

ABSTRACT

Title of dissertation: TWO-DIMENSIONAL SEMIPARAMETRIC
DENSITY RATIO MODELINGS
AND ITS APPLICATIONS

Eun young Kim, Master of Arts, 2008

Dissertation directed by: Professor Benjamin Kedem
Department of Statistics

Bivariate semiparametric inference based on a two-dimensional density ratio model is discussed and applied in testing the significance of risk factors regarding testicular germ cell tumors (TGCT). The results from the joint analysis of height and weight data from a case-control study show that jointly these two factors are significant for TGCT. In addition, joint distributions of genes for diffuse large B-cell lymphoma (DLBCL) are provided.

TWO-DIMENSIONAL SEMIPARAMETRIC DENSITY RATIO
MODELINGS AND ITS APPLICATIONS

by

Eun young Kim

Thesis 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
Master of Arts
2008

Advisory Committee:
Professor Benjamin Kedem, Chair/Advisor
Professor Chris Laskowski
Professor Paul Smith

© Copyright by
Eun young Kim
2008

Table of Contents

List of Tables	iii
List of Figures	iv
1 Introduction	1
2 Statistical Formulation	3
2.1 A Two-Dimensional Density Ratio Model	3
2.2 Connection With Logistic Regression	4
2.3 Estimation	5
2.4 Hypothesis Testing	7
2.5 Simulation Results	8
2.6 A Generalization	9
3 Application to Testicular Germ Cell Cancer	14
4 Application to Diffuse Large B-cell Lymphoma	21
4.1 Joint distribution of gene (9, 112), (9, 148), and (100, 112)	24
5 Summary	34
Bibliography	35

List of Tables

2.1	True parameters and their estimates obtained as averages from 1000 runs. The figures within the parentheses are the corresponding standard deviations. “Uniform” refers to the uniform distribution on the unit disk	13
3.1	Joint probabilities of height (H) and weight (W) of case and control groups.	20
4.1	Estimation and hypothesis testing results for some pairs. The figures within the parentheses are the corresponding standard deviations. . .	23

List of Figures

2.1	True (right) and estimated (left) g_1	10
2.2	True (right) and estimated (left) g_2	11
2.3	Several vertical cuts from the plots in Figures 2.1 and 2.2. The dashed (continuous) line depicts the estimated (true) densities.	12
3.1	Plots of estimated densities of case (g_1) and control (g_2).	17
3.2	Height distribution: Overlay of vertical slice cuts from Figure 3.1. Weight (y value) was fixed at 52.555 (top left), 77.555 (top right), 107.555 (bottom left), and 117.555 (bottom right) kg. The dashed (continuous) line depicts the estimated g_1 (g_2).	18
3.3	Weight distribution: Overlay of vertical slice cuts from Figure 3.1. Height (x value) was fixed at 161.4 (top left), 171.4 (top right), 191.4 (bottom left), and 196.4 (bottom right) cm. The dashed (continuous) line depicts the estimated g_1 (g_2).	19
4.1	Plots of estimated densities of the GC148 gene 9 and 112 (g_1) and the AC148 gene 9 and 112 (g_2).	25
4.2	Gene 9 and 112: Overlay of vertical slice cuts from Figure 4.1. Gene 112 (y value) was fixed at -1.79 (top left), -0.79 (top right), 0.46 (bottom left), and 0.96 (bottom right). The dashed (continuous) line depicts the estimated g_1 (g_2).	26
4.3	Gene 9 and 112: Overlay of vertical slice cuts from Figure 4.1. Gene 9 (x value) was fixed at -0.84 (top left), -0.34 (top right), 0.16 (bottom left), and 0.66 (bottom right). The dashed (continuous) line depicts the estimated g_1 (g_2).	27
4.4	Plots of estimated densities of the GC148 gene 9 and 148 (g_1) and the AC148 gene 9 and 148 (g_2).	28
4.5	Gene 9 and 148: Overlay of vertical slice cuts from Figure 4.4. Gene 148 (y value) was fixed at -1.41 (top left), -0.81 (top right), -0.21 (bottom left), and 0.39 (bottom right). The dashed (continuous) line depicts the estimated g_1 (g_2).	29

4.6	Gene 9 and 148: Overlay of vertical slice cuts from Figure 4.4. Gene 9 (x value) was fixed at -0.84 (top left), -0.34 (top right), 0.16 (bottom left), and 0.66 (bottom right). The dashed (continuous) line depicts the estimated g_1 (g_2).	30
4.7	Plots of estimated densities of the GC148 gene 100 and 112 (g_1) and the AC148 gene 100 and 112 (g_2).	31
4.8	Gene 100 and 112: Overlay of vertical slice cuts from Figure 4.7. Gene 112 (y value) was fixed at -1.79 (top left), -0.29 (top right), 1.21 (bottom left), and 1.71 (bottom right). The dashed (continuous) line depicts the estimated g_1 (g_2).	32
4.9	Gene 100 and 112: Overlay of vertical slice cuts from Figure 4.7. Gene 100 (x value) was fixed at -0.14 (top left), 0.96 (top right), 0.86 (bottom left), and 1.36 (bottom right). The dashed (continuous) line depicts the estimated g_1 (g_2).	33

Chapter 1

Introduction

Consider the $m = q + 1$ random samples, $(x_{11}, \dots, x_{1n_1}), \dots, (x_{q1}, \dots, x_{qn_q}), (x_{m1}, \dots, x_{mn_m})$, with probability density functions g_i

$$x_{ij} \sim g_i, \quad i = 1, \dots, q, m, \quad j = 1, \dots, n_i, \quad (1.1)$$

where $g_m \equiv g$ is called the *reference* probability density, and where the g_i satisfy the *density ratio model*

$$\frac{g_j(x)}{g(x)} = \exp(\alpha_j + \beta_j' \mathbf{h}(x)), \quad j = 1, \dots, q. \quad (1.2)$$

The distortion function $\mathbf{h}(x)$ is assumed a known vector-valued function whose choice depends on the data. For example $h(x) = (x, \log x)'$ when the data are reminiscent of the gamma distribution. None of the densities is known, and neither are the parameters α_j, β_j . As such, this is a semiparametric model, and the statistical objective is to estimate the reference density g and all the parameters from the combined data

$$\mathbf{t} = \{(x_{11}, \dots, x_{1n_1}), \dots, (x_{q1}, \dots, x_{qn_q}), (x_{m1}, \dots, x_{mn_m})\}'. \quad (1.3)$$

The density ratio model has been studied and successfully applied by many authors including Prentice and Pyke (1979) (case-control studies), Qin and Zhang (1997) (logistic model validation), Qin (1998) (case-control studies), Gilbert et al

(1999) (AIDS vaccine trials), Zhang (2000) (goodness of fit), Fokianos et al (2001) (analysis of variance), Fokianos (2004) (kernel density estimation), Phue et al (2006) (microarrays evaluation), Kedem and Wen (2007) (cluster detection), Kedem et al (2008) (mortality rate forecasting). The picture which emerges from all this and related work is that, under the density ratio model, by combining all the samples we get both better estimates and more powerful tests. For example, the reference g is estimated from all the $n = n_1 + n_2 + \dots + n_m$ observations, that is *the combined data*, and not just from the m th sample, thus yielding a more efficient estimate. This increase in efficiency has been studied rigorously in Gilbert (2000) and Fokianos (2004). Gagnon (2005) showed that in the case of $m = 2, q = 1$, the semiparametric test (2.14) in Section 2.4 for testing $H_0 : \beta_1 = 0$ is a useful semiparametric alternative to the t -test, and that in some cases it is more efficient. Wen (2007), and Kedem and Wen (2007), demonstrated that in cluster detection the likelihood ratio test obtained under (1.2) competes well with specialized tests obtained under the assumptions of specific distributions.

In this thesis we extend the one-dimensional case-control density ratio model to a two-dimensional model, and apply the latter to detect a distributional difference of testicular germ cell tumor data, and diffuse large B-cell lymphoma gene data

Our approach is semiparametric. It depends on a density ratio model studied and applied mostly in its one-dimensional version. This necessitates first a certain bivariate extension discussed in Section 2.

Chapter 2

Statistical Formulation

A clue for generalizing (1.2) can be obtained from the ratio of two bivariate normal densities with different means but the same covariance matrices.

