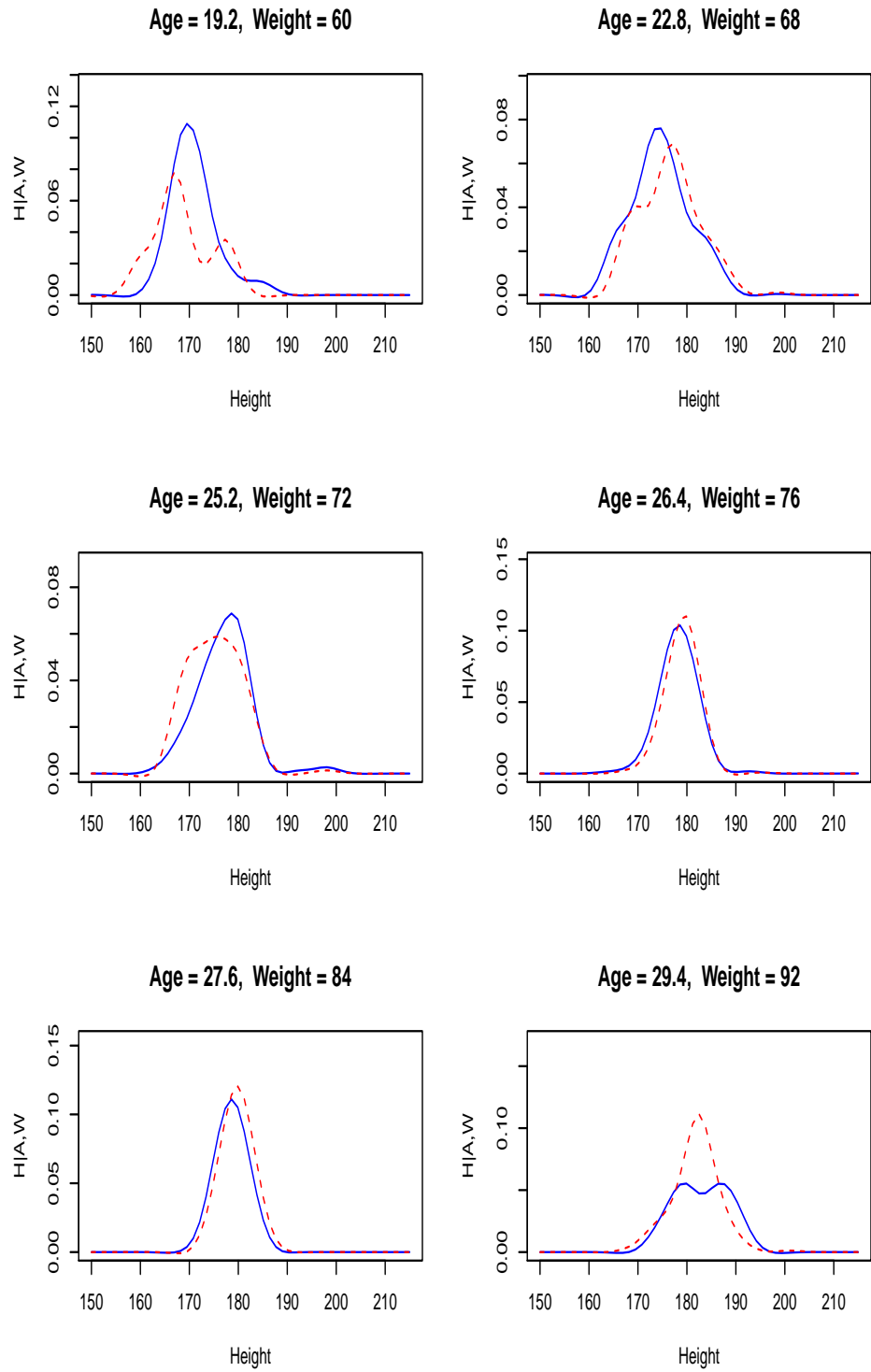


Figure 7.9: Pdf plots conditional on Age and Weight. Control: solid line; Case: dashed line.

Chapter 8

Analysis of Microarray Data of Colon Cancers

8.1 Introduction

Microarray data are now routinely used to analyze how gene expression profiles relate to certain diseases. Extensive literature reports how to identify individual genes according to their significantly differential expressions using different strategies. These strategies lead to numerous *R* packages such as *qvalue* and *limma* which are involved in this study. The popularity of these packages is due to their handling of pairwise multiple tests. However, it is also believed that a given disease may relate to expression profiles of groups made of several genes instead of individual genes (Guillot et al., 2007).

In this chapter, we apply the OSF DR model to a case-control microarray data using similar approaches as those applied to the TGCT data. The artificial sample employed in this study is a sample from a normal distribution with a mean and variance that are computed from the pooled sample of both case and control. This analysis has there are two goals :

1. Differentiate the case and control by identifying the most significantly differentially expressed gene groups.

2. Detect the association of member genes among the groups in either the case or the control.

For sets of multiple genes, the computation cannot be carried out on all possible sets (think about how big is $\binom{5000}{2}$ if we know $\binom{500}{2} = 124,750$). Guillot (2007) believed that genes involved in the best sets of 2 or 3 genes were those ranking high in terms of univariate differential expression. We restricted the search in a list of top 200 genes in differential expression and ranked genes according to their occurrence in the best sets of genes. We decided to use the same strategy to alleviate our computation burden.

Note the choice of 2 or 3 genes in a designated gene group is arbitrary and is done entirely for demonstration and data visualization purposes. The method and strategy employed here are also applicable to gene groups containing more than 3 genes. It is our purpose to demonstrate an application of the density ratio model with OSF to high dimensional data and to provide an alternative strategy to tackle the microarray interpretation.

8.2 Description of Microarray Data of Colon Cancer

We use a colon microarray data contain 68 subjects (30 controls and 38 cases) and 5339 genes. The data format is shown in Table 8.1.

Table 8.1: Part of colon cancer microarray data : 5339 genes and 68 subjects: 30 in control, 38 in case

ID	Type	238493_at	1562133_x.at	1559616_x.at	235687_at	...
1	Control	2.8539	-0.6096	-1.2855	-1.0888	...
2	Control	0.5906	0.0024	-0.6383	-0.7287	...
3	Control	0.6176	-0.4620	-0.7241	-0.3482	...
4	Control	0.8220	-0.5143	-0.6485	-1.0201	...
5	Control	2.0971	-0.6756	-0.8039	-0.5683	...
6	Control	0.1188	-0.6273	0.0188	-0.8524	...
...	
63	Case	1.2171	-0.2321	-1.5178	-0.2983	...
64	Case	0.1281	-0.3300	-0.7632	-0.0570	...
65	Case	3.6614	-0.3115	-1.8012	-0.7528	...
66	Case	1.5992	0.5000	-1.1992	-0.4584	...
67	Case	0.4309	0.2062	-0.5090	-0.7522	...
68	Case	1.5078	-0.1746	-0.7188	-0.6654	...

8.3 Microarray Data Analysis

To analyze that the microarray data by the OSF DR model we follow the following steps:

1. Use Package *qvalue* to rank the significance of differentially expressed genes in the case and control. The top 200 genes have been identified according to the significance of their differential expression between the case and control.
2. Select the most significant gene groups (every group contains 2 or 3 genes in this study) by the OSF DR models. The artificial samples involved are sampled from a normal distribution with the pooled mean and variance from both the case and control. The significance is based on their *q*-values, which are elaborated in the R package *qvalue*. They can be obtained from the *p*-value

of the null hypothesis: case and control have the same distribution.

3. Visualize the association between the member genes in a two-gene group, the scatterplots of the top 4 pair genes are plotted. Moreover, their joint pdfs are also plotted as contour plots.
4. Plot scatterplots for the top 4 of the most differentially expressed three-gene groups.

8.4 Results and Discussion

The ranking of the significantly differentially expressed gene groups can be performed according to the procedure specified in the previous section: identify first 200 most significantly expressed individual genes; Then use the two-sample multivariate OSF DR model to test every pair of genes among 200 individual genes ($\binom{200}{2} = 19,900$ pairs), which leads to 19,900 p -values. Finally, rank the q values according to the R *qvalue* package.

It brought our attention that the association among the member gene in gene groups is not related to the significance of differential expression. The scatterplots of the top 4 gene pairs are illustrated in Figure 8.1. Their linear regression line and GAM line have been added to illustrate the trends. Obviously although the first pair is the most significantly expressed gene pair, the trends for both case and control are similar from either linear regression or GAM. However, the third pair seems to have a very different association among the member genes in the case and the control. We believe that these association information could be important

for the colon cancer study and diagnoses. The gene pairs which have significantly differential expressions between case and control deserve special attention if there are different associations among the member gene between the case and control. One may believe intuitively that similar association among the member genes in either group is due to some effects unrelated to the colon cancer regardless how significant their differential expressions are. Indeed, it is also reasonable to investigate the association among member genes even in some less differentially expressed gene groups.

