Semiparametric Regression in Testicular Germ Cell Data

Anastasia Voulgaraki\textsuperscript{1}, Benjamin Kedem\textsuperscript{1}, and Barry I. Graubard\textsuperscript{2}

\textsuperscript{1}University of Maryland, College Park, MD 20742
avoulgar@math.umd.edu, bnk@math.umd.edu

\textsuperscript{2}National Cancer Institute, Rockville, MD 20852
graubarb@exchange.nih.gov

May 2010

Abstract

It is possible to approach regression analysis with random covariates from a semiparametric perspective where information is combined from multiple multivariate sources. The approach assumes a semiparametric density ratio model where multivariate distributions are “regressed” on a reference distribution. Each multivariate distribution and a corresponding conditional expectation-regression-of interest is then estimated from the combined data from all sources. Graphical and quantitative diagnostic tools are suggested to assess model validity. The method is applied in quantifying the effect of age on weight of germ cell testicular cancer patients. Comparisons are made with both multiple regression and nonparametric kernel regression.

Keywords: Multivariate density ratio model, kernel, regression, random covariates, diagnostic, Nadaraya-Watson.
1 Introduction

The purpose of this paper is to address the relationship between weight, height, and age of germ cell testicular cancer patients. We approach this problem through a nonlinear regression method based on the density ratio model. The method points to the importance of the inclusion of age as a covariate in the prediction of weight, as expressed by a significant reduction in mean square error (MSE) and mean absolute error (MAE), in both case and control groups. This effect is not clearly discernible in a straightforward application of multiple regression.

Before the actual application to germ cell testicular data in Section 4, we describe the method and some of its underpinnings, and suggest appropriate graphical and quantitative diagnostic tools in Section 2. To gain insight into the potential of the approach and the use of the diagnostic tools, we report simulation results in Section 3.

1.1 Background and Preliminaries

Suppose we have \( m = q + 1 \) data sources, such as \( q \) case groups and a control group, each giving a sample of random vectors from an unknown multivariate distribution. Assume that each vector consists of a random response and its random covariates. Given this setup, the semiparametric multivariate density ratio model can be used in providing an alternative to classical regression with random covariates, as well as to kernel nonparametric regression. This approach falls under the general rubric of fusion or integration of information from multiple sources, and it does not depend on the normal assumption.

In the density ratio model one distribution serves as a reference or baseline, and all other distributions are exponential tilts of the reference. In its one dimensional form the model is motivated by the classical one-way analysis of variance with \( m = q + 1 \) independent normal random samples, and logistic regression (Fokianos et al 2001, Qin and Zhang 1997). In its multivariate form, the model is motivated by classical classification given multivariate normal samples, and generalized logistic regression (Anderson 1971, Prentice and Pyke 1979).

Formally, in the one-dimensional case there are \( m = q + 1 \) random samples, \((x_{11}, \ldots, x_{1n_1}), \ldots, (x_{q1}, \ldots, x_{qn_q}), (x_{m1}, \ldots, x_{mn_m})\), with probability density functions \( g_i \)

\[
x_{ij} \sim g_i, \quad i = 1, \ldots, q, m, \quad j = 1, \ldots, n_i, \quad (1)
\]
where \( g_m \equiv g \) is called the reference probability density, and where the \( g_i \) satisfy the density ratio model

\[
\frac{g_j(x)}{g(x)} = \exp(\alpha_j + \beta_j^t h(x)), \quad j = 1, \ldots, q.
\] (2)

Assuming that the distortion function \( h(x) \) is a known vector-valued function, the objective is to estimate the reference density \( g \) and the parameters \( \alpha_j, \beta_j \) from the combined data

\[
t = \{(x_{11}, \ldots, x_{1n_1}), \ldots, (x_{q1}, \ldots, x_{qn_q}), (x_{m1}, \ldots, x_{mn_m})\}'. \quad (3)
\]

The density ratio model has been applied in various problems including kernel density estimation (Fokianos 2004, Cheng and Chu 2004, Qin and Zhang 2005) analysis of variance (Fokianos et al 2001), AIDS vaccine trials (Gilbert et al 1999), mortality rate prediction (Kedem et al 2008), case-control studies (Prentice and Pyke 1979, Qin 1998), logistic model validation (Qin and Zhang 1997), cluster detection (Wen and Kedem 2009), and goodness of fit (Zhang 2000). A two-dimensional case-control application has been made recently in Kedem et al (2009).