2.1 A Two-Dimensional Density Ratio Model

Corresponding to (1.1), suppose we have $m = q + 1$ two-dimensional data sets,

$$\begin{aligned}(x_{11}, y_{11}), (x_{12}, y_{12}), \dots, (x_{1n_1}, y_{1n_1}) &\sim g_1(x, y) \\(x_{21}, y_{21}), (x_{22}, y_{22}), \dots, (x_{2n_2}, y_{2n_2}) &\sim g_2(x, y) \\&\cdot \\&\cdot \\&\cdot \\(x_{q1}, y_{q1}), (x_{q2}, y_{q2}), \dots, (x_{qn_q}, y_{qn_q}) &\sim g_q(x, y) \\(x_{m1}, y_{m1}), (x_{m2}, y_{m2}), \dots, (x_{mn_m}, y_{mn_m}) &\sim g_m(x, y)\end{aligned}$$

where $g_j(x, y)$ is the probability density of $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, with

$$\boldsymbol{\mu}_j = \begin{pmatrix} \mu_{jx} \\ \mu_{jy} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}, \quad j = 1, \dots, m.$$

Then, choosing $g_m(x, y)$ as a reference density we have (Anderson 1971),

$$\frac{g_j(x, y)}{g_m(x, y)} = \exp\left[(\boldsymbol{\mu}_j - \boldsymbol{\mu}_m)' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_j' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j - \boldsymbol{\mu}_m' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_m)\right], \quad (2.1)$$

where $\mathbf{x} = (x, y)'$. We see that (2.1) is a special case of the general form

$$\frac{g_j(x, y)}{g_m(x, y)} = \exp(\alpha_j + \boldsymbol{\beta}'_j \mathbf{x}) \quad (2.2)$$

where

$$\alpha_j = -\frac{1}{2}(\boldsymbol{\mu}'_j \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j - \boldsymbol{\mu}'_m \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_m)$$

$$\boldsymbol{\beta}_j = \begin{pmatrix} \beta_{j1} \\ \beta_{j2} \end{pmatrix} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_m)$$

Observe that $\boldsymbol{\beta}_j = \mathbf{0}$ implies $\alpha_j = 0$, $j = 1, \dots, q$. It follows that the hypothesis $H_0 : \boldsymbol{\mu}_1 = \dots = \boldsymbol{\mu}_m$ is equivalent to $H_0 : \boldsymbol{\beta}_1 = \dots = \boldsymbol{\beta}_q = \mathbf{0}$. That is, equidistribution: all the g_i are equal.

In what follows we shall adopt the form (2.2) but drop the normal assumption.

Thus, by the *two dimensional density ratio model* we mean the model

$$\frac{g_j(\mathbf{x})}{g(\mathbf{x})} = \exp(\alpha_j + \boldsymbol{\beta}'_j \mathbf{x}), \quad j = 1, \dots, q \quad (2.3)$$

with reference $g \equiv g_m$, scalar α_j , two-dimensional $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2})'$, and $\mathbf{x} = (x, y)'$. Other than the “tilt” $\exp(\alpha_j + \boldsymbol{\beta}'_j \mathbf{x})$, no other distributional assumptions are made.

2.2 Connection With Logistic Regression

A special case of (2.3) is obtained within the case-control framework studied in Prentice and Pyke (1979). Let $D = i$ denote the i th disease incidence, $i = 1, \dots, q$, and let $D = m$ indicate disease-free state. The probabilities $\pi_i = P(D = i)$ satisfy $\sum_{i=1}^m \pi_i = 1$. Let $P(D = i | \mathbf{x})$ denote the conditional probability that an individual with covariate vector \mathbf{x} has disease $D = i$, where $\mathbf{x} \sim f(\mathbf{x})$.

Define $g_i(\mathbf{x}) = P(\mathbf{x} \mid D = i)$, $i = 1, \dots, q, m$, and assume the generalized logistic regression model,

$$P(D = i|\mathbf{x}) = \frac{\exp(\alpha_i^* + \boldsymbol{\beta}'_i \mathbf{x})}{1 + \sum_{i=1}^q \exp(\alpha_i^* + \boldsymbol{\beta}'_i \mathbf{x})}, \quad i = 1, \dots, q, m \quad (2.4)$$

where $\alpha_m^* = 0$ and $\boldsymbol{\beta}_m = \mathbf{0}$. Then, from Bayes' Theorem,

$$\frac{g_i(\mathbf{x})}{g_m(\mathbf{x})} = \exp(\alpha_i + \boldsymbol{\beta}'_i \mathbf{x})$$

where $\alpha_i = \alpha_i^* + \log(\pi_m/\pi_i)$. Holding $g \equiv g_m$ as a reference density we obtain (2.3).

The connection with logistic regression points to the fact that the $\boldsymbol{\beta}_i$ can be estimated from a model such as (2.4), which provides only part of the puzzle regarding the infinite dimensional parameters $g_i(\mathbf{x})$. On the other hand, the density ratio model (2.3) allows direct semiparametric inference about both the $\boldsymbol{\beta}_i$ and the multivariate $g_i(\mathbf{x})$. Moreover, (2.3) plays on the intuitively appealing notion of a baseline behavior and deviation from it.

2.3 Estimation

To estimate the parameters and the reference g , or equivalently the reference distribution function G , the data are first combined in a single matrix of dimension $n \times 2$ where $n = n_1 + n_2 + \dots + n_m$. For $i = 1, \dots, q, m$, and $j = 1, \dots, n_i$, let $\mathbf{x}_{ij} = (x, y)'_{ij}$, and denote by \mathbf{t} the combined data,

$$\begin{aligned} \mathbf{t} &= (\mathbf{x}'_{11}, \dots, \mathbf{x}'_{1n_1}, \mathbf{x}'_{21}, \dots, \mathbf{x}'_{2n_2}, \dots, \mathbf{x}'_{q1}, \dots, \mathbf{x}'_{qn_q}, \mathbf{x}'_{m1}, \dots, \mathbf{x}'_{mn_m})' \\ &= (\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n)'. \end{aligned} \quad (2.5)$$

It is convenient to define $\mathbf{t}_i = (t_{ix}, t_{iy})'$, and also to switch between the \mathbf{t}_i and \mathbf{x}_{kl} as needed.

We shall follow Qin (1998) and Qin and Zhang (1997) and references therein. To obtain the maximum likelihood estimator of $G(x, y)$, we optimize over the class of two-dimensional step functions with jumps p_i at $\mathbf{t}_1, \dots, \mathbf{t}_n$,

$$p_i = G(t_{ix}, t_{iy}) - G(t_{i-1,x}, t_{iy}) - G(t_{ix}, t_{i-1,y}) + G(t_{i-1,x}, t_{i-1,y}), \quad i = 1, \dots, n.$$

With $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)'$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_q)'$, the likelihood is then given by,

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, G) = \prod_{i=1}^n p_i \prod_{k=1}^{n_1} \exp(\alpha_1 + \beta_{11}x_{1k} + \beta_{12}y_{1k}) \cdots \prod_{k=1}^{n_q} \exp(\alpha_q + \beta_{q1}x_{qk} + \beta_{q2}y_{qk}) \quad (2.6)$$

subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n w_1(\mathbf{t}_i)p_i = 1, \dots, \quad \sum_{i=1}^n w_q(\mathbf{t}_i)p_i = 1 \quad (2.7)$$

where

$$w_j(\mathbf{t}_i) = \exp(\alpha_j + \boldsymbol{\beta}'_j \mathbf{t}_i) = \exp(\alpha_j + \beta_{j1}t_{ix} + \beta_{j2}t_{iy}), \quad j = 1, \dots, q.$$

The relative sample sizes $\rho_j = n_j/n_m$, $j = 1, \dots, q$, play an important role in the ensuing estimation and hypothesis testing as seen from the log-likelihood l ,

$$\begin{aligned} l &\equiv \log L(\boldsymbol{\alpha}, \boldsymbol{\beta}, G) \\ &= - \sum_{i=1}^n \log[1 + \rho_1 w_1(\mathbf{t}_i) + \cdots + \rho_q w_q(\mathbf{t}_i)] + \sum_{j=1}^{n_1} (\alpha_1 + \beta_{11}x_{1j} + \beta_{12}y_{1j}) \\ &\quad + \cdots + \sum_{j=1}^{n_q} (\alpha_q + \beta_{q1}x_{qj} + \beta_{q2}y_{qj}) + \text{Constant}. \end{aligned} \quad (2.8)$$

Maximizing the log-likelihood subject to the constraints (2.7) by the same profiling method as in Qin and Zhang (1997) and Fokianos et al (2001), we first get an

expression for p_i as in (2.11) below, and then get the score equations,

$$\frac{\partial l}{\partial \alpha_j} = - \sum_{i=1}^n \frac{\rho_j w_j(\mathbf{t}_i)}{1 + \rho_1 w_1(\mathbf{t}_i) + \cdots + \rho_q w_q(\mathbf{t}_i)} + n_j = 0 \quad (2.9)$$

$$\frac{\partial l}{\partial \beta_j} = - \sum_{i=1}^n \frac{\rho_j w_j(\mathbf{t}_i) \mathbf{t}_i}{1 + \rho_1 w_1(\mathbf{t}_i) + \cdots + \rho_q w_q(\mathbf{t}_i)} + \sum_{i=1}^{n_j} \begin{pmatrix} x_{ji} \\ y_{ji} \end{pmatrix} = 0 \quad (2.10)$$

The solution of the score equations provides maximum likelihood estimators $\hat{\alpha}_j$ and $\hat{\beta}_j$, $j = 1, \dots, q$. It can be shown that the maximum likelihood estimators are asymptotically normal with variance-covariance matrix which is essentially the same as in Fokianos et al (2001) and Kedem and Wen (2007).

By substitution

$$\hat{p}_i = \frac{1}{n_m} \cdot \frac{1}{1 + \rho_1 \hat{w}_1(\mathbf{t}_i) + \cdots + \rho_q \hat{w}_q(\mathbf{t}_i)} \quad (2.11)$$

$$\hat{G}(\mathbf{t}) = \frac{1}{n_m} \cdot \sum_{i=1}^n \frac{I(\mathbf{t}_i \leq \mathbf{t})}{1 + \rho_1 \hat{w}_1(\mathbf{t}_i) + \cdots + \rho_q \hat{w}_q(\mathbf{t}_i)} \quad (2.12)$$

where $\hat{w}_j(\mathbf{t}_i) = \exp(\hat{\alpha}_j + \hat{\beta}_j' \mathbf{t}_i)$, and $I(B)$ is the indicator of the event B . Asymptotic properties of \hat{G} have been studied by Gilbert (2000), Zhang (2000), and more recently Lu (2007).