Figures 8.2 and 8.3 show their joint pdfs. These contour pictures can illustrate not only their significance of expression (comparing the contour centers) but also visualize their association between the two genes in the pair. It is believed that the shapes and curvatures are related to the association. A metric could be developed in the future to quantify their associations based on these contour plots.

Figure 8.4 illustrates the scatterplots of first 4 groups from the most significantly expressed three-gene groups. Table 8.2 shows the top 10 individual genes, 4 gene pairs and 4 three-gene groups according to their rankings in terms of differential expression between the case and control. It is obvious:

1. Genes involved in the best sets are not those ranking high in terms of univariate differential expression.
2. For 3-gene groups, some genes appear more frequently in many of the best sets: such as *1554970_at* which appears in groups 1 and 4, all first 4 groups include *206349_at*. This trend does not appear in the best 2-gene groups.

Table 8.2: Ranking of gene or gene groups according to their significantly differential expressions between case and control of colon cancer microarray data.

number gene	Rank	gene or genes in gene groups
1	1	<i>220423_at</i>
1	2	<i>224144_at</i>
1	3	<i>206025_at</i>
1	4	<i>236285_at</i>
1	5	<i>206349_at</i>
1	6	<i>208288_at</i>
1	7	<i>204844_at</i>
1	8	<i>217240_at</i>
1	9	<i>208105_at</i>
1	10	<i>1553857_at</i>
1
2	1	{ <i>218623_at</i> , <i>205656_at</i> }
2	2	{ <i>218623_at</i> , <i>209934_s_at</i> }
2	3	{ <i>218623_at</i> , <i>203305_at</i> }
2	4	{ <i>203042_at</i> , <i>203305_at</i> }
2
3	1	{ <i>1554970_at</i> , <i>217240_at</i> , <i>206349_at</i> }
3	2	{ <i>230412_at</i> , <i>206349_at</i> , <i>224144_at</i> }
3	3	{ <i>1555612_s_at</i> , <i>230412_at</i> , <i>206349_at</i> }
3	4	{ <i>1554970_at</i> , <i>1555612_at</i> , <i>206349_at</i> }
3

8.5 Summary

This chapter analyzes microarray data from colon cancer and identifies the most significantly differentially expressed gene pairs or three-gene groups. For the gene pairs, the associations between the two genes are also illustrated. We believe that the associations between genes may shed more light on the colon cancer diagnose.

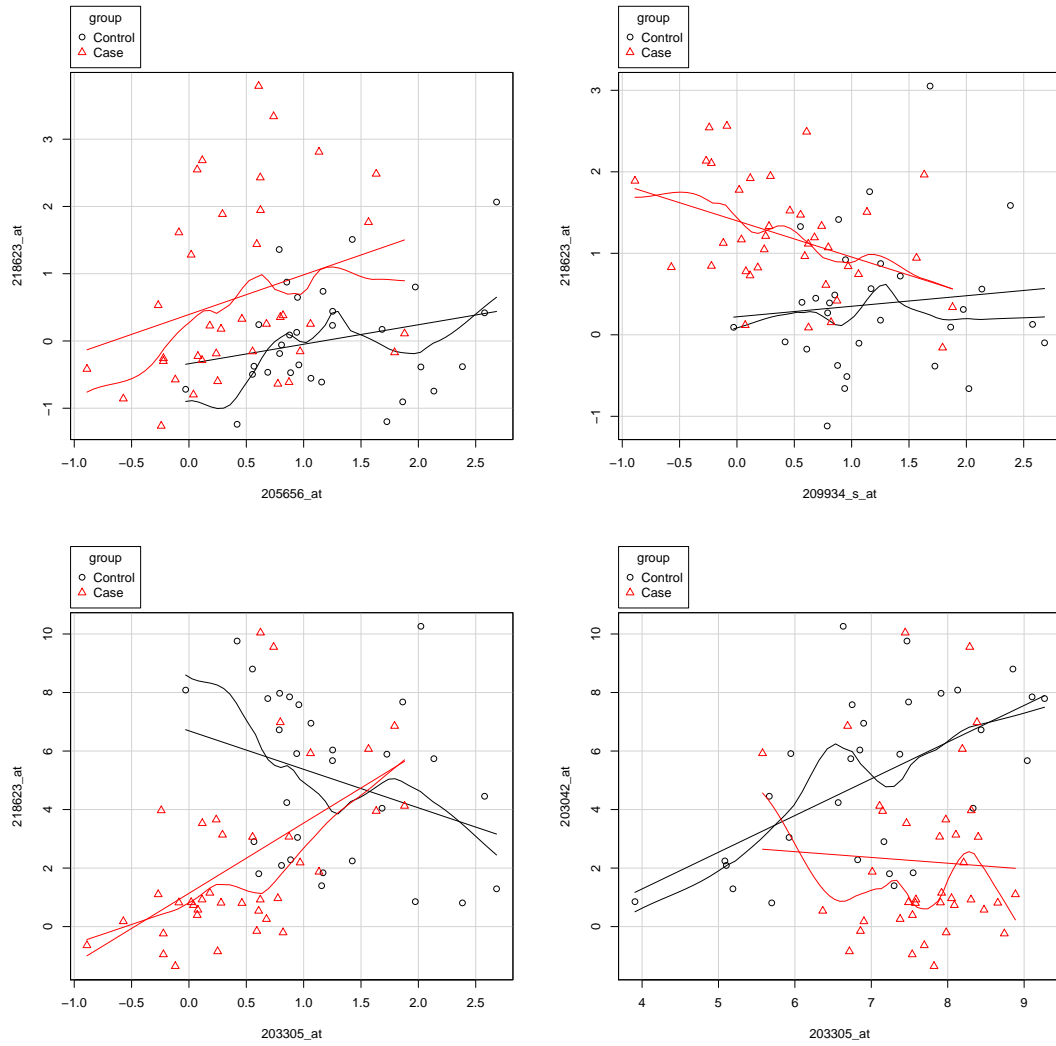


Figure 8.1: The scatterplots of 4 gene pairs which are most significantly differentially expressed. The linear regression line and GAM line are also plotted to illustrate the trend.

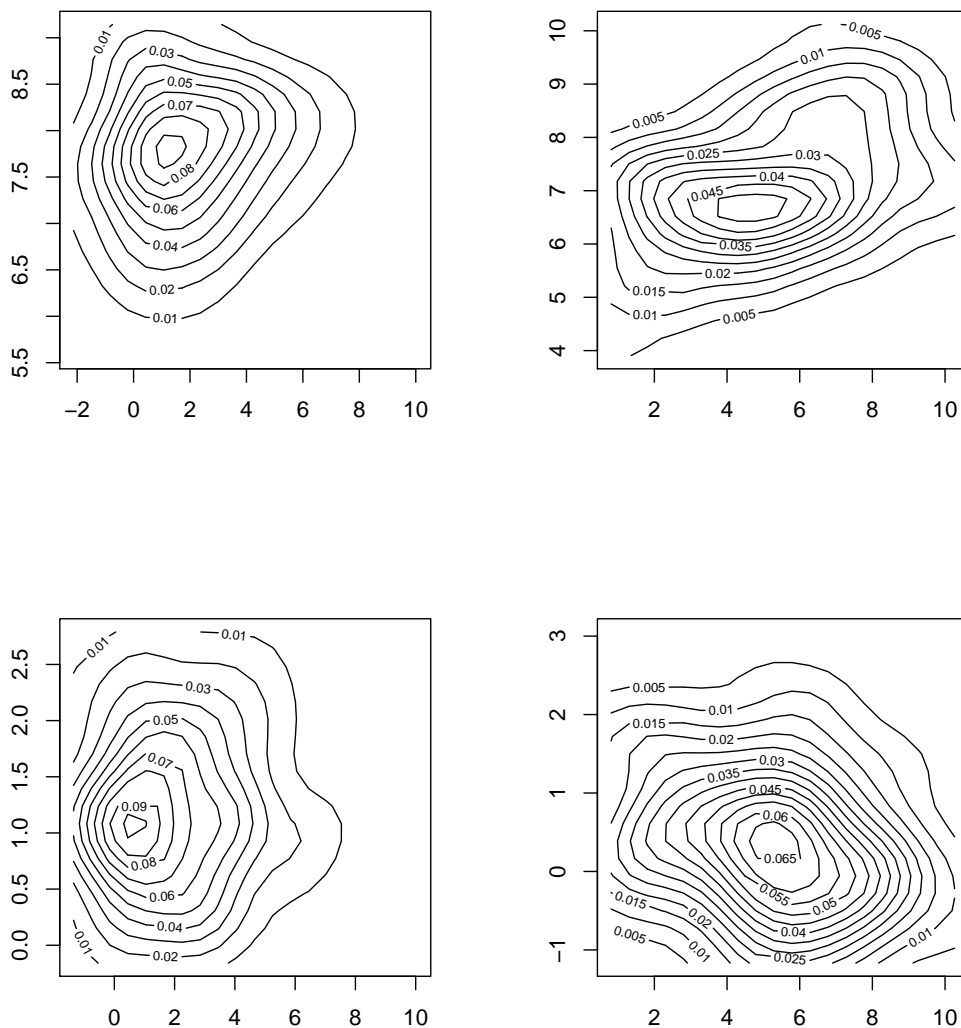


Figure 8.2: The contour plots of joint *pdf* for No.1 and No.2 gene pairs which are most significantly differentially expressed. Left: Control; Right: Case. Top: group 1; Bottom: group 2.