In this paper the \( L \)-dimensional formulation of the model is used in the estimation of the conditional expectation of a response given covariate information. Specifically, for each of the \( m \) data sources, we use the \( L \)-dimensional density ratio model in predicting, via the estimated conditional expectation, the response variable given the corresponding covariate information, and propose measures of goodness of fit and diagnostic plots to check the validity of the model. A comparison with linear multiple regression and the Nadaraya-Watson kernel nonparametric regression is made using both real and simulated data.

### 1.2 Motivation

The \( L \)-dimensional formulation of the model was motivated by an extension of a previous analysis of two risk factors, body weight and height, of germ cell testicular cancer to including three or more risk factors or covariates (Kedem et al. 2009). We wanted to include age in the analysis with height and weight as age is both an important risk factor and potential confounder since the incidence of testicular cancer varies by age, peaking around 25-35 years for the most common types of testicular cancer, and age correlates with body weight (McGlynn and Cook, 2010; Ogden et al. 2004). The use of a two-dimensional density ratio model in the previous analysis uncovered
an important contribution of body weight in the presence of height that was not observed in logistic regression analyses (McGlynn et al. 2006). The proposed extension of the density ratio model provides an opportunity to explore the interrelationships of height and weight with testicular cancer while controlling for age by estimating the conditional expectation of weight given height and age.

2 Statistical Formulation

2.1 The L-Dimensional Density Ratio Model

Suppose we have \( m = q + 1 \) data sets or samples of \( L \)-dimensional vectors, where each vector consists of \( L - 1 \) covariates and one response, and assume that the \( i \)th sample size is \( n_i \). Thus, for \( i = 1, \ldots, q, m, j = 1, \ldots, n_i \) we have

\[
(x_{ij1}, x_{ij2}, \ldots, x_{ij(L-1)}, y_{ij}) \sim g_i(x_1, \ldots, x_{(L-1)}, y).
\]

We choose \( g \equiv g_m(x_1, \ldots, x_{(L-1)}, y) \) as a reference or baseline probability density function (pdf), and let each \( g_i(x_1, \ldots, x_{(L-1)}, y), i = 1, \ldots, q \) be an exponential distortion or tilt of the reference distribution,

\[
\frac{g_i(x)}{g(x)} = \exp(\alpha_i + \beta_i'x), \quad i = 1, \ldots, q
\] (4)

where \( x = (x_1, \ldots, x_{(L-1)}, y)' \) and \( \beta_i = (\beta_{i1}, \ldots, \beta_{iL})' \). Since the \( g_i(x), i = 1, \ldots, q, m \) are probability densities, \( \beta_i = 0 \) implies \( \alpha_i = 0, j = 1, \ldots, q \). It follows that the hypothesis \( H_0 : \beta_1 = \cdots = \beta_q = 0 \) implies equidistribution: all the \( g_i \) are equal. Model (4) is referred to as a \textit{density ratio model}.

To estimate the parameters and the reference \( g \), or equivalently the reference distribution function \( G \), we follow the same procedure described in Fokianos et al (2001), Qin and Zhang (1997), and Qin (1998). First the data are combined in a single vector \( t \) of length \( n = n_1 + n_2 + \cdots + n_m \),

\[
t = ((x_{ij1}, x_{ij2}, \ldots, x_{ij(L-1)}, y_{ij}) : i = 1, \ldots, q, m, j = 1, \ldots, n_i)' = (t_1', t_2', \ldots, t_n')'
\] (5)

where \( t_i = (t_{ix1}, \ldots, t_{ix_{L-1}}, t_{iy})' \). The idea is to approximate the reference distribution function by a step function \( G \) (same notation is used) with jumps \( p_i \) at all the observed points (Vardi 1982,1985). For the three-dimensional case:

\[
p_i = G(t_{ix_1}, t_{ix_2}, t_{iy}) - G(t_{i-1,x_1}, t_{ix_2}, t_{iy}) - G(t_{ix_1}, t_{i-1,x_2}, t_{iy})
\]
Thus, the \( p_i \) are the jumps in the \( L \)-dimensional step function \( G \) at \( t_1, \ldots, t_n \).