2.4 Hypothesis Testing

Several test statistics for testing the equidistribution hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = \mathbf{0}$ have been discussed in Fokianos et al (2001), and in Kedem and Wen (2007), including the likelihood ratio (LR),

$$LR \equiv -2[l(0, 0) - l(\hat{\alpha}, \hat{\beta})]$$

$$\begin{aligned}
&= -2 \sum_{i=1}^n \log[1 + \rho_1 \hat{w}_1(\mathbf{t}_i) + \dots + \rho_q \hat{w}_q(\mathbf{t}_i)] \\
&+ 2 \sum_{i=1}^q \sum_{j=1}^{n_i} [\hat{\alpha}_i + \hat{\beta}_{i1} x_{ij} + \hat{\beta}_{i2} y_{ij}] + 2n \log[1 + \sum_{i=1}^q \rho_i] \quad (2.13)
\end{aligned}$$

Under H_0 , the likelihood ratio is asymptotically approximately distributed as χ^2 with $2q$ degrees of freedom, and H_0 is rejected for large values of LR.

Another useful test connected with the one-dimensional version of (2.3), with scalar β_1 , $q = 1$, and $\rho_1 = n_1/n_2$, is the test based on the \mathcal{X}_1 statistic,

$$\mathcal{X}_1 = n \frac{\rho_1 \hat{V}ar(t) \hat{\beta}_1^2}{(1 + \rho_1)^2} \quad (2.14)$$

where $\hat{V}ar(t)$ is the estimated variance of the reference distribution given by,

$$\sum_{i=1}^n t_i^2 \hat{p}_i - \left(\sum_{i=1}^n t_i \hat{p}_i \right)^2.$$

Under the hypothesis $H_0 : \beta_1 = 0$, \mathcal{X}_1 is asymptotically χ^2 with one degree of freedom. See Fokianos et al (2001) for the derivation of \mathcal{X}_1 , and Gagnon (2005) for its efficiency. In our analysis of the TGCT data we apply both the likelihood ratio and \mathcal{X}_1 .

2.5 Simulation Results

To illustrate the estimation procedure discussed above, consider first the case of $q = 1$, $n_1 = 200$, $n_2 = 100$, reference g from $N((0, 0)', \Sigma)$, and g_1 from $N((1, 1)', \Sigma)$, where

$$\Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}.$$

The true parameter values are $\alpha = -0.3$, $\beta_1 = 0.2$, and $\beta_2 = 0.4$. The estimated parameter values are $\hat{\alpha} = -0.3142789$, $\hat{\beta}_1 = 0.2309552$, $\hat{\beta}_2 = 0.3941714$, with standard deviations 0.085, 0.101, 0.131, respectively. The true and estimated g_1 and g_2 are shown in Figures 2.1 and 2.2, respectively. Two-dimensional kernel density estimation is used for smoothing. Since it is difficult to see differences between the true and estimated densities, several vertical slices or cut traces from Figures 2.1 and 2.2 are provided in Figure 2.3.

Further results from bivariate normal and from the uniform distribution on the unit disc are given in Table 2.1. The estimates are averages from 1000 runs, and they seem to be fairly precise.

2.6 A Generalization

A generalization of (2.2) is obtained when the covariance matrices are not equal and $g_i \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Then

$$\frac{g_j(x, y)}{g_m(x, y)} = \exp(\alpha_j + \boldsymbol{\beta}'_j \mathbf{x}) \quad (2.15)$$

where

$$\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \beta_{j3}, \beta_{j4}, \beta_{j5})' \quad \text{and} \quad \mathbf{x} = (x^2, x, y^2, y, xy)'$$

Since the form of (2.15) is identical to (2.2), the inferential results are the same.

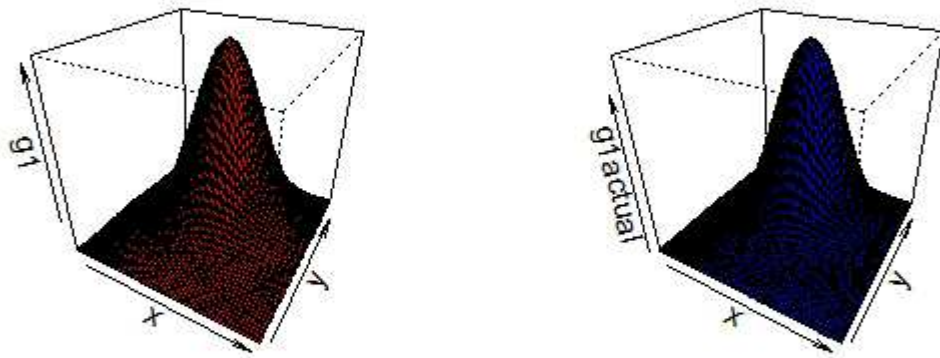


Figure 2.1: True (right) and estimated (left) g_1 .

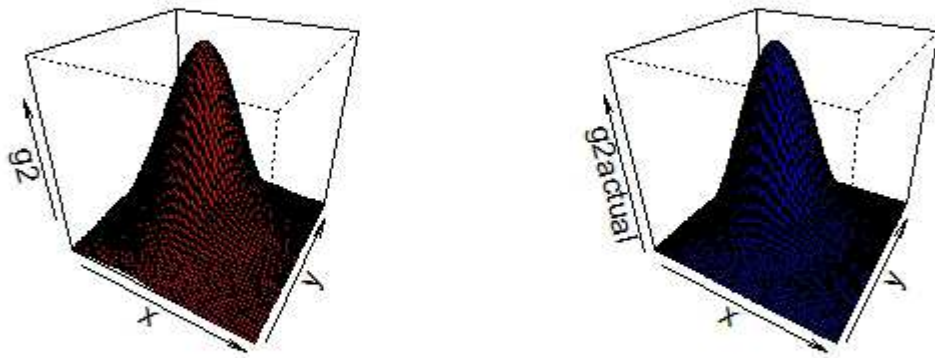


Figure 2.2: True (right) and estimated (left) g_2 .

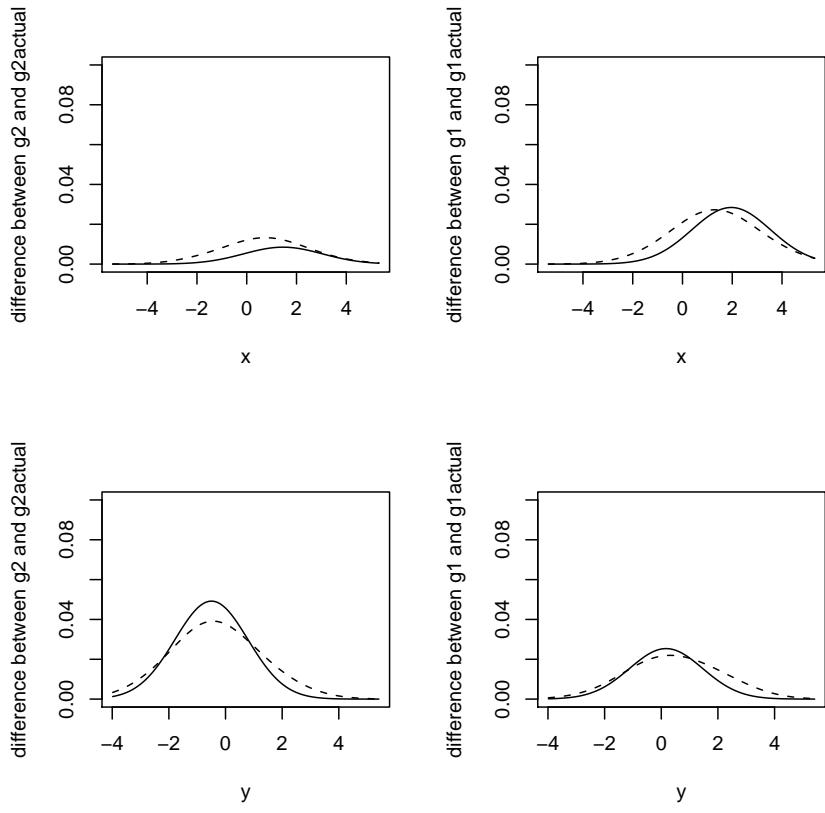


Figure 2.3: Several vertical cuts from the plots in Figures 2.1 and 2.2. The dashed (continuous) line depicts the estimated (true) densities.

Table 2.1: True parameters and their estimates obtained as averages from 1000 runs. The figures within the parentheses are the corresponding standard deviations.