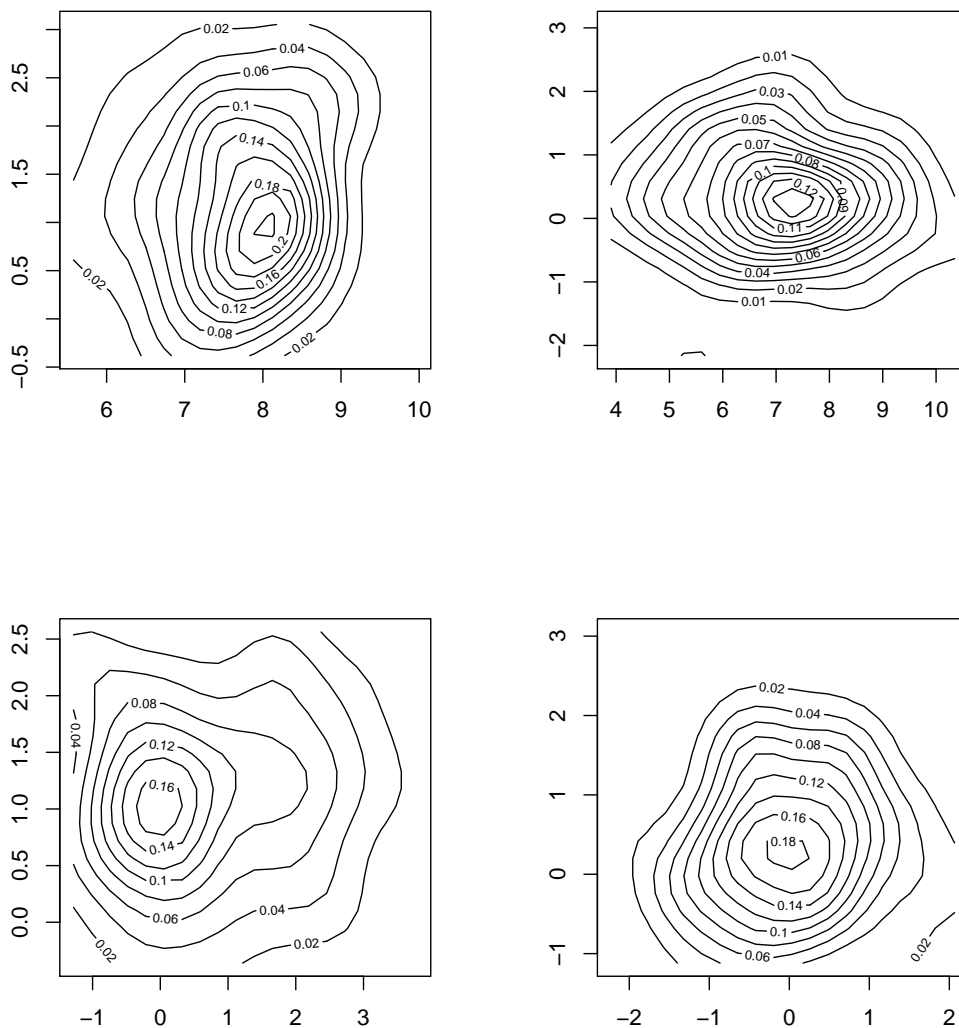


Figure 8.3: The contour plots of joint pdf for No. 3 and No. 4 gene pairs which are most significantly differentially expressed. Left: Control; Right: Case. Top: group 3; Bottom: group 4.

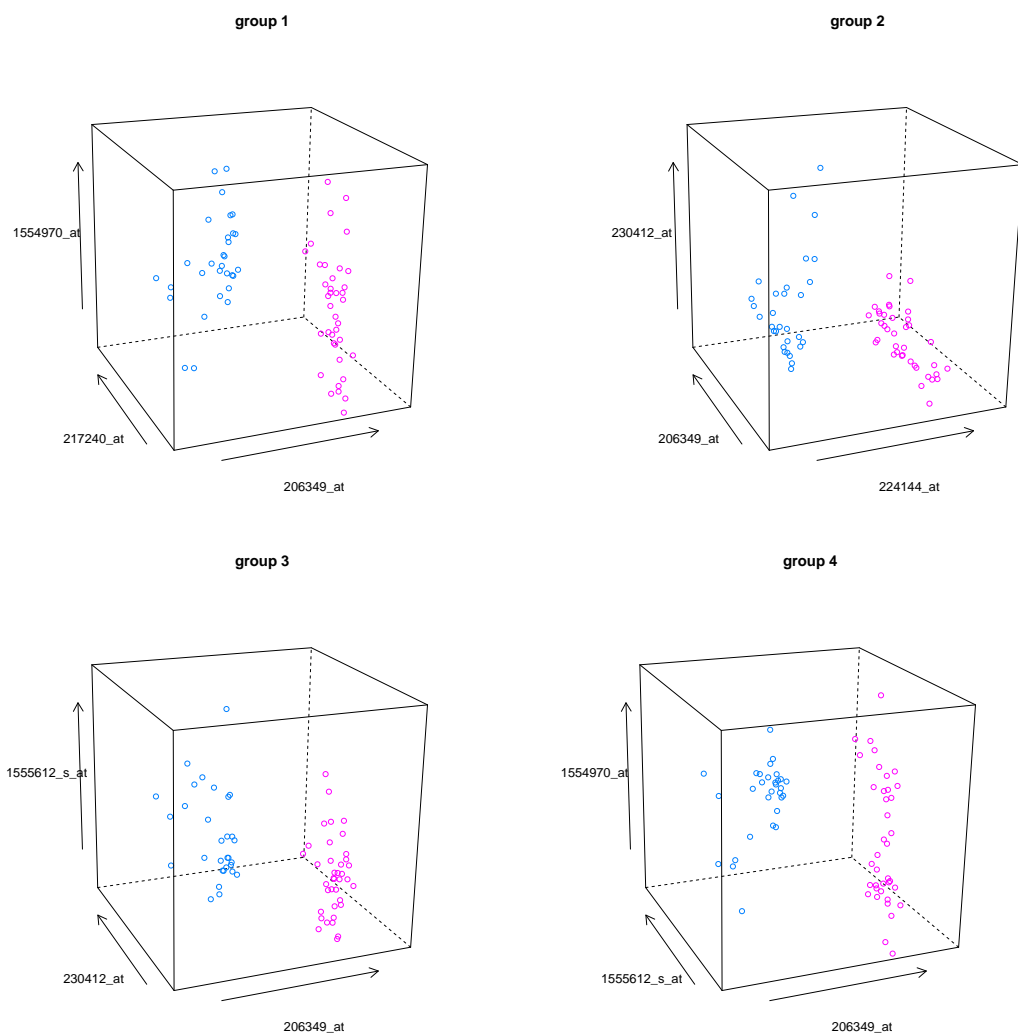


Figure 8.4: The scatterplots of first 4 three-gene groups which are most significantly differentially expressed. The **right** are the observation from the control

Appendix A

R package: Density Ratio

The R package *DensityRatio* is developed by Wen Zhou, University of Maryland. It computes estimates for parameters, distributions and densities in either typical density models or out-of-sample fusion density ratio models (OSF DR). All figures in this dissertation are graphed by this packages. Please find a brief summary of the functions included in the *DensityRatio* package, along with a detailed description, function codes and examples in this appendix.

A.1 Summary of functions in the *DensityRatio* package

- **CONDITION** calculates and plots the conditional kernel densities for a pair of 3-covariate samples. It gives conditional pdfs, conditional means and variances also. The algorithm of choosing bandwidth used to compute the kernel pdf is from Voulgaraki et al (2012).

```
CONDITION(Y1,Y2,index_1=10, index_2=10,gridsize=20,variable="A")
```

Y1: First 3-covariate samples,

Y2: Second 3-covariate samples,

gridsize: The number of grid points,

```

variable: Column name of the samples,
index_1: index for the first column,
index_2: index for the second column.

#### Example #####

data(TGCT)

Y1 <- Control

Y2 <- Case

CONDITION(Y1,Y2,index_1=10, index_2=10,variable='C')

```

- **CONTOUR** plots the kernel density for two 3-covariate samples. It depends on CONTOUR.0, CONTOUR.1, CONTOUR.2 and CONTOUR.3.

```

CONTOUR <- function(Y1,Y2,gridsize=20,variable="A",a)

Y1: First of 3-covariate samples,
Y2: Second of 3-covariate samples,
gridsize: The number of grid points.
variable: Column name of the samples,
a: index for the column,

### example ###

data(TGCT)

Y1 <- Control

Y2 <- Case

CONTOUR(Y1,Y2, variable='C',a=10)

```

- **SE** is the most important function in this package. It requires a matrix input consisting of several artificial samples and a real data reference sample as its columns. It is a multiple univariate version of function SP2K. It outputs the estimates for the parameters and probabilities assuming a density ratio model. It also calculates confidence intervals for OSF DR, AC, Wilson, and EP (Wald) methods. The function SP.AC.EP is the direct application of function SE. To run use:

X: matrix consisting of multiple univariate samples of the same length, the last component being the reference sample.