The empirical likelihood (cf Owen 2001) is then a function of \( \hat{p_i}, \alpha = (\alpha_1, \ldots, \alpha_q)' \) and \( \beta = (\beta_1', \ldots, \beta_q')' \):

\[
L(\alpha, \beta, G) = \prod_{i=1}^{n} p_i \prod_{k=1}^{n_1} \exp(\alpha_1 + \beta_1 x_{1k1} + \cdots + \beta_1(L-1) x_{1k(L-1)} + \beta_1 L y_{1k}) \times \cdots \times \prod_{k=1}^{n_q} \exp(\alpha_q + \beta_q x_{qk1} + \cdots + \beta_q(L-1) x_{qk(L-1)} + \beta_q L y_{qk})
\]

subject to the constraints

\[
\sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} w_1(t_i)p_i = 1, \ldots, \quad \sum_{i=1}^{n} w_q(t_i)p_i = 1
\]

where

\[
w_j(t_i) = \exp(\alpha_j + \beta_j' t_i), \quad j = 1, \ldots, q.
\]

Estimates for \( \hat{\alpha}_j \) and \( \hat{\beta}_j \) are obtained by solving the score equations:

\[
\frac{\partial l}{\partial \alpha_j} = -\sum_{i=1}^{n} \frac{\rho_j w_j(t_i)}{1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)} + n_j = 0 \quad (8)
\]

\[
\frac{\partial l}{\partial \beta_j} = -\sum_{i=1}^{n} \frac{\rho_j w_j(t_i) t_i}{1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)} + \sum_{i=1}^{n_j} (x_{ji1}, \ldots, y_{ji})' = 0 \quad (9)
\]

for \( j = 1, \ldots, q \) and \( \rho_j = n_j/n_m \). Then

\[
\hat{p_i} = \frac{1}{n_m} \cdot \frac{1}{1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)} \quad (10)
\]

\[
\hat{G}(t) = \frac{1}{n_m} \cdot \frac{n}{\sum_{i=1}^{n} \frac{I(t_i \leq t)}{1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)}} \quad (11)
\]

where \((t_i \leq t)\) is defined pointwise, \( w_j(t_i) = \exp(\hat{\alpha}_j + \hat{\beta}_j' t_i) \), and \( I(B) \) is the indicator of the event \( B \). It can be shown that the estimators \( \hat{\theta} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_q, \hat{\beta}_1, \cdots, \hat{\beta}_q)' \) are asymptotically normal

\[
\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, \Sigma) \quad (12)
\]
as $n \to \infty$, where $\hat{\theta}_0$ denotes the true parameters and $\Sigma = S^{-1}VS$ is defined in the appendix.

Notice that $g_j(t_i), j = 1, \ldots, q$ can be estimated as exponential tilts of $\hat{p}_i$. Thus, under the $L$-dimensional density ratio model we can predict the response $y$ given the covariate information $x_1, x_2, \ldots, x_{(L-1)}$ for any of the $m$ data sets as follows:

$$\hat{E}_j(y \mid x_1, \ldots, x_{(L-1)}) = \sum_{i} y_i \frac{\hat{g}_j(x_1, \ldots, x_{(L-1)}, y_i)}{\sum_{y_i} \hat{g}_j(x_1, \ldots, x_{(L-1)}, y_i)}, j = 1, \ldots, q, m. \quad (13)$$

The $\hat{g}_j$ are kernel density estimates in the sense of Fokianos (2004),

$$\hat{g}_j(z_0) = \frac{1}{hL} \sum_{i=1}^{n_j} \hat{p}_i \hat{w}_j(t_i) K((t_i - z_0)/h), j = 1, \ldots, m. \quad (14)$$

where $z_0$ is $L$-dimensional. From Fokianos (2004) the asymptotic mean integrated square error (AMISE) of $\hat{g}_j$ converges to 0 as $n \to \infty, h \to 0$ and $nhL \to \infty$. This implies that $\hat{E}_j(y \mid x_1, \ldots, x_{(L-1)})$ is a consistent estimator of $E(y \mid x_1, \ldots, x_{(L-1)})$ in the $j$th population in the sense of convergence in probability at least for bounded data (see the appendix).