“Uniform” refers to the uniform distribution on the unit disk

Densities	n_1, n_2	$\alpha, \beta_{11}, \beta_{12}$	$\hat{\alpha}, \hat{\beta}_{11}, \hat{\beta}_{12}$
$g_1 \sim N((0, 0)', \Sigma)$	100, 100	0, 0, 0	-0.00004, -0.00117, -0.00117
$g_2 \sim N((0, 0)', \Sigma)$			(.01326), (.08918), (.10968)
	200, 100	0, 0, 0	.00457, .001834, -0.000235
			(.01129), (.07771), (.09682)
	100, 200	0, 0, 0	-0.00542, -0.00206, .00089
			(.01188), (.08157), (.09649)
$g_1 \sim N((1, 1)', \Sigma)$	100, 100	-0.3, 0.2, 0.4	-.30343, .20442, .41204
$g_2 \sim N((0, 0)', \Sigma)$			(.08500), (.10149), (.13139)
	200, 100	-0.3, 0.2, 0.4	-.30055, .20105, .4104467
			(.06266), (.08630), (.11300)
	100, 200	-0.3, 0.2, 0.4	-.30984, .20315, .40365
			(.08502), (.08681), (.10708)
$g_1 \sim \text{uniform}$	100, 100	0, 0, 0	.00375, -.05830, .01340
$g_2 \sim \text{uniform}$			(.15897), (.98302), (.30291)
	200, 100	0, 0, 0	.00475, -.03706, -0.00743
			(.14079), (.97358), (.27732)
	100, 200	0, 0, 0	-0.002497, .09041, .01230
			(.14215), (.97791), (.27832)

Chapter 3

Application to Testicular Germ Cell Cancer

Testicular germ cell tumor (TGCT) is a common cancer among young US men, mainly in the age group 15-35 years. Well known risk factors are cryptorchidism, prior history of TGCT, and family history of TGCT. Men with seminoma, history of TGCT, or family history of TGCT have a higher rate of TGCT. Other possible risk factors include body size, dairy consumption, and age at puberty (McGlynn et al 2003, McGlynn et al 2007.) Using standard logistic regression with categories for body size, it was determined in McGlynn et al (2007) that increased height was significantly related to risk, giving an odds ratio (OR) of 1.83 with 95% confidence interval (CI) of (1.36, 2.45), where this OR is for men with height greater than 182.88 cm compared to those with less than or equal to 172.72 cm. On the other hand body mass index (weight in kilograms divided by height in meters squared) was not significant (OR = 1.06, 95% CI: 0.66, 1.69), where this OR is for men with body mass index greater than or equal to 30 compared to to those with less than 18.5. Moreover, there was no association found for age at puberty, voice changing, and dairy consumption.

We shall further analyze the data in McGlynn et al (2007), applying a likelihood ratio test, obtained from semiparametric considerations, in testing equidistribution between the case and control groups. In doing so, since height and weight

are related, we consider the joint two-dimensional height-weight data rather than height by itself, and estimate the bivariate distribution of height and weight. Such a bivariate analysis may lead to results which cannot be found from marginal considerations alone, or even using logistic regression analyses with both variables modeled together.

The TGCT data referred to in the Introduction consist of pairs of heights and weights of 1691 individuals, of which 928 are cases and 763 are in the control group. The range of height in the case group is from 152.4 to 215.9 cm, and in the control group from 160.02 to 203.2 cm. The range of weight in the case group is from 38.555 to 127.006 kg, and in the control group from 50.802 to 131.542 kg. We assume the density ratio model (2.3) which in the present case is written as,

$$\frac{g_1(x, y)}{g_2(x, y)} = \exp(\alpha_1 + \boldsymbol{\beta}'_1 \mathbf{x}) \quad (3.1)$$

where g_1 is the distribution of the case group, and g_2 is the reference distribution of the control group, and the hypothesis of interest is $H_0 : \boldsymbol{\beta}_1 = 0$. Supporting (3.1) is the the fact that the correlations between height and weight in the two groups are 0.5050266 and 0.5210018 respectively.

Before applying our two-dimensional density ratio model to the TGCT data, it is interesting to see what the one-dimensional analog of (3.1) gives when applied to height and body mass index separately, as was done in McGlynn et al (2007). Accordingly, the hypothesis of equidistribution reduces to testing the vanishing of the scalar β_1 , $H_0 : \beta_1 = 0$, with $n_1 = 928$ and $n_2 = 763$. Regarding height, the likelihood ratio test (2.13) with one degree of freedom gives a p -value of 0.00011,

whereas it is 0.88213 for body mass index. Similarly, the \mathcal{X}_1 test in (2.14) gives the p -values 0.00009 and 0.88208, respectively. Consequently, height is a significant risk factor, but body mass index is not, in agreement with McGlynn et al (2007), who approached the problem very differently using the odds ratio from logistic regression analysis.

Going back to the joint (height,weight) case, if the entire data set is used with $n_1 = 928$ and $n_2 = 763$, then we obtain from the score equations (2.9) and (2.10),

$$(\hat{\alpha}, \hat{\beta}_{11}, \hat{\beta}_{12}) = (4.67597, -0.02523, -0.00198) \quad (3.2)$$

with respective standard errors (1.151588079, 0.007432646, 0.004618770), indicating difference between the two groups. Indeed, the likelihood ratio (2.13) is equal to 15.10806, and with 2 degrees of freedom the corresponding p -value is 0.00052. Thus, when height and weight are considered jointly, we reject the null hypothesis of equidistribution (difference) quite conclusively. This conclusion is illustrated graphically in Figure 3.1, and its one-dimensional manifestations given in Figures 3.2 and 3.3. Joint probabilities of height and weight are provided in Table 3.1.

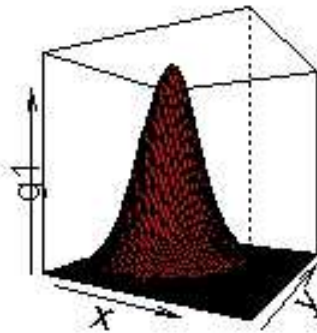
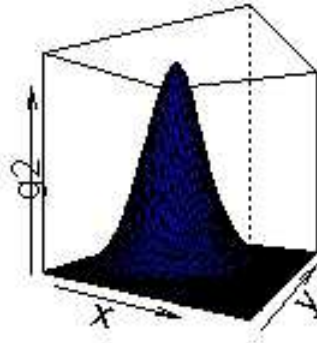


Figure 3.1: Plots of estimated densities of case (g_1) and control (g_2).

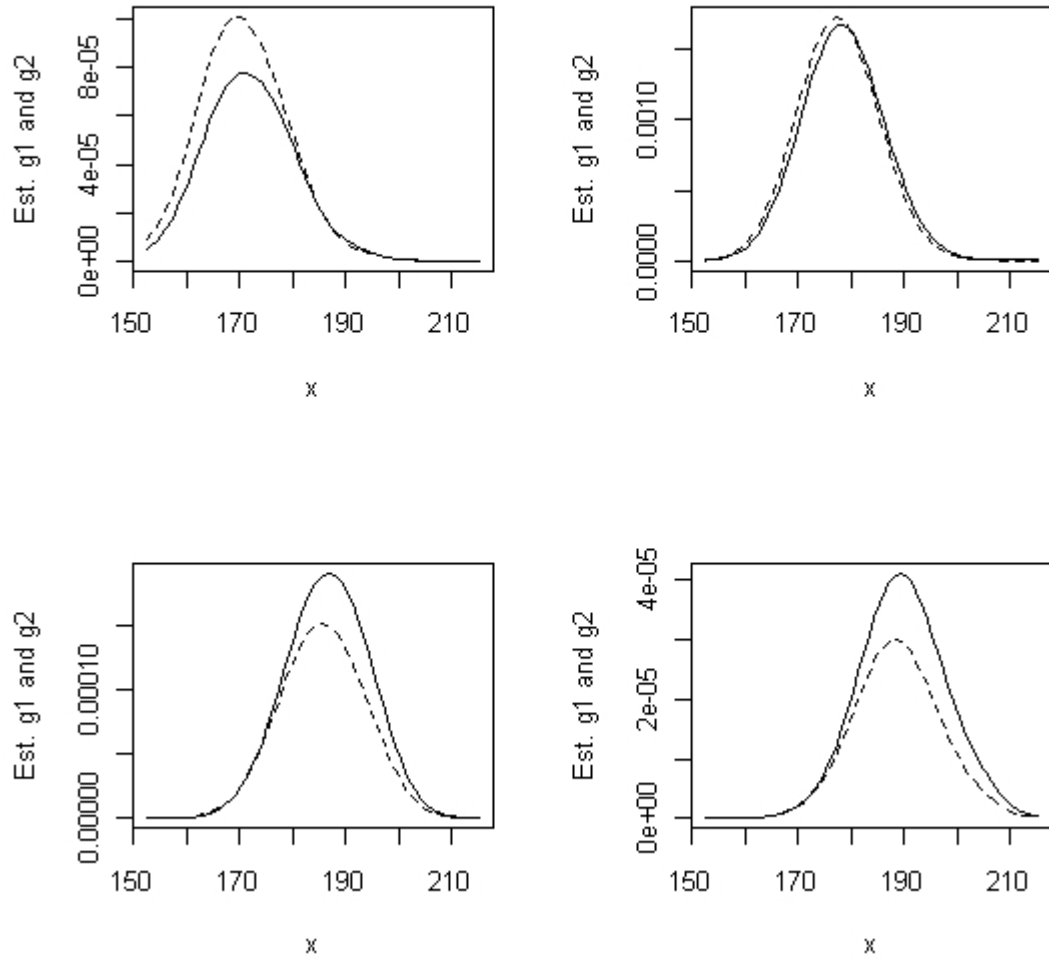


Figure 3.2: Height distribution: Overlay of vertical slice cuts from Figure 3.1. Weight (y value) was fixed at 52.555 (top left), 77.555 (top right), 107.555 (bottom left), and 117.555 (bottom right) kg. The dashed (continuous) line depicts the estimated g_1 (g_2).