T: threshold value

Example

```
X <- cbind(rnorm(100,1,1),runif(100,-3,3),rnorm(100))
```

```
SE(X)
```

Note: rnorm(100) is the "real data" reference sample.

- **SIMULATION.FUSION** This executes repeated out of sample fusion using hybridization for simulation purposes. We obtain confidence intervals for ROSF, AC, and EP (Wald) methods.

The reference sample is simulated from any one of the following x_0 distributions: normal, gamma, uniform, logistic and exponential distribution.

```
SIMULATION.FUSION(dist='normal',N=NULL,a=NULL,b=NULL,
```

```
K=100,alpha=0.05,W=c(0.4,0.4))
```

dist: specify the distribution

K: Number of runs.

N: number of observations in exact sample.

Note: `x_1` is simulated 50 times from different uniform distributions as explained on page 51 (Section 5.2 of Chapter 5).

W: a vector which is the relative weight of SP vs AC

```
SIMULATION.FUSION(dist='normal',N=50,a=0,b=1,K=10)
```

```
#           SP    AC    HB
# Mean      0.064  0.09  0.077
# Width     0.079  0.15  0.122
# Coverage  70.000 90.00 90.000
```

```
### runif(50,0,50)
```

```
SIMULATION.FUSION(dist='uniform',N=50,a=0,b=50,K=5)
```

- **SP2K** is from Professor Benjamin Kedem. It implements a two-sample univariate density ratio model. It compute estimate of parameters, probabilities and threshold probabilities. It also perform equidistribution test and model validations. It is the base for function `SE`, which implements a multivariate univariate sample version density ratio model and the function `TWOSAMPLE.DR`, which implements two-sample multivariate version of density ratio models.

```
SP2K(x1,x2,Increment=0.05,BandWidth=0.5,T),
```

```
x1:vector of sample value,univariate sample only,
```

x2:vector or sample value (reference sample),univariate only,

Increment:Increment controls the grid at which est.

g,g1,G,G1 are evaluated, default value is 0.05.

BandWidth:Bandwidth for calculation kernel smoothing density,

it controls smoothness of kernel est. of g,g1

and the default value is 0.5.

T: Threshold

Example

```
SP2K(rnorm(100,1,1),rnorm(100),Increment=0.05,BandWidth=0.5,T=1.645)
```

- **TWOSAMPLES.DR** implements two sample density ratio models for both univariate and multivariate samples. It is the base for functions **CONDITION** and **CONTOUR**. It also performs an equidistribution test.

```
TWOSAMPLES.DR(x1,x2)
```

x1:First sample}

x2:Second sample (reference sample)

```
### EXAMPLE 1 (Univariate) #####
```

```
TWOSAMPLES.DR(rnorm(100,1,1),rnorm(100))
```

```
### EXAMPLE 2 (Multivariate) #####
```

```
x1 <- cbind(rnorm(80,1,1),rnorm(80,0.5,1))
```

```
x2 <- cbind(rnorm(100,0,2),rnorm(100))
```

```
TWOSAMPLES.DR(x1,x2)
```

DensityRatio

January 31, 2013

CONDITION

Conditional Densities for Two 3-Covariate Samples

Description

The function `CONDITION` calculates and plots the conditional kernel densities for two 3-covariate samples. The samples should have identical column names. The function also computes conditional mean and variance.

Usage

```
CONDITION(Y1,Y2,index_1=10, index_2=10,gridsize=20,variable="A")
```

Arguments

<code>Y1</code>	First of the two 3-covariate samples, which is required to have a column name
<code>Y2</code>	Second of the two 3-covariate samples, which is required to have the exactly same column name as that of the first sample
<code>gridsize</code>	The number of grid points.
<code>variable</code>	Column name of the samples, three levels only: 'A', 'B', 'C'
<code>index_1</code>	index for the first column
<code>index_2</code>	index for the second column

Details

The function `CONDITION` calculates and plots the conditional kernel densities for two 3-covariate samples, it needs the functions: `CONDITION.0`, `CONDITION.1` and `CONDITION.2`.

Value

<code>M</code>	Conditional mean.
<code>SE</code>	Conditional variance.

Note

This function needs `CONDITION.0`, `CONDITION.1`, `CONDITION.2`, `DR.KS`

Author(s)

Wen Zhou

See Also

DR.KS, CONDITION.0, CONDITION.1,CONDITION.2

Examples

```
#### Example 1 #####
### 1. Load the data.

data(TGCT)

### 2. Create the two 3-covariate samples.

Y1 <- Control[1:100,]
Y2 <- Case[1:100,]

### 3. Read the function CONDITION's arguments.

args(CONDITION)

### 4. Plot the conditional pdf of C condition on A and B. level b and c

CONDITION(Y1,Y2,index_1=10, index_2=10,variable='C')

#### example 2 #####
Y1 <- cbind(rnorm(100,1,1),rnorm(100,0.5,1),rnorm(100))
Y2 <- cbind(rnorm(100,2,1),rnorm(100,0.5,1.2),rnorm(100))
colnames(Y1) <- c("Col_1","Col_2","Col_3")
colnames(Y2) <- c("Col_1","Col_2","Col_3")

CONDITION(Y1,Y2,index_1=10,index_2=10)
```

CONTOUR

Contour plot for kernel densities of two 3_covariate samples

Description

The function CONTOUR plots the kernel density for two 3-covariance samples.

Usage

```
CONTOUR <- function(Y1,Y2,gridsize=20,variable="A",a)
```

Arguments

Y1	First of the two 3-covariate samples, which is required to have a column name
Y2	Second of the two 3-covariate samples, which is required to have the exactly same column name as that of the first sample
gridsize	The number of grid points.
variable	Column name of the samples, three levels only: 'A', 'B', 'C'
a	index for the column, scalar or vector. The choose of a value depends on gridsize.

Details

The function CONTOUR plots the kernel densities for two 3-covariance samples.

Value

value value of the last variable

Note

It depends on CONTOUR.0, CONTOUR.1, CONTOUR.2

Author(s)

Wen Zhou

See Also

CONTOUR.0, CONTOUR.1, CONTOUR.2

Examples

```
### Code ###

#####
### Function: CONTOUR
###                               Wen Zhou 8/24/2012
#####

CONTOUR <- function(Y1,Y2,gridsize=20,variable="A",a) {
  Name <- colnames(Y1)
  LIST <- CONTOUR.0(Y1,Y2,gridsize)
  L1 <- switch(variable,
    A=LIST[[1]],
    B=LIST[[2]],
    C=LIST[[3]])

  L2 <- switch(variable,
    A=c(Name[2],Name[3],Name[1]),
    B=c(Name[1],Name[3],Name[2]),
    C=c(Name[1],Name[2],Name[3]))

  CONTOUR.1(L1[[1]],L2,a=a)
  dev.new()
```



```

CONTOUR.1(L1[[2]],L2,a=a)

}

#####
### Function: CONTOUR.0, CONTOUR.1
###                               Wen Zhou  8/24/2012
#####
### a is a scalar or vector
CONTOUR.0 <- function(Y1,Y2,gridsize=20) {

  fhat <- function(Y1,Y2, gridsize=20) {
    RRR1 <- DR.KS(x2=Y1, gridsize=gridsize)
    RRR2 <- DR.KS(x2=Y2, gridsize=gridsize)
    list(RRR1,RRR2)
  }

  L_C <- fhat(Y1,Y2,gridsize=gridsize)
  L_B <- fhat(cbind(Y1[,1],Y1[,3],Y1[,2]),cbind(Y2[,1],Y2[,3],Y2[,2]),gridsize=gridsize)
  L_A <- fhat(cbind(Y1[,2],Y1[,3],Y1[,1]),cbind(Y2[,2],Y2[,3],Y2[,1]),gridsize=gridsize)

  list(L_A=L_A,L_B=L_B,L_C=L_C)

}

CONTOUR.1 <- function(S,L=NULL,gridsize=51,a) {
  k <- length(a)
  if(k==1) {
    contour(S$grid[[1]],S$grid[[2]],S$estimate[, ,a],xlab=L[1],
            ylab=L[2],cex.main=0.8,
            main=paste(L[3], ' = ',round(S$grid[[3]][a]),sep=''))
  }
  else{
    n <- ifelse(k/2-ceiling(k/2)==0,k/2,(k+1)/2)
    par(mfrow=c(n,2))

    for(i in 1:k) {
      contour(S$grid[[1]],S$grid[[2]],S$estimate[, ,a[i]],
              xlab=L[1],ylab=L[2],
              main=paste(L[3], ' = ',round(S$grid[[3]][a[i]],2),sep=''))
    }
    return(round(S$grid[[3]][a],2))
  }
}

### example 1 ###

### 1. Samples.

data(TGCT)

### 2. Create the two 3-covariate samples.