As in the Nadaraya-Watson kernel estimate (Nadaraya 1964, Watson 1964), the estimated conditional expectation (13) is of the form $\sum_i w_i y_i$, where the $w_i$ are positive weights which sum to 1, except that here the $w_i$ also depend on the $y_i$.

### 2.2 Diagnostic Plots and Measures of Goodness-of-Fit

The density ratio model motivates graphical and quantitative diagnostic tools for measuring both goodness-of-fit of the model and the quality of the regression (13). Goodness-of-fit tests have been proposed by Gilbert (2004), Qin and Zhang (1997), and Zhang (1999, 2001, 2002), where the appropriateness of the model is judged by the closeness of the estimated reference distribution to the corresponding empirical distribution. Bondell (2007) suggests a reformulation of this in terms of the corresponding kernel density estimates. We suggest data analytic tools to measure discrepancies stemming from both all case and control (reference) groups.

Graphical evidence of goodness-of-fit can be obtained from the plots of $\hat{G}_i$ versus the corresponding empirical multivariate distribution function $\tilde{G}_i$, $i = 1, \ldots, q, m$, evaluated at some selected $L$-dimensional points as to obtain two dimensional plots. Figures 1 and 2 in the next section are examples of this.
We found the following measure of goodness-of-fit useful. Consider the
ith sample of size \( n_i \), and let \( x \) be the number of times the estimated semi-
parametric cdf falls in the \( 1 - \alpha \) confidence interval obtained from the cor-
responding empirical cdf, both evaluated at the sample points. Define
\[
R_{\alpha,k}^2 = 1 - \exp \left\{ -\left( \frac{x}{n_i} \right)^k \right\}
\] (15)
where \( k > 0 \). Observe that:

- \( R_{\alpha,k}^2 \) takes values between 0 and 1, being close to 1 when \( x \) approaches
  \( n_i \) and close to 0 when \( x \) is close to 0.

- \( R_{\alpha,k}^2 \) is a flexible criterion that can be adjusted by changing the pa-
  rameters \( \alpha \) and \( k \).

- Computing \( R_{\alpha,k}^2 \) is both simple and fast.

We describe next three natural alternatives to \( R_{\alpha,m}^2 \). First, as in multiple
regression, goodness-of-fit may be approached by residual analysis. In this
vein, consider the decomposition
\[
E[y - E(y)]^2 = E[y - E(y \mid x)]^2 + E[E(y \mid x) - E(y)]^2.
\]
Therefore, from the approximations \( \bar{y} \approx E(y) \) and \( \hat{y} \equiv E(y \mid x) \),
\[
\frac{1}{n} \sum (y_i - \bar{y})^2 \approx \frac{1}{n} \sum (y_i - \hat{y}_i)^2 + \frac{1}{n} \sum (\hat{y}_i - \bar{y})^2
\]
we define “\( R^2 \)” as in linear regression:
\[
R_1^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \tag{16}
\]
Next, define
\[
R_2^2 = \text{corr}(y, \hat{y})^2 \tag{17}
\]
Lastly, following Qin and Zhang (1997), define
\[
R_3^2 = \exp(-\sqrt{n} \cdot \max \mid \hat{G}_i - \tilde{G}_i \mid) \tag{18}
\]
Clearly, \( R_3^2 \) takes values between 0 and 1. Alternatives to \( R_3^2 \) are \( \exp(-\sqrt{n} \cdot \text{median} \mid \hat{G}_i - \tilde{G}_i \mid) \) or \( \exp(-\frac{1}{n} \sum \mid \hat{G}_i - \tilde{G}_i \mid^2) \).