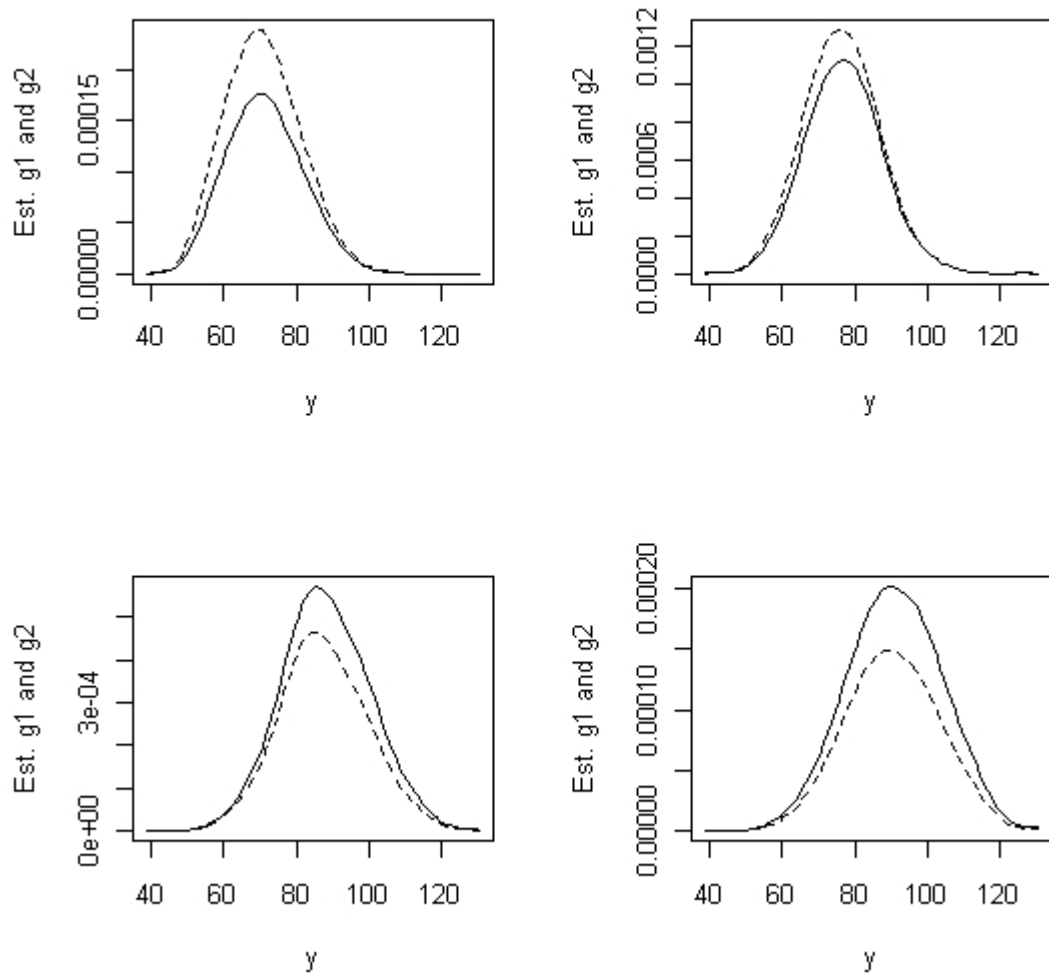


Figure 3.3: Weight distribution: Overlay of vertical slice cuts from Figure 3.1. Height (x value) was fixed at 161.4 (top left), 171.4 (top right), 191.4 (bottom left), and 196.4 (bottom right) cm. The dashed (continuous) line depicts the estimated g_1 (g_2).

Table 3.1: Joint probabilities of height (H) and weight (W) of case and control groups.

Probabilities	Case group	Control group
$\Pr(H \leq 155, W \leq 59)$	0.000769	0.000374
$\Pr(H \leq 165, W \leq 59)$	0.005750	0.003490
$\Pr(H \leq 178, W \leq 65)$	0.066406	0.051604
$\Pr(H \leq 185, W \leq 70)$	0.161664	0.133651
$\Pr(H \leq 180, W \leq 80)$	0.375041	0.315808
$\Pr(H \leq 180, W \leq 90)$	0.520636	0.448033
$\Pr(H \leq 187, W \leq 95)$	0.818147	0.769016
$\Pr(H \leq 200, W \leq 100)$	0.957770	0.943891
$\Pr(H \leq 203, W \leq 119)$	0.996346	0.993959

Chapter 4

Application to Diffuse Large B-cell Lymphoma

Diffuse large B-cell lymphoma (DLBCL), the most common subtype of non-Hodgkin's lymphoma, is clinically heterogeneous. Using DNA microarrays, a systematic characterization of gene expression in B-cell malignancies was conducted. It was shown that there is diversity in gene expression among the tumours of DLBCL patients, apparently reflecting the variation in tumour proliferation rate, host response and differentiation state of the tumour. Two molecularly distinct forms of DLBCL, which had gene expression patterns indicative of different stages of B-cell differentiation, were identified. One type expressed gene characteristic of germinal centre B cells (germinal centre B-like DLBCL); the second type expressed genes normally induced during *in vitro* activation of peripheral blood B cells (activated B-like DLBCL). Patients with germinal centre B-like DLBCL had a significantly better overall survival than those with activated B-like DLBCL. The molecular classification of tumours on the basis of gene expression can identify previously undetected and clinically significant subtypes of cancer. See Alizadeh et al. (2000).

From the original investigations in their study, 148 genes expressed in the germinal centre B cells were found to be the most useful for classification. The 148 genes were divided into two groups: the first containing 24 samples referred to GC and the second containing 23 samples referred to AC. These 148 genes have been num-

bered from 1 to 148 for convenience. For each group, genes 9, 36, 69, 100, 112, and 148 were observed. Since some of the expressions in the samples are missing, the GC and AC data sample sizes are at most 24 and 23, respectively. The data used in this analysis are \log_2 transformed. This transformation essentially stabilizes the variance of the data and produces rough symmetry. See Gagnon et al. (2008).

The one-dimensional semiparametric method was applied to each of the genes 9, 36, 69, 100, 112, and 148 to see how the GC148 data are different from the AC148 data. The results are given in a table in Gagnon(2008). For example, gene 9 is similarly distributed in the GC148 and AC148 with a p-value 0.226, however gene 69 is distributed differently in the GC148 and AC148 with a p-value 0.000. Now we shall apply a two-dimensional semiparametric analysis to the two groups of data GC148 and AC148.

Here we take two genes together to form two dimensional data. For instance, suppose gene 9 (x -values) and gene 36 (y -values) from the GC148 follow $g_1(x, y)$ and gene 9 (x -values) and gene 36 (y -values) from the AC148 follow $g_2(x, y)$. To compare the two-dimensional analysis with the one in Gagnon (2008) we use the AC148 as the reference data. Recall the two-dimensional model is

$$\frac{g_1(x, y)}{g_2(x, y)} = \exp(\alpha_1 + \boldsymbol{\beta}'_1 \mathbf{x}).$$

Since we have 6 genes, there are 15 pairs of gene data. Estimation and hypothesis testing results for some pairs are given in Table 4.1. To test equidistribution, $H_0 : \boldsymbol{\beta}_1 = 0$, we used the likelihood ratio test as we did for the TGCT data. Graphs of estimated g_1 and g_2 are shown on Figure 4.1 to 4.9.

Table 4.1: Estimation and hypothesis testing results for some pairs. The figures within the parentheses are the corresponding standard deviations.

Gene	$\hat{\alpha}, \hat{\beta}_{11}, \hat{\beta}_{12}$	LR	p-value
9, 100	-0.17532, 0.66091, 0.700492 (0.18246), (0.95280), (0.96080)	2.282791	0.31937
9, 112	-0.03742, 0.49971, 1.66484 (0.21212), (1.12509), (1.08909)	11.62944	0.00298
9, 148	-0.03387, 0.86908, 0.08002 (0.18876),(0.84483), (0.51725)	1.486727	0.47551
100, 112	-0.36744, 0.67480, 1.65708 (0.40977),(1.05499), (0.55627)	16.43722	0.00026
100, 148	-0.14261, 0.61199, -0.07898 (0.27576), (1.19593), (0.51192)	0.93505	0.62655
112, 148	-0.35921, 1.76323, -0.33106 (0.33682), (0.70348), (0.65036)	16.23034	0.00029

4.1 Joint distribution of gene (9, 112), (9, 148), and (100, 112)

In Gagnon (2008), the equidistribution hypotheses regarding gene 9 and gene 112 give p-values 0.226 and 0.000, respectively, which indicates that gene 9 of the GC148 and AC148 are similarly distributed, but gene 112 is not. When gene 9 and 112 are jointly treated, the corresponding p-value is 0.002, as given in the Table 4.1. Thus we reject the null hypothesis. Gene 148 gives a p-value of 0.821 when it was individually treated. When gene 9 and 148 are jointly analyzed the p-value is 0.475. Thus we accept the null hypothesis. Gene 100 has a p-value .265 in Gagnon, but when it was jointly analyzed with gene 112, we reject the null hypothesis with the p-value 0.000.

From gene pairs as given in Table 4.1, the two-dimensional density ratio modeling accords with the one-dimensional approach. If a gene was significantly different, when it is jointly analyzed with another gene, regardless of its significance, they are still differently distributed.