```

```
Y1 <- Control[1:100,]
Y2 <- Case[1:100,]

### 3. Read the function CONTOUR's arguments.
head(CONTOUR)
### 4. Run

CONTOUR(Y1,Y2, variable='C',a=c(10,12,14,16))

### Two contour plots

#### example 2 #####
Y1 <- cbind(rnorm(100,1,1),rnorm(100,0.5,1),rnorm(100))
Y2 <- cbind(rnorm(100,2,1),rnorm(100,0.5,1.2),rnorm(100))
colnames(Y1) <- c("Col_1","Col_2","Col_3")
colnames(Y2) <- c("Col_1","Col_2","Col_3")

CONTOUR(Y1,Y2, variable='C',a=c(10,12,14,16))
```

DATA DESCRIPTION	<i>Dataset (TGCT)</i>
------------------	-----------------------

Description

Dataset included.

Details

Case, Control

Author(s)

Wen Zhou, Benjamin Kedem

Maintainer: Wen Zhou <wenzhou@math.umd.edu>

References

Kedem, Qin, Zhang

See Also

ks

Examples

```
data(TGCT)
head(Control)
head(Case)
```

Description

The function MULTISAMPLES.DR estimates the parameters, their standard deviation, confidence interval for threshold probability and test the equidistribution for multiple univariate samples according to density ratio model.

Usage

```
SE(X, T=1.645, tilt="normal")
```

Arguments

X	matrix made from multiple same length univariate sample, the last component is the reference sample
T	threshold value
tilt	tilt function, only normal and gamma available

Details

This function estimates parameters and their standard deviation, threshold probability and its confidence interval and test the equidistribution of all samples according to density ratio models. The same sizes should be equal.

Value

par	estimate for parameters.
p	Jump probability.
Prob	confidence intervals for SP, AC, EP

Note

This is the multiple sample version of SP2XXSQQ.

Author(s)

Wen Zhou

See Also

TWOSAMPLES.DR

Examples

```
#####
### Function: SE compute the confidence intervals for SP(OSF),AC, EP.
### X is made from sample vectors, T is threshold
###                               Wen Zhou           1/31/2013
#####

SE <- function(X,T=1.645,tilt="normal") {

### X = cbind(x1,x2,..., xm, x0)

m <- ncol(X) -1
n <- nrow(X)
N <- (m+1)*n

### t = c(x1',x2',...,xm',x0')' # fusion of samples plus reference.

t <- NULL
for (i in 1:(m+1)){
t <- c(t,X[,i]) } # t has N elements

### 1. Functions involves ###

### a. H(x)=c(1,h(x))=h(x)=(1,x,x^2); x is a n X 1 vector
if (tilt=="normal")
{ H <- function(x) rbind(rep(1,n), x , x^2)}
else if (tilt=="gamma") {
H <- function(x) rbind(rep(1,n), x , log(x))
}

else { stop("tilt is either normal or gamma")}

### b. w = exp(alpha + beta*x + gamma*x^2), w is a n X 1 vector.
### It is w(theta, x)*I(x <= a)

w <- function(theta,x,a=max(x)) {
IN <- exp(colSums(theta*H(x)))
ifelse(x<=a,IN,0) }

### c. MINUS log-likelihood function:
###

minusloglike <- function(theta){
A <- B <- rep(0,n)
Theta <- matrix(rep(0,3*m),m)

for(i in 1:m) {
Theta[i,] <- theta[(3*i-2):(3*i)]
A <- A + w(Theta[i,],t)
B <- B + log(w(Theta[i,],X[,i])) }

### log(1 + sum (w)) - sum( alpha + beta*x + gamma*x^2)
```

```

sum(log(rep(1,n) + A)) - sum(B) }

### 2. Parameter Estimation: parameter, W ###

### Maximizing loglikelihood by minimizing MINUS loglikelihood

min.func <- nlmnb( start=rep(c(-0.02,.2,.2),m),obj = minusloglike)

### Parameter estimates: alpha_hat, beta_hat, gamma_hat, p_hat

parameter <- matrix( min.func$par, 3,m)
row.names(parameter)=c("alpha","beta","gamma")

### c. W(t) = cbind(w1(t), w2(t), ... , wm(t))
### Weight Matrix: nrow=m X n; ncol=m

W <- function(x,a=max(x)) {
weight <- matrix(rep(0,m*N),N)

for (j in 1:m) {
weight[,j] <- w(parameter[,j],x,a=max(x))}
return(weight) }

### p=dG,

p <- ( n*(1 + rowSums(W(t))) )^(-1)

#### 3. S calculation

### Lu' Dissertation,P24: A, B, C calculations

### (I0)_ij = int w_i*w_j dG/(1 + colSums(W))
### = sum(w_i*w_j*p*np)=sum(n*w_i*w_j*p^2)
### since dG=p/(1+colSums(W))

I0 <- t(W(t))
I1 <- t(W(t))
I2 <- t(W(t))
I3 <- t(W(t))
I4 <- t(W(t))

Amat <- I0 # Lu' Dissertation: P23

### Bmat: made by m*m blocks, every block result from
### h(x)=(x,x^2)';
### odd row is I1, even row is I2

Bmat <- matrix(rep(0,2*m*m), m*2)

for (i in 1:m) {
Bmat[2*i-1,] <- I1[i,]
Bmat[2*i,] <- I2[i,]
}

### Cmat: made by m*m blocks, every block resulted from
### h(x)h'(x) = matrix((x,x^3, x^3, x^4), nrow=2)

```

```

C_01 <- matrix(rep(0,2*m*m), m*2)

### odd row is I2, even row is I3

for (i in 1:m) {
  C_01[2*i-1,] <- I2[i,]
  C_01[2*i,]   <- I3[i,]
}

C_02 <- matrix(rep(0,2*m*m), m*2)

### odd row is I3, even row is I4

for (i in 1:m) {
  C_02[2*i-1,] <- I3[i,]
  C_02[2*i,]   <- I4[i,]
}

### Cmat: 2*m X 2*m, odd column is C_01, even column is C_02

Cmat <- matrix(rep(0,4*m*m), m*2)
for (i in 1:m) {
  Cmat[,2*i-1] <- C_01[,i]
  Cmat[,2*i]   <- C_02[,i]
}

### E1 = int w_i*t dG = sum(w_i*t*p)

E1 <- colSums(p*W(t)*t)
E2 <- colSums(p*W(t)*t^2)
E3 <- colSums(p*W(t)*t^3)
E4 <- colSums(p*W(t)*t^4)

### E_0: vector m X 1 blocks, every blocks is a 2X1 vector resulted
###      from h(x,x^2)

E_0 <- rbind(E1,E2)

### construct Emat (diagonal matrix with 2X1 blocks as diagonal element)

L <- rep(0,m); L <- as.list(L)
for (i in 1:m){
  L[[i]] <- E_0[,i] }

library(Matrix)
Emat <- bdiag(L)
Emat <- as.matrix(Emat)
### construct Emat_bar ( diagonal matrix with 2X2 matrix as diag element )

### EE_01 is 1 X m blocks, every block made by 4X1 vector:

EE_01 <- rbind(E2,E3,E3,E4)

### construct block diagonal matrix with matrix(EE_01[,i],2) as diag blocks.

LL <- rep(0,m); LL <- as.list(LL)

```

```

for (i in 1:m) {
LL[[i]] <- matrix(EE_01[,i],2)}

Emat_bar <- bdiag(LL)

### S calculation

rho <- diag(m); RR <- rho

S11 <- rho - rho
S12 <- rho
S12 <- as.matrix(S12)
S21 <- t(S12)
S22 <- RR
S22 <- as.matrix(S22)

S1 <- cbind(S11,S12); S2 <- cbind(S21,S22)

S <- rbind(S1,S2)/m

### 4. Computation of Standard Deviation of Parameters ###

### Sigma calculation: Lu's Dissertation P28

J <- function(num_1) matrix(rep(1,num_1*num_1),num_1)
O <- function(num_1,num_2) matrix(rep(0,num_1*num_2),num_2)

mat_0 <- rbind(J(m),0(m,2*m) )

mat <- cbind(mat_0,0(m*2,m*3))

Sigma <- solve(S) - (m+1)*mat
dia_SS <- diag(Sigma)
dia_S <- dia_SS*(dia_SS > 0)
par_SE_0 <- sqrt(dia_S/N)

par_SE <- matrix(par_SE_0,3)
row.names(par_SE)=c('alpha','beta','gamma')

### 5. Computation of variance of probability ###

### Guanhua Lu' Dissertation P50, A_t, B_t Calculation

### A_t = int w_j I(t <=x) dG/(1+sum(W))=sum(w_j I(t<=x)*np^2

A_t <- function(x) {
  colSums(np^2*W(t,x)) }

### B_t'(B_tp) has 1 X m blocks, every block is 1X2 vector due to h(x)

B_tp <- function(x) {
B_t1 <- function(x) {
  colSums(np^2*W(t,x)*t) }
B_t2 <- function(x) {
  colSums(np^2*W(t,x)*t^2) }

B_tt <- rbind(B_t1(x),B_t2(x))