The following simulation study suggests that \( R_{\alpha,k}^2 \) is a more pragmatically
indicator of goodness-of-fit compared to \( R_1^2, R_2^2, \) and \( R_3^2 \). We note that the
coefficient of determination suggested in Nagelkerke (1991) was not found
sufficiently sensitive in the present context.
3 Some Simulation Results

In the present simulation study $m = 2$, and $g_2$ denotes the reference distribution. We considered the following bivariate cases (runs).

1. $g_1 \sim N((0, 0)', \Sigma)$, $g_2 \sim N((0, 0)', \Sigma)$ with $\Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$, $n_1 = 40$, $n_2 = 30$.

2. $g_1 \sim N((0, 0)', \Sigma)$, $g_2 \sim N((1, 1)', \Sigma)$ with $\Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$, $n_1 = 200$, $n_2 = 200$.

3. $g_1$ from standard two dimensional Multivariate Cauchy and $g_2$ from two dimensional Multivariate Cauchy with $\mu = (1, 1)'$, $V = \begin{pmatrix} 5 & 5 \\ 5 & 10 \end{pmatrix}$, $n_1 = 200$, $n_2 = 200$.

4. $g_1$ from standard two dimensional Multivariate Cauchy and $g_2$ from uniform distribution on the triangle $(0, 0), (6, 0), (-3, 4)$, and $n_1 = 200$, $n_2 = 200$.

The normal distribution follows the density ratio model, but this is not true for the Cauchy and the uniform distributions. Hence we expect to see straight lines in the diagnostic plots and high $R^2$'s, as defined above, in cases (1) and (2). On the other hand, we expect to see deviations from straight lines in the diagnostic plots and lower $R^2$'s in cases (3) and (4).

Figures 1-2 show the estimated $\hat{G}_1$ and $\hat{G}_2$ (where $\hat{G}_1$ is the exponential tilt of $\hat{G}_2$ defined in (11)) versus the empirical cdf $\tilde{G}_1$ and $\tilde{G}_2$, respectively, all obtained from the simulated case-control data, and evaluated at selected 2D points. As expected, in cases (1), (2), there is almost a perfect agreement between $\hat{G}_i$ versus $\tilde{G}_i$, $i = 1, 2$, whereas Figure 2 shows clearly that the density ratio model is not appropriate for the data from cases (3) and (4).

A comparison of the different measures of goodness of fit is given in Table 1. Apparently here $R^2_1$ and $R^2_2$ are misleading as measures of goodness of fit. They are erroneously higher at the cases where the simulated distributions do not follow the density ratio model. It seems $R^2_3$ is more appropriate than both $R^2_1$ and $R^2_2$ but it is sensitive to outliers and can give low values even for data that follow the density ratio model (e.g. case 2). On the other hand, the proposed measure $R^2_{\alpha,k}$ classifies correctly the four cases, giving high values for runs (1) and (2) and low values for (3) and (4). The values
Figure 1: Case-control plots of $\hat{G}_i$ vs. $\tilde{G}_i$, $i = 1, 2$, cases (1) and (2)

Figure 2: Case-control plot of $\hat{G}_i$ vs. $\tilde{G}_i$, $i = 1, 2$, cases (3) and (4)
of $R_{\alpha,k}^2$ in Table 1 were calculated with $k = 2$ and $1 - \alpha = 90\%$. In general, $R_{\alpha,k}^2$ gets closer to $R_3^2$ by lowering $1 - \alpha$.