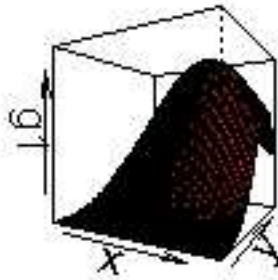
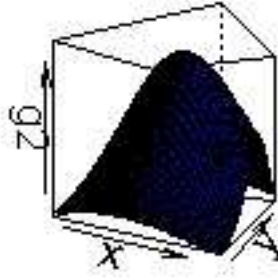


Figure 4.1: Plots of estimated densities of the GC148 gene 9 and 112 (g_1) and the AC148 gene 9 and 112 (g_2).

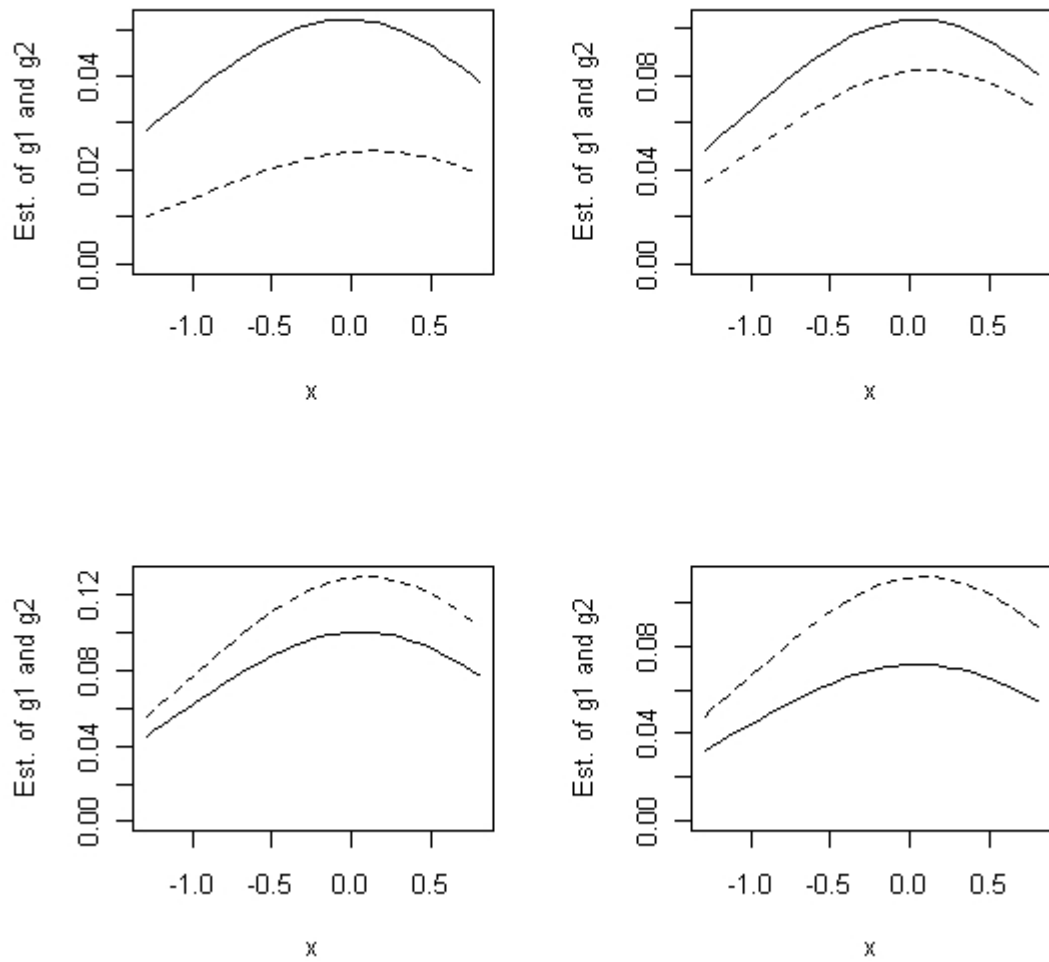


Figure 4.2: Gene 9 and 112: Overlay of vertical slice cuts from Figure 4.1. Gene 112 (y value) was fixed at -1.79 (top left), -0.79 (top right), 0.46 (bottom left), and 0.96 (bottom right). The dashed (continuous) line depicts the estimated g_1 (g_2).

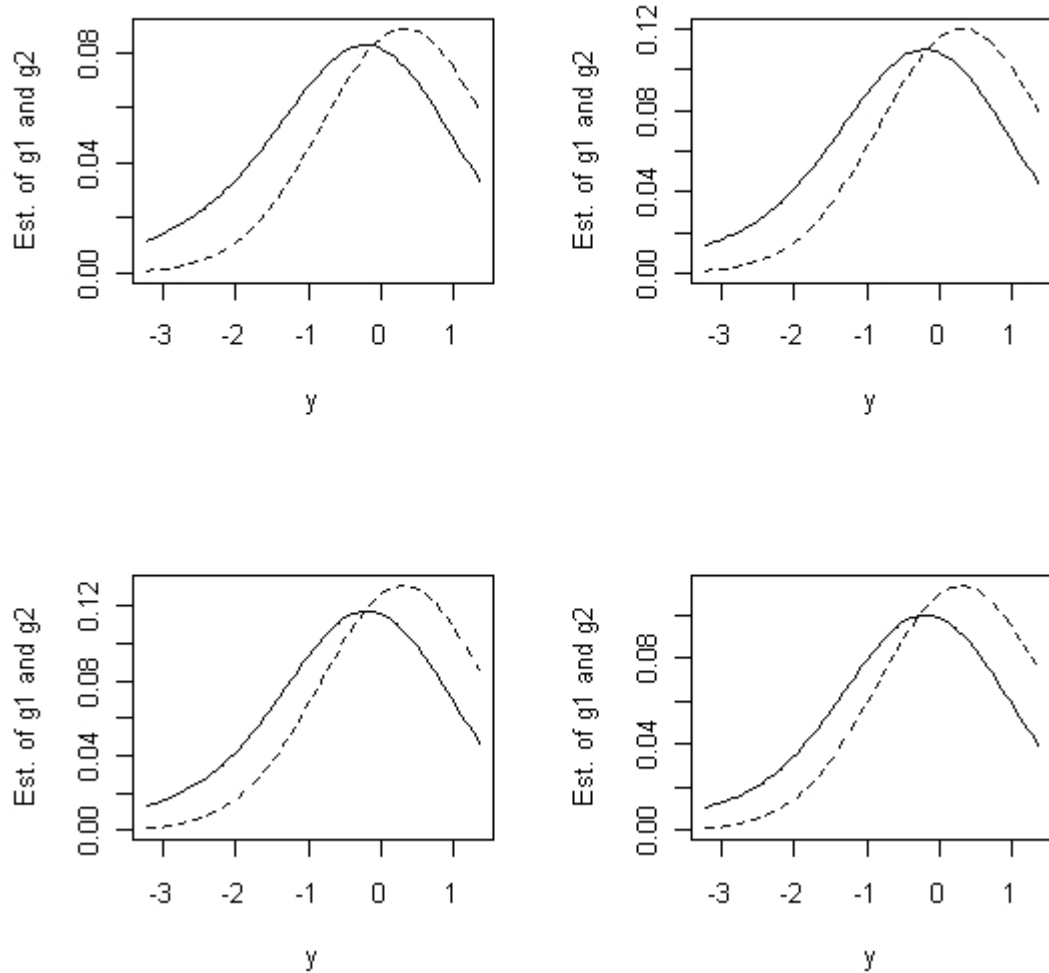


Figure 4.3: Gene 9 and 112: Overlay of vertical slice cuts from Figure 4.1. Gene 9 (x value) was fixed at -0.84 (top left), -0.34 (top right), 0.16 (bottom left), and 0.66 (bottom right). The dashed (continuous) line depicts the estimated g_1 (g_2).

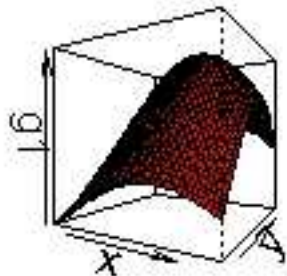
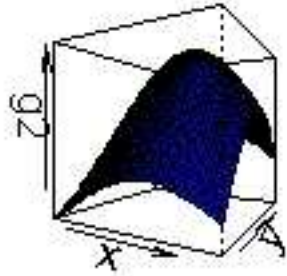


Figure 4.4: Plots of estimated densities of the GC148 gene 9 and 148 (g_1) and the AC148 gene 9 and 148 (g_2).