```

```

B_tp <- rep(0,2*m)

for (i in 1:m) {
  B_tp[2*i-1] <- B_tt[1,i]
  B_tp[2*i] <- B_tt[2,i]
}
return(B_tp)
}

### Calculation of G(t) and Covariance

G_Lu <- function(x) sum(p[t<=x])
G_EP <- function(x) sum(X[,m+1]<=x)/n

### Guanhua Lu's Dissertation P72, Covariance of G(t)

A_B <- function(x) c(A_t(x), B_tp(x))

Term_1 <- function(x) G_Lu(x)*(1-G_Lu(x))
Term_2 <- function(x) sum(A_t(x))
Term_3 <- function(x) A_B(x)

Var_Lu <- function(x) m*(Term_1(x)-Term_2(x)) + Term_3(x)

Var_EP <- function(x) G_EP(x)*(1-G_EP(x))

### make sure Variance always are nonnegative
### Confidence Interval Calculation

SE_Lu <- function(x) ifelse(Var_Lu(x)>0,sqrt(Var_Lu(x)/N),0)
SE_EP <- function(x) sqrt(Var_EP(x)/n) # not m*n

U_EP <- function(T) 1 - G_EP(T)
U_Lu <- function(T) 1 - G_Lu(T)

U_Lu_L <- U_Lu(T) - 1.96*SE_Lu(T)
U_Lu_U <- U_Lu(T) + 1.96*SE_Lu(T)

U_EP_L <- U_EP(T) - 1.96*SE_EP(T)
U_EP_U <- U_EP(T) + 1.96*SE_EP(T)

Lu <- c(U_Lu(T), SE_Lu(T), U_Lu_L,U_Lu_U)

EP <- c(U_EP(T), SE_EP(T), U_EP_L,U_EP_U)

### Agresti Coull (AC) Interval Calculation

p0 <- sum(X[,m+1] > T)
z <- 1.96
n_AC <- n + z^2
p_AC <- (p0 + z^2/2)/n_AC
meanAC <- p_AC

```



```

SE_AC <- sqrt(p_AC*(1-p_AC)/n_AC)
LAC <- p_AC - 1.96*SE_AC #L
UAC <- p_AC + 1.96*SE_AC #U

AC <- c(p_AC,SE_AC,LAC,UAC)

#### Wilson Score Interval Calculation
p.w <- p0/n
term1 <- (p0 + z^2/2)/(n + z^2)
term2 <- z*sqrt(n)/(n + z^2)*sqrt(p.w*(1-p.w) + z^2/(4*n))

WL <- max(0, term1-term2)
WU <- min(1,term1+term2)

pW <- term1
SE_W <- WU-WL

Wilson <- c(term1, (WU-WL)/(2*z),WL,WU)

Prob <- rbind(Lu,AC,Wilson,EP)
colnames(Prob)=c("Prob", "SE", "L", "U")
rownames(Prob)=c("Lu", "AC", "Wilson", "EP")

### 6. LR test for equidistribution and gamma=0 ###

### H_0: all the samples are the same

logL <- - minusloglike(min.func$par)
logL0 <- - minusloglike(rep(0,3*m))
LR <- -2*(logL0 - logL)
pvalue_LR <- 1 - pchisq(LR,3*m)

### H_0: gamma=0

H_test <- function(x) rbind(rep(1,n), x)

### b. w = exp(alpha + beta*x + 0*x^2), w is a n X 1 vector.
### It is w(theta, x)*I(x <= a)

w <- function(theta,x,a=max(x)) {
  IN <- exp(colSums(theta*H_test(x)))
  ifelse(x<=a,IN,0) }

### c. MINUS log-likelihood function:

minusloglike_0 <- function(theta){
A <- B <- rep(0,n)
Theta <- matrix(rep(0,2*m),m)

  for(i in 1:m) {
    Theta[i,] <- theta[(2*i-1):(2*i)]
    A <- A + w(Theta[i,],t)
    B <- B + log(w(Theta[i,],X[,i])) }

### log(1 + sum (w)) - sum( alpha + beta*x + gamma*x^2)

sum(log(rep(1,n) + A)) - sum(B) }

```

```

min.func_0 <- nlm(b(start=rep(c(-0.02,.2),m),obj=minusloglike_0)
parameter_0 <- matrix(min.func_0$par,2,m)
row.names(parameter_0)=c("alpha","beta")

logL00 <- -minusloglike_0(min.func_0$par)
LR0 <- -2*(logL00 - logL)
pvalue_LR0 <- 1- pchisq(LR0,m)

list(par=parameter,par_SE=par_SE,par0=parameter_0,
S=S,pvalue_LR=pvalue_LR,pvalue_LRgamma=pvalue_LR0,Prob=Prob)  }

### Example 1 ###

x1 <- rnorm(50,1,1)
x2 <- rnorm(50)
x3 <- rnorm(50,2,1)
x4 <- rnorm(50)
X <- cbind(x1,x2,x3,x4)
SE(X,T=1.645)$Prob

### example 2 ###

SP_AC_EP <- function(N=100,K=100,alpha=0.05) {

### 1.obtain K intervals

Prob2 <- list()
for (i in 1:K) {

### Specify samples

DATA <- cbind(rnorm(N,0.5,1),rnorm(N,1,1),rnorm(N),rnorm(N))
Prob2[[i]] <- SE(DATA)$Prob
}

SP <- matrix(rep(0,4*K),K); AC <- SP; EP <- SP
for (i in 1:K) SP[i,] <- Prob2[[i]][1,]
for (i in 1:K) AC[i,] <- Prob2[[i]][2,]
for (i in 1:K) EP[i,] <- Prob2[[i]][3,]

  Min <- -0.01 + min(rbind(SP[, 3], AC[, 3]))
  Max <- 0.03 + max(rbind(SP[, 4], SP[, 4]))
  par(mfcol = c(3, 2), oma = c(2, 1, 1, 3), mar = c(2, 4, 1, 1))

### 2. plot confidence intervals

plotCI(SP[,3], SP[,4], alpha, ylim = c(Min, Max), ylab = "SP Intervals")
plotCI(AC[,3], AC[,4], alpha, ylim = c(Min, Max), ylab = "AC Intervals")
plotCI(EP[,3], EP[,4], alpha, ylim = c(Min, Max), ylab = "EP Intervals")

### 3. plot histograms

Max_His <- max(4*AC[,2])
hist(4*SP[, 2], breaks = 10, main = "", xlim = c(0,Max_His))
hist(4*AC[, 2], breaks = 10, main = "", xlim = c(0,Max_His))
hist(4*EP[,2], breaks = 10, main = "", xlim = c(0,Max_His))

```

```

### 4.calculate coverage and mean width

MC <- function(SP) {
  SP_M <- mean(SP[,4] - SP[,3])
  SP_C <- sum((SP[,3]< alpha)&(SP[,4]>alpha))*100/nrow(SP)
  c(SP_M, SP_C)
}

lapply(list(SP,AC,EP), MC)

}

```

SIMULATION.FUSION *CI Estimation of Threshold Probability with Repeated FUSION*

Description

The function SIMULATION.FUSION estimates a family of confidence intervals for density ratio models and Agresti Coull methods

Usage

```
SIMULATION.FUSION(dist='normal',N=NULL,a=NULL,b=NULL,K=100,alpha=0.05,W=c(0.4,0.4))
```

Arguments

dist	specify the distribution which is used to simulate the data. it has arguments:normal(a,b),gamma(a,b),k exponential(a),uniform(a,b),lognorm(a,b)
K	times of simulations
N	number of observatons
alpha	probability
W	a vector which is the relative weight of SP vs AC

Details

This function plots a family of confidence intervals according to the specified distributon using density ratio models,a family of Agresti Coull CIs and their hybrid CIs.