Table 1: Comparison of goodness of fit measures for case and control.

| Run | Group | $R_{1}^2$ | $R_{2}^2$ | $R_{3}^2$ | $R_{10,2}^2$ |
|-----|-------|---------|---------|---------|-------------|
| (1) Case | 0.1947 | 0.6196 | 0.7702 | 1 |
| Ctrl | 0.3123 | 0.4812 | 0.7422 | 1 |
| (2) Case | 0.0290 | 0.0470 | 0.3281 | 0.9998 |
| Ctrl | 0.1214 | 0.2356 | 0.3651 | 0.9999 |
| (3) Case | 0.6948 | 0.8441 | 0.1390 | 0.1469 |
| Ctrl | 0.6792 | 0.7537 | 0.1294 | 0.1219 |
| (4) Case | 0.4978 | 0.5662 | 0.0340 | 0.0999 |
| Ctrl | 0.4277 | 0.4372 | 0.0305 | 0.0001 |

Figure 3 shows the estimated $E[Y \mid X]$ using equation (13) for the first two cases. The prediction line is apparently influenced by the endpoints but otherwise it is a smooth curve. Superimposed is the line obtained from simple linear regression. From Table 2, except in one case, the semiparametric prediction gives lower MSE and MAE than linear regression.

Table 2: Comparison of the semiparametric regression (13) and linear regression for simulations (1) and (2) in terms of MSE and MAE.

| Simulation | MSE | MAE |
|------------|-----|-----|
|            | Sem. Prediction | Linear Reg. | Sem. Prediction | Linear Reg. |
| Simulation 1 | $G_1$ | 1.4534 | 1.3251 | 0.8583 | 0.8758 |
|             | $G_2$ | 1.1540 | 1.3363 | 0.8640 | 0.9693 |
| Simulation 2 | $G_1$ | 0.8187 | 0.8411 | 0.7224 | 0.7302 |
|             | $G_2$ | 1.2909 | 1.4825 | 0.9820 | 0.9919 |

4 Application to Testicular Germ Cell Cancer

Testicular germ cell tumor (TGCT) is a common cancer among U.S. men, mainly in the age group of 15-35 years (McGlynn et al 2003). In McGlynn et al (2007) it was shown that increased risk was significantly related to height, whereas body mass index was not significant. In Kedem et al
(2009), using the two dimensional semiparametric model, it was shown that jointly height and weight are significant risk factors. The TGCT data consist of age (years), height (cm) and weight (kg) of 1691 individuals, of which $n_1 = 763$ are cases and $n_2 = 928$ belong to the control group. In the present application $m = 2$ and the chosen bandwidth is $h = 0.3$. We use the semiparametric regression (13) to predict weight given the covariates age and height for both the case and the control groups, holding the control distribution as reference. The measures of goodness of fit discussed in Section (2.2) are applied to assess the model and the predictions, and the results are compared with those from multiple regression and the Nadaraya-Watson regression.

Before applying the three-dimensional density ratio model to the TGCT data, it is interesting to apply the two-dimensional model to get prediction of weight given only height. As Figure 4 shows, the density ratio model is a suitable model for the TGCT data: there is almost a perfect agreement between the plots of the estimated semiparametric $\hat{G}_i$ and the corresponding empirical $\tilde{G}_i$, $i = 1, 2$. Figure 5 shows the estimated $E[Y \mid X]$ using equation (13) for the case and control groups. Superimposed is the regression line obtained from linear regression under the normal assumption. The residual plots in Figure 6 are centered around zero. The MSE values for control are 89.33101 for the semiparametric model and 92.26375 for the regression model; for case the MSE values are 99.10386 (semiparametric) and 99.51023
Figure 4: 2D problem: Plots of $\hat{G}_i$ versus $\tilde{G}_i$, $i = 1, 2$ evaluated at (height, weight) pairs for the case and control groups from the TGCT data.

Figure 5: Comparison of $\hat{E}[\text{weight} \mid \text{height}]$ in (13) and linear regression.
Figure 6: Residual plots for the TGCT case and control groups from $\hat{E}[\text{weight} \mid \text{height}]$ in (13).

(regression). The corresponding MAE values are 7.212146 (semiparametric) and 7.295542 (regression) for control, and 7.801397 (semiparametric) and 7.784165 (regression) for case. The value of $R_{20,1}^2$ is 1 for both case and control.

From the preceding results, in the 2D problem the two models give similar results. However, from Table 3, the introduction of the covariate age results in a significant reduction in MSE and MAE from the semiparametric model, whereas the multiple regression MSE and MAE stay almost unchanged.