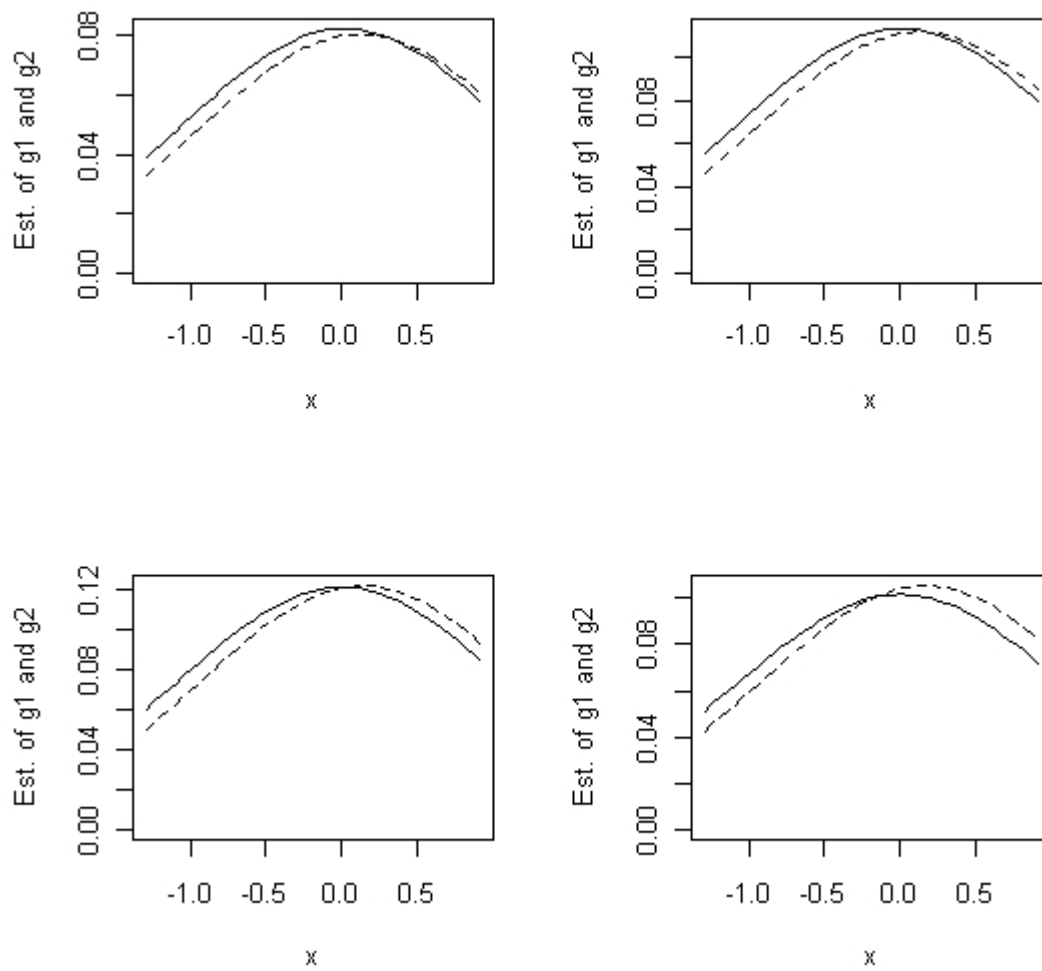


Figure 4.5: Gene 9 and 148: Overlay of vertical slice cuts from Figure 4.4. Gene 148 (y value) was fixed at -1.41 (top left), -0.81 (top right), -0.21 (bottom left), and 0.39 (bottom right). The dashed (continuous) line depicts the estimated g_1 (g_2).

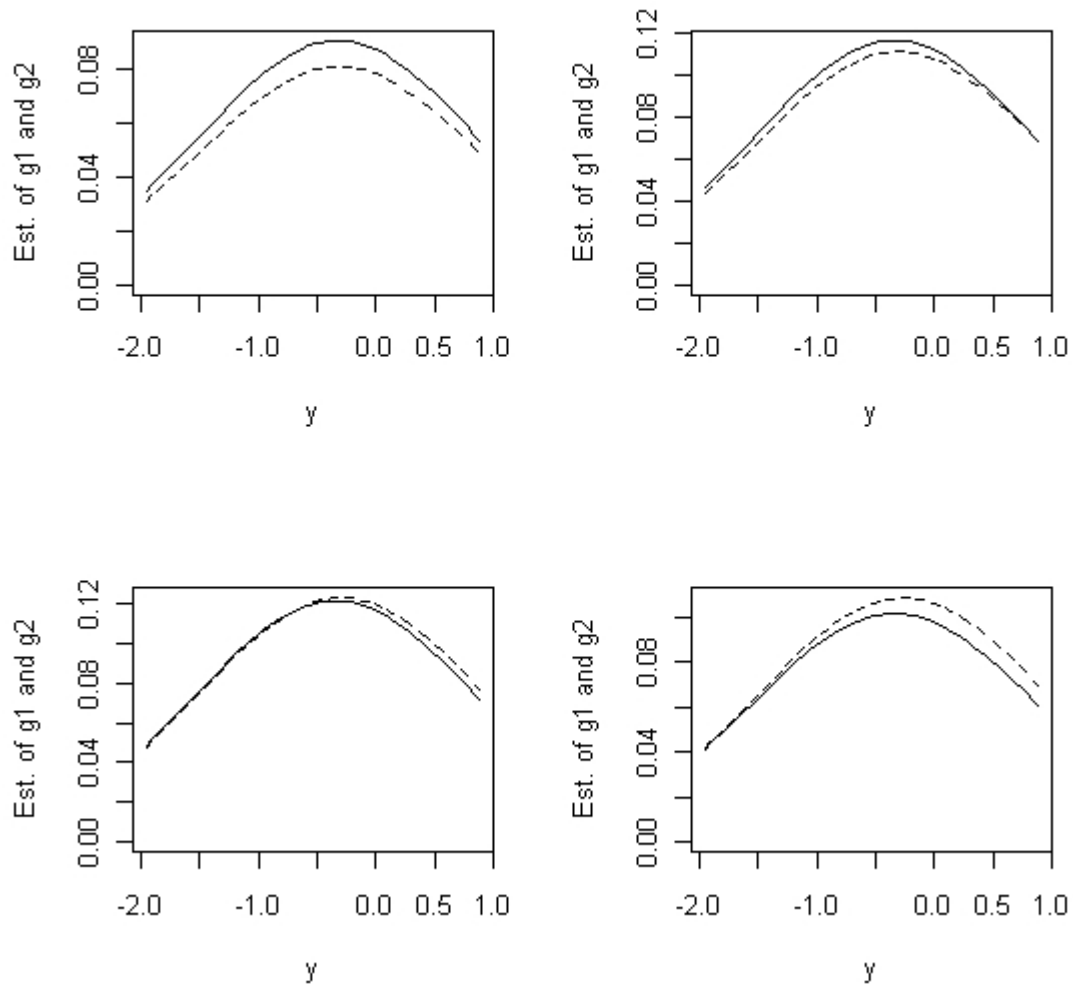


Figure 4.6: Gene 9 and 148: Overlay of vertical slice cuts from Figure 4.4. Gene 9 (x value) was fixed at -0.84 (top left), -0.34 (top right), 0.16 (bottom left), and 0.66 (bottom right). The dashed (continuous) line depicts the estimated g_1 (g_2).

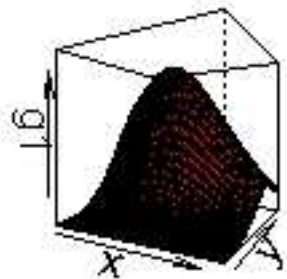
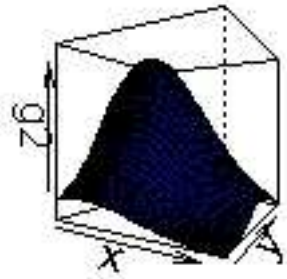


Figure 4.7: Plots of estimated densities of the GC148 gene 100 and 112 (g_1) and the AC148 gene 100 and 112 (g_2).