Value

SP	Semiparametric CI
AC	Agresti Coull CI
HB	Hybrid CI according to the weight of SP and AC

Note

Please see Fusion method

Author(s)

Wen Zhou

See Also

FUSION.CI,FUSION.CI.PLOT,plotCI,AC

Examples

```
### Code ###

#####
### SIMULATION.FUSION
#####

SIMULATION.FUSION <- function(dist='normal',N=NULL,a=NULL,b=NULL,K=100,
alpha=0.05,W=c(0.4,0.4)) {
DATA <- FUSION.CI(dist,N,a,b,K,alpha)
FUSION.CI.PLOT(DATA,W,alpha) }

#####
###      Function: FUSION.CI
###      Wen Zhou      5/16/2012
#####

FUSION.CI <- function(dist='normal',N=NULL,a=NULL,b=NULL,K=100,alpha=0.05) {

  ### Semiparametric (SP) Interval Calculation
  SP_vec <- matrix(numeric(4*K),K)
  AC_vec <- matrix(numeric(4*K),K)

  for(k in 1:K){

  Temp <- switch(dist,
    normal=list(x2=rnorm(N,a,b),T=qnorm(1-alpha,a,b)),
    gamma=list(x2=rgamma(N,a,b),T=qgamma(1-alpha,a,b)),
    log=list(x2=rlogis(N,a,b),T=qlogis(1-alpha,a,b)),
    exp=list(x2=rexp(N,a),T=1-qexp(1-alpha,a)),
    uniform=list(x2=runif(N,a,b), T=qunif(1-alpha,a,b)),
    lnorm=list(x2=rlnorm(N,a,b), T= qlnorm(1-alpha,a,b)))

  x2 <- Temp$x2; T <- Temp$T

  L <- 0; U <- 0
  P <- 0; meanP <- 0; index <- 0
  for(j in 1:50){
    for(i in 1:49){
      x1 <- runif(N,-2-i,3+i)
      P[i] <- SP2(x1,x2,T)$Upper_G ### SP2X(x1,x2,T)
```

```

    }
    meanP[j] <- mean(P)
    L[j] <- sort(P)[1]
    U[j] <- sort(P)[48]
  }
  SP_vec[k,] <- c(mean(meanP),mean(L),mean(U),mean(U-L))

  ### Agresti Coull (AC) Interval Calculation
  AC_vec[k,] <- AC(x2,T)
}

list(SP=SP_vec, AC=AC_vec)
}
#####
### SAMPLE.FUSION.CI
#####

SAMPLE.FUSION.CI <- function(x2,K,T=1.645) {

  ### Semiparametric (SP) Interval Calculation
  SP_vec <- matrix(numeric(4*K),K)
  AC_vec <- matrix(numeric(4*K),K)

  L <- 0; U <- 0; P <- 0; meanP <- 0
  for(j in 1:K){
    for(i in 1:49){
      x1 <- runif(length(x2),-2-i,3+i)
      P[i] <- SP2(x1,x2,T)$Upper_G ### SP2:simplified SP2K
    }
    meanP[j] <- mean(P)
    L[j] <- sort(P)[1]
    U[j] <- sort(P)[48]

    AC_vec[j,] <- AC(x2,T)
  }
  SP_vec <- cbind(meanP,L,U,U-L)

  list(SP=SP_vec, AC=AC_vec)
}

### Agresti Coull (AC) Interval Calculation
AC <- function(x2,T) {
  X <- sum(x2 > T)
  n <- length(x2)
  z <- 1.96
  nT <- n+z^2
  pT <- (X+z^2/2)/nT
  meanAC <- pT
  LAC <-ifelse(pT<1, pT - z*sqrt(pT*(1-pT)/nT),pT) #L
  UAC <-ifelse(pT<1, pT + z*sqrt(pT*(1-pT)/nT),pT) #U

  return( c(pT,LAC,UAC,UAC-LAC))
}

#####
### SAMPLE.FUSION

```

```
#####

SAMPLE.FUSION <- function(x2,K=100,W=c(0.4,0.4),T) {
  DATA <- SAMPLE.FUSION.CI(x2,K,T)
  alpha <- sum(x2 > T)/length(x2)
  FUSION.CI.PLOT(DATA,W,alpha)
}

#####
### FUSION.CI.PLOT
#####

### DATA <- FUSION.CI output

FUSION.CI.PLOT <- function(DATA,W=c(0.4,0.4),alpha=0.05) {

  SP <- DATA$SP
  AC <- DATA$AC

  #### plot CI and Hist

  HBL <- W[1]*SP[,2] + (1-W[1])*AC[,2]
  HBU <- W[2]*SP[,3] + (1-W[2])*AC[,3]

  ### Average width for SP,AC,HB

  Width <- c(mean(SP[,4]),mean(AC[,4]),
             mean(HBU-HBL))

  ### Mean for SP,AC, HB

  Mean <- c(mean(SP[,1]),mean(AC[,1]),
            mean(cbind(SP[,1],AC[,1])))

  ### coverage for SP, AC,HB.

  HBL_C <- sum((HBL<alpha)&(HBU>alpha))*100/length(HBL)
  SP_C <- sum((SP[,2]<alpha)&(SP[,3]>alpha))*100/length(HBL)
  AC_C <- sum((AC[,2]<alpha)&(AC[,3]>alpha))*100/length(HBL)

  Coverage <- c(SP_C,AC_C,HB_C)

  HHBB <- rbind(Mean, Width,Coverage)
  colnames(HHBB) <- c("SP","AC","HB")
  rownames(HHBB) <- c("Mean","Width","Coverage")
  small=round(HHBB,digits=3)

  Min <- -0.01 + min(rbind(SP[,2],AC[,2]))
  Max <- 0.03 + max(rbind(SP[,3],SP[,3]))

  par(mfcol=c(3,2),oma=c(2,1,1,3),mar=c(2,4,1,1))
  plotCI(SP[,2],SP[,3],alpha,ylim=c(Min,Max),ylab="SP Intervals")
  plotCI(AC[,2],AC[,3],alpha,ylim=c(Min,Max),ylab="AC Intervals")
  plotCI(HBL,HBU,alpha,ylim=c(Min,Max),ylab="Hybrid Intervals")

  ### Histogram Plot
```

```

Max_His <- max(AC[,3])
hist(SP[,4],prob=TRUE,breaks=10,main="",xlim=c(0,Max_His))
hist(AC[,4],prob=TRUE,breaks=10,main="",xlim=c(0,Max_His))
hist(HBU-HBL,prob=TRUE,breaks=10,main="",xlim=c(0,Max_His))

return(small)
}

#####
#### FUSION.CI.CONTOUR output contour and table
#### a, b are vectors. alpha confidence level.
####
####           Wen Zhou           8/24/2012
#####

FUSION.CI.CONTOUR <- function(DATA,a=seq(0,1,0.1),b=seq(0,1,0.1),alpha=0.05) {

### internal function 1 hybrid(X)

hybrid <- function(W) {
HBL <- W[1]*DATA$SP[,2] + (1-W[1])*DATA$AC[,2]
HBU <- W[2]*DATA$SP[,3] + (1-W[2])*DATA$AC[,3]
HB <- sum((HBL<alpha)&(HBU>alpha))*100/length(HBL)
return(HB)
}

### internal function 2 Mat(fun,a,b)
### fun should accept size-2 vector

Mat <- function(fun,a,b) {

  z <- expand.grid(a,b)

A <- matrix(apply(z,1,fun),ncol=length(b))

contour(a,b,A)
  return(A)
}

### program:

Mat(hybrid,a,b)

}

#####
### Function plotCI
###           Wen Zhou           8/24/2012
#####
plotCI <- function(L, U,a,ylim=ylim,ylab=ylab) {

plot(L, xlim=c(0,length(L)),
      ylab=ylab, ylim=ylim,type='n')

for (i in 1:length(L)) {

```

```

if (L[i] > a || U[i]<a) {
segments(i, L[i],i,U[i],lwd=2)
}
else{
segments(i, L[i],i,U[i],col="pink")
}
}

abline(h=a, col='purple',lty=2,lwd=2)
list(CIs=cbind(L,U),a=a)
}

### rnorm(50,0,1)

SIMULATION.FUSION(dist='normal',N=50,a=0,b=1,K=10)
#           SP      AC      HB
# Mean      0.064  0.09  0.077
# Width     0.079  0.15  0.122
# Coverage  70.000 90.00 90.000

system.time(SIMULATION.FUSION('normal',N=50,a=0,b=1,K=10))
#  user system elapsed
# 107.018   7.997 116.421

### runif(50,0,50)

SIMULATION.FUSION(dist='uniform',N=50,a=0,b=50,K=5)

```

SP2K

Estimation of CI for the Threshold Probability

Description

The function SP2K estimates the parameters according the density ratio model and test the null hypothesis of equidistribution; and also estimate the kernel density and confidence intervals of threshold with three methods: SP, AC and EP.

Usage

```
SP2K(x1,x2,Increment=0.05,BandWidth=0.5,T)
```

Arguments

x1	vector of sample value,univariate sample only
x2	vector or sample value (reference sample),univariate only
Increment	Increment controls the grid at which est. g, g_1, G, G_1 are evaluated, default value is 0.05
BandWidth	Bandwidth for calculation kernel smoothing density,it controls smoothness of kernel est. of g, g_1 and the default value is 0.5
T	Threshold

Details

This function estimates parameters according to density ratio models, calculates the threshold probability, compares the two samples, plot both G_1, G_2 and g_1, g_2 and compare the kernel densities g_1 and g_2 and their histograms. it tests the equidistribution of g_1 and g_2 and also calculate the CIs for the threshold probability with three methods: SP, AC and EP

Value

par	estimate for parameters, alpha, beta, gamma.
pval_LRgamma	p values for null hypothesis: $\gamma=0$. The test is likelihood test.
pval_LR	p values for null hypothesis: $\alpha=0, \beta=0, \gamma=0$, the test is likelihood test. It can be interpreted as equidistribution test.
Sum_p	summation of probability for every observations of reference sample x2, the value should be 1.
Sum_p1	summation of probability for every observations of reference sample x1, the value should be 1.
ALPHA	value to check the integral relations
S	Asymptotic variance matrix
Prob	Threshold probability and its confidence intervals

Note

This is the program developed by Professor Benjamin Kedem in 2006.