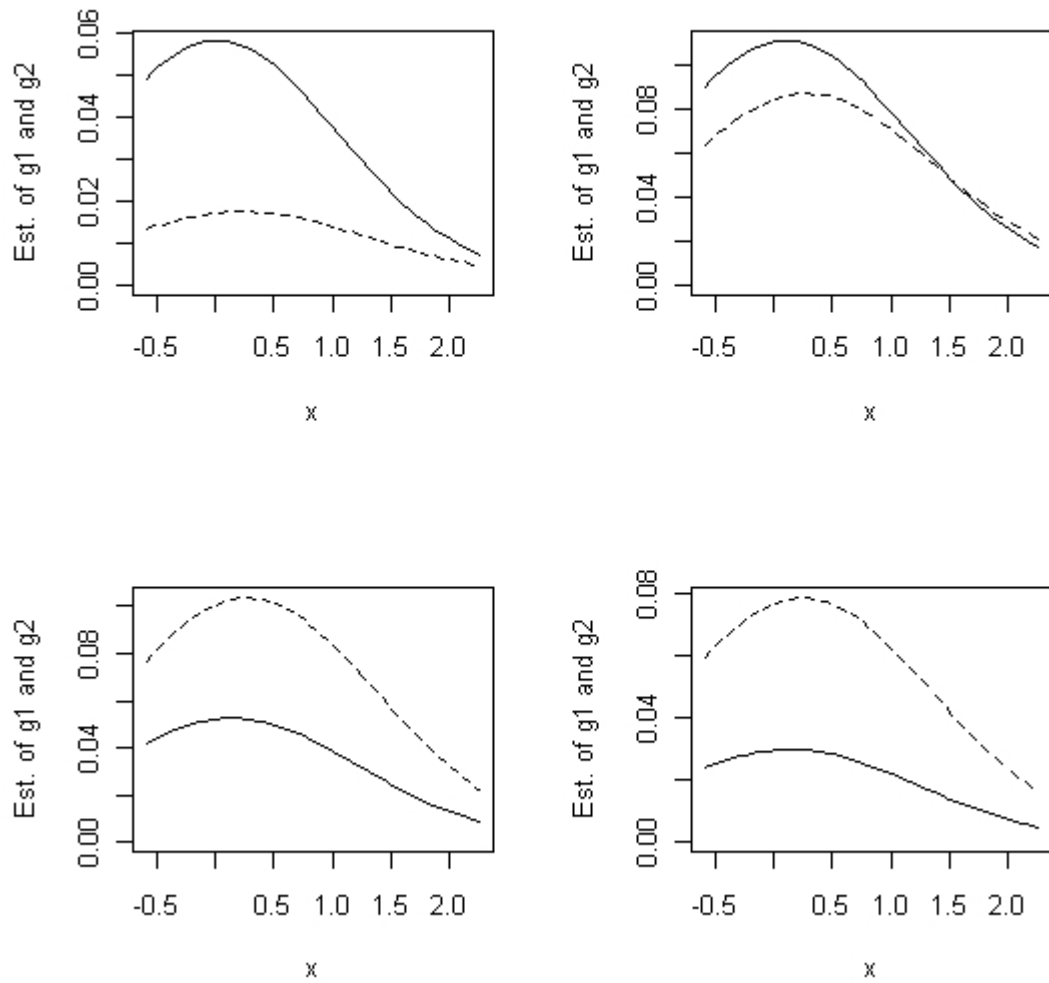


Figure 4.8: Gene 100 and 112: Overlay of vertical slice cuts from Figure 4.7. Gene 112 (y value) was fixed at -1.79 (top left), -0.29 (top right), 1.21 (bottom left), and 1.71 (bottom right). The dashed (continuous) line depicts the estimated g_1 (g_2).

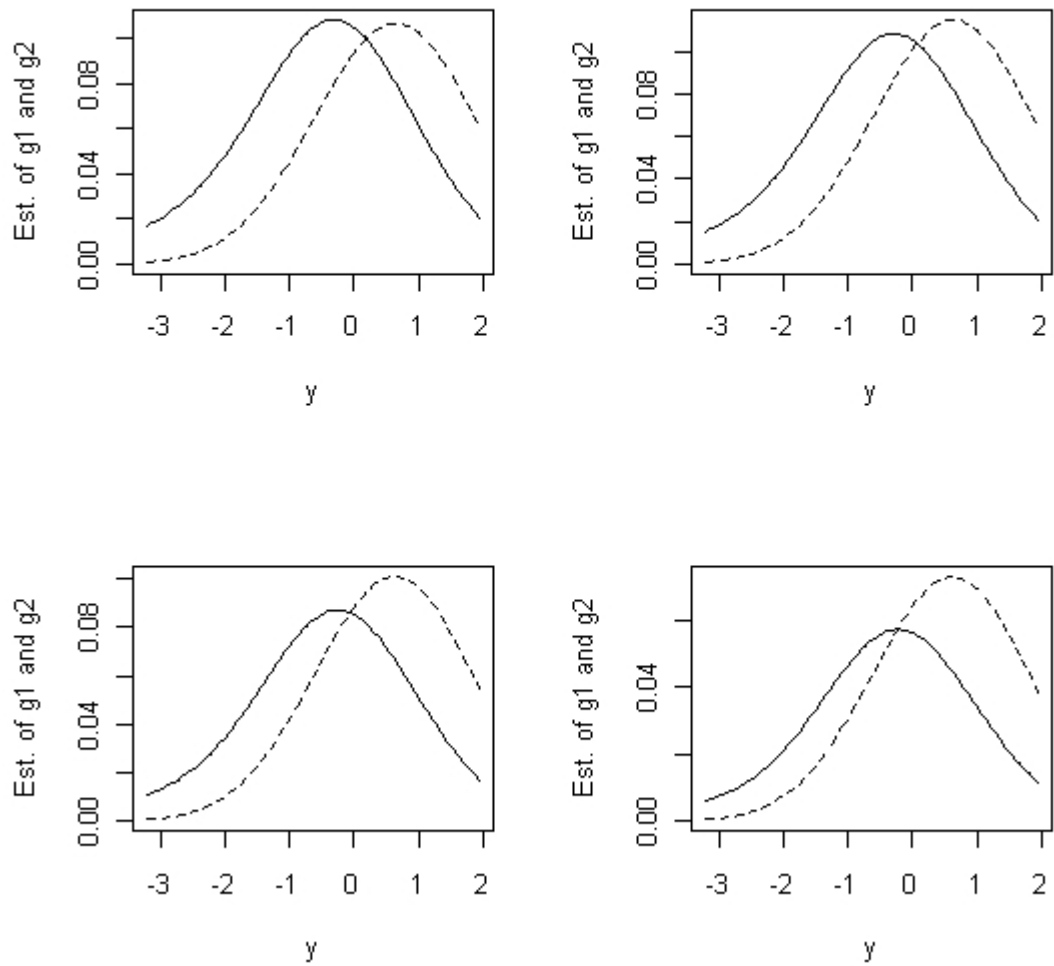


Figure 4.9: Gene 100 and 112: Overlay of vertical slice cuts from Figure 4.7. Gene 100 (x value) was fixed at -0.14 (top left), 0.96 (top right), 0.86 (bottom left), and 1.36 (bottom right). The dashed (continuous) line depicts the estimated g_1 (g_2).

Chapter 5

Summary

A bivariate semiparametric methodology for the joint analysis of risk factors was provided. Regarding the TGCT data discussed in McGlynn et al (2007), it was shown that height and weight are significant risk factors when both factors are considered jointly. Even though body mass was not a significant factor in McGlynn et al (2007) and one-dimensional analysis, I found out that when height and weight are considered jointly, they are significant factors.

Regarding the gene data, the fact that the genes are correlated gave the impetus for our two-dimensional analysis, whose results confirm those in Gagnon (2008).

The two-dimensional density ratio method has several advantages.

- It provides a way for determining the difference between two or more multivariate distributions, and in particular testing multivariate equidistribution.
- The reference $G(x, y)$ and all the parameters are estimated from the combined data \mathbf{t} , and not just from the reference sample, leading to more precise estimates.
- The implementation of the method is surprisingly simple and fast. It took 6 seconds to obtain the results from the TGCT data set with 1691 observations.
- Most important two-dimensional analysis enable us to compute joint probability.

Bibliography

- [1] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J. Jr, Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., Staut, L.M. (2000). Distinct type of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503-511.
- [2] Anderson, T.W. (1971). *An Introduction to Multivariate Statistical Analysis*. (2nd ed.) Wiley, New York.
- [3] Fokianos, K. (2004). Merging information for semiparametric density estimation. *Journal of the Royal Statistical Society B*, **66**, 941-958.
- [4] Fokianos, K., Kedem, B., Qin, J., and Short, D. A. (2001). A semiparametric approach to the one-way layout. *Technometrics*, **43**, 56-65.
- [5] Gagnon R., Kedem B., Ying Q. (2008). On the efficiency of a semiparametric approach to the one-way layout. *Journal of Statistical Theory and Practice*, **2**, 385-406.
- [6] Gagnon, R. (2005). *Computational Aspects of Power Efficiency and State Space Models*. Ph.D. Dissertation, University of Maryland, College Park.
- [7] Gilbert, P. B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Annals of Statistics*, **28**, 151-194.
- [8] Gilbert, P.B., Lele, S.R., and Vardi, Y. (1999), Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials, *Biometrika*, **86**, 27-43.
- [9] Kedem, B. and Wen, S. (2007). Semi-parametric cluster detection. *Journal of Statistical Theory and Practice*, **1**, 49-72.
- [10] Kedem, B., Lu, G., Wei, R., and D. Williams (2008). Forecasting mortality rates via density ratio modeling. *Canadian Journal of Statistics*, **36**, 193-206.
- [11] Lu, Guahua (2007). *Asymptotic Theory for Multiple-sample Semiparametric Density Ratio Models and Its Application to Mortality Forecasting*. Ph.D. Dissertation, University of Maryland, College Park.

- [12] McGlynn, K.A., Devesa, S.S., Sigurdson, A.J., Brown, L.M., Tsao, L., B.S., and Tarone, R.E. (2003). Trends in the incidence of testicular germ cell tumors in the United States. *Cancer*, **97**, 63-70.
- [13] McGlynn, K.A., Sakoda, L.C., Rubertone, M.V., Sesterhenn, I.A., Lyu, C., Graubard, B.I., and Erickson, R.L. (2007). Body size, dairy consumption, puberty, and risk of testicular germ cell tumors. *American Journal of Epidemiology*, **165**, 355-363
- [14] Phue, J.N., Kedem, B., Jaluria, P., and Shiloach, J. (2006). Evaluating microarrays using a semiparametric approach: Application to the central carbon metabolism of *Escherichia coli* BL21 and JM109. *GENOMICS*, **89**, 300-305.
- [15] Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403-411.
- [16] Qin, J. (1998), Inferences for case-control and semiparametric two-sample density ratio models, *Biometrika*, **85**, 619-630.
- [17] Qin, J. and Zhang, B. (1997). A goodness of fit test for logistic regression models based on case-control data. *Biometrika*, **84**, 609-618.
- [18] Wen, S. (2007). *Semi-parametric Cluster Detection*. Ph.D. Dissertation, University of Maryland, College Park.
- [19] Zhang, B. (2000). A goodness of fit test for multiplicative-intercept risk models based on case-control data. *Statistica Sinica*, **10**, 839-866.