Author(s)

Benjamin Kedem

See Also

TWOSAMPLES.DR, MULTISAMPLES.DR

Examples

```
### Example 1 ###
SP2K(rnorm(100, 1, 1), rnorm(100), Increment=0.05, BandWidth=0.5, T=1.645)
```

TWOSAMPLES.DR

Estimation of the jump probabilities and test the equidistribution for two samples

Description

The function TWOSAMPLES.DR estimates the parameters and tests the equidistribution of two samples according density ratio model. The sample can be either univariate or multivariate.

Usage

```
TWOSAMPLES.DR(x1,x2)
```

Arguments

x1	First sample
x2	Second sample (reference sample)

Details

This function estimates parameters according to density ratio models; tests the equidistribution by the likelihood ratio test.

Value

par	estimate for parameters, alpha,beta,gamma.
Sum_p	summation of probability for every observations of reference sample x2, the value should be 1.
Sum_p1	summation of probability for every observations of reference sample x1, the value should be 1 .
pval	p value for the test: $G1=G2$

Note

This is simplified version of SP2K and MULTISAMPLES.DR.

Author(s)

Wen Zhou

See Also

MULTISAMPLES.DR

Examples

```
### Code ###

#####
### Function: TWOSAMPLES.DR Compute jump probability and test equidistribution.
### Two sample case: x1,x2 are either univariate or multivariate samples.
###                               Wen Zhou           8/21/2012
#####

TWOSAMPLES.DR <- function(x1,x2) {

### Univariate Case ###

if (is.vector(x1) && is.vector(x2)) {

n1 <- length(x1); n2 <- length(x2)
rho <- n1/n2; n <- n1+n2
```

```

t <- c(x1,x2) #Data fusion.

###MINUS log-likelihood
minusloglike <- function(theta) {
sum(log(1+rho*exp(theta[1] + theta[2]*(t) + theta[3]*(t)^2))) -
sum(theta[1]+theta[2]*(x1)+theta[3]*(x1)^2)}
###Maximizing loglikelihood by minimizing MINUS loglikelihood
min.func <- nlm(b, start=c(-0.02,.2,.2),obj = minusloglike)

###Parameter estimates
Theta <- min.func$par
names(Theta)=c("alpha","beta","gamma")

###Reference dist. p=dG and its distortion p1=w*dG
p <- 1/(n2*(1+rho*exp(Theta[1] + Theta[2]*(t) + Theta[3]*(t)^2)))
p1 <- p*exp(Theta[1]+Theta[2]*(t)+Theta[3]*(t)^2)

###The LR Test of Equidistribution (beta,gamma)=(0,0)
LR <- -2*(minusloglike(Theta) - minusloglike(c(0,0,0)))
pval <- 1 - pchisq(LR,2)
list(par=Theta,p=p,sum.p=sum(p),p1=p1,sum.p1=sum(p1),p_value=pval)
}

### Multivariate Case ###

else {
# a. Sample Size
n1=nrow(x1); n2=nrow(x2)
rho1=n1/n2; n=n1+n2
t=rbind(x1,x2)
# b. A: vector X matrix
A <- function(theta, x) {
B <- cbind(rep(1,nrow(x)),x)
colSums(theta*t(B)) }
# c. likelihood function
minusloglike = function (theta){
sum(log(1+rho1*exp(A(theta,t))))-sum(A(theta,x1))
}
# 2. Parameter Estimation and distribution Estimation (p,p1)
init <- rep(0.2,ncol(x2)+1)
min.func = nlm(b, start=init, obj=minusloglike)
# Parameter estimates
Theta=min.func$par
p = 1/(n2*(1+ rho1*exp(A(Theta,t))))
p1=exp(A(Theta,t))*p

# 3. Test H_0: G1=G2
LR <- -2*(minusloglike(Theta) - minusloglike(numeric(length(Theta))))
pval <- 1 - pchisq(LR,2)
list(par=Theta,p=p,sum.p=sum(p),p1=p1,sum.p1=sum(p1),p_value=pval)
}

}

### EXAMPLE 1 (Univariate) #####

```

```
TWOSAMPLES.DR(rnorm(100,1,1),rnorm(100))

### EXAMPLE 2 (Multivariate) #####
x1 <- cbind(rnorm(80,1,1),rnorm(80,0.5,1))
x2 <- cbind(rnorm(100,0,2),rnorm(100))
TWOSAMPLES.DR(x1,x2)
```

Index

*Topic **Density Ratio**

CONDITION, 1
CONTOUR, 2
SE, 6
SIMULATION.FUSION, 14
TWO SAMPLES.DR, 20

*Topic **Likelihood Test**

CONDITION, 1
CONTOUR, 2
SE, 6
SIMULATION.FUSION, 14
TWO SAMPLES.DR, 20

*Topic **package**

DATA DESCRIPTION, 5

CONDITION, 1
CONTOUR, 2

DATA (DATA DESCRIPTION), 5
DATA DESCRIPTION, 5

SE, 6
SIMULATION.FUSION, 14
SP2K, 19

TWO SAMPLES.DR, 20

Bibliography

- A. Agresti and B. A. Coull. Approximate is better than “exact” for interval estimation of binomial proportions. *American Statistician*, 52:119–126, 1998.
- J. A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59:19–35, 1972.
- L. D. Brown, T. T. Cai, and A. DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133, 2001.
- L. D. Brown, T. T. Cai, and A. DasGupta. Confidence intervals for a binomial proportion and asymptotic expansions. *Annals of Statistics*, 30(1):160–201, 2002.
- C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413, 1934.
- D. R. Cox. Some sampling problems in technology. *New Developments in Survey Sampling (N. L. Johnson and H. Smith. Jr., eds)*, Wiley, New York, pages 506–517, 1969.
- B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 1:1–26, 1979.
- K. Fokianos. Merging information for semiparametric density estimation. *Journal of the Royal Statistical Society, Series B*, 66:941–958, 2004.
- K. Fokianos and J. Qin. A note on monte carlo maximization by the density ratio model. *Journal of Statistical Theory and Practice*, 2:355–367, 2008.
- K. Fokianos, B. Kedem, J. Qin, and D. Short. A semiparametric approach to the one-way layout. *Technometrics*, 43:56–65, 2001.
- P. B. Gilbert. Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Annals of Statistics*, 28:151–194, 2000.
- P. B. Gilbert. Goodness-of-fit tests for semi parametric biased sampling models. *Journal of the Statistical Planning and Inference*, 118:51–81, 2004.
- R. D. Gill, Y. Vardi, and J. A. Wellner. Large sample theory of empirical distribution in biased sampling models. *Annals of Statistics*, 16(3):1069–1112, 1988.
- G. Guillot, M. Olsson, M. Benson, and M. Rudemo. Discrimination and scoring using small sets of genes for two-sample microarray data. *Mathematical Biosciences*, 5(2):195–203, 2007.
- B. Kedem, G. Lu, R. Wei, and P. D. Williams. Forecasting mortality rates via density ratio modeling. *Canadian Journal of Statistics*, 36:193–206, 2008.
- B. Kedem, E-Y. Kim, A. Voulgaraki, and B. I. Graubard. Two-dimensional semiparametric density ratio modeling of testicular germ cell data. *Statistics in Medicine*, 28:2147–2159, 2009.
- P. S. Laplace. *Theorie analytique des probabilites*. Paris, Ve. Courcier, 1812.

- G. Lu. *Asymptotic Theory for Multiple-Sample Semiparametric Density Ratio Models and its Application to Mortality Forecasting*. Ph.D. Dissertation, University of Maryland, College Park., 2007.
- N. Metropolis and S. Ulam. The monte carlo method. *Journal of American Statistical Association*, 44:335–341, 1949.
- R. L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411, 1979.
- J. Qin. Empirical likelihood in biased sample problem. *Annals of Statistics*, 21:1182–1196, 1993.
- J. Qin. Inferences for case-control data and semi parametric two-sample density ratio models. *Biometrika*, 85:619–630, 1998.
- J. Qin. Empirical likelihood ratio based confidence intervals for mixture proportions. *Annals of Statistics*, 21:1368–1384, 1999.
- J. Qin and J. F. Lawless. Empirical likelihood and general estimating equations. *Annals of Statistics*, 22:300–325, 1994.
- J. Qin and B. Zhang. A goodness of fit test for logistic regression models based on case-control data. *Biometrika*, 84:609–618, 1997.
- Y. Vardi. Nonparametric estimation in the presence of length bias. *Annals of Statistics*, 10:616–620, 1982.
- Y. Vardi. Empirical distribution in selection bias models. *Annals of Statistics*, 13:178–203, 1985.
- A. Voulgaraki, B. Kedem, and B. I. Graubard. Semiparametric regression in testicular germ cell data. *Annals of Applied Statistics*, 6(3):1185–1208, 2012.
- E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal American Statistical Association*, 22:209–212, 1927.
- B. Zhang. A goodness of fit test for multiplicative-intercept risk model based on case-control data. *Statistica Sinica*, 10:839–865, 2000.