Tables 4 and 5 give some predicted values for weight given age and height for the two models. These, the diagnostic plots in Figures 7 and 8, and the fact $R_{20,1}^2 = 1$ for both case and control tell us that the semiparametric model is quite appropriate for the TGCT data.

We end this section by noting that, as expected, $\hat{E}(y \mid x)$ in (13) tends to be close to the average of y’s which correspond to the same x. This is demonstrated in Table 6 which gives the case-control weight predictions (13) and the actual weights. Empty entries in the table correspond to subjects with the same height and age (i.e. same x), but possibly different weights. The averaging property can be seen by averaging the run of weights in the “empty cells” and the run upper bound. Thus, for example, the control-weights corresponding to age 22 and height 175.26 average to 74.3894 and the prediction is 75.67005.
Table 3: Case-control MSE and MAE for weight given height and age.

| Case          | Control |
|---------------|---------|
| MSE           | MAE     | MSE     | MAE     |
| Semiparametric model | 77.09442 6.794231 | 73.97413 6.45978 |
| Multiple regression  | 96.36657 7.678669 | 90.29083 7.24368 |

Table 4: Predicted control values of weight given height and age.

| Case          | Control |
|---------------|---------|
| Age | Height | Weight | Semiparametric | Mult. Regression |
|-------|--------|--------|----------------|------------------|
| 26    | 193.04 | 102.058 | 99.09662       | 92.47554         |
| 24    | 167.64 | 72.575  | 71.28688       | 70.00329         |
| 29    | 180.34 | 65.771  | 81.45212       | 82.42360         |
| 38    | 185.42 | 81.647  | 85.17266       | 89.46406         |
| 34    | 195.58 | 89.811  | 86.53122       | 97.03194         |
| 27    | 162.56 | 58.967  | 59.08479       | 66.51540         |

Table 5: Predicted case values of weight given height and age.

| Control          | |
|------------------|---------------------------|
| Age | Height | Weight | Semiparametric | Mult. Regression |
|-------|--------|--------|----------------|------------------|
| 29    | 180.34 | 90.718 | 81.32563       | 82.06293         |
| 39    | 175.26 | 77.111 | 76.02246       | 80.36549         |
| 19    | 172.72 | 63.503 | 72.09521       | 73.58821         |
| 33    | 177.80 | 83.915 | 84.51664       | 80.97707         |
| 31    | 190.50 | 102.058| 90.09859       | 90.67494         |
| 25    | 165.10 | 58.967 | 59.11650       | 68.90777         |
Figure 7: Case-control plots of $\hat{G}_i$ versus $\tilde{G}_i$, $i = 1, 2$ for the 3D TGCT problem: the $\hat{G}_i, \tilde{G}_i$ are evaluated at selected (age, height, weight) triplets.

Figure 8: Residual plots from the regression of weight given height and age using (13) for the case and control groups from the TGCT data.
Table 6: Case-control weight and $\hat{E}[weight/height, age]$.

| Age | Height | Weight | $E[W \mid H, A]$ | Weight | $E[W \mid H, A]$ |
|-----|--------|--------|-------------------|--------|-------------------|
| 27  | 162.56 | 58.967 | 59.08212         | 58.967 | 59.08479         |
| 28  | 162.56 | 77.111 | 69.85564         | 65.771 | 69.90699         |
| 30  | 165.10 | 68.039 | 70.01298         | 72.575 | 70.02422         |
| 37  | 165.10 | 69.40  | 66.70674         | 63.503 | 66.72657         |
| 25  | 167.64 | 86.183 | 77.18912         | 72.575 | 77.4467          |
|     |        |        |                   | 90.718 | 63.503           |
| 30  | 167.64 | 72.575 | 80.2309          | 88.451 | 80.37712         |
| 18  | 170.18 | 61.235 | 67.6877          | 72.575 | 67.7608          |
| 32  | 170.18 | 70.307 | 72.34239         | 81.647 | 72.46667         |
|     |        |        |                   | 63.503 |                 |
| 37  | 172.72 | 74.843 | 80.73858         | 88.451 | 80.84287         |
| 40  | 172.72 | 70.307 | 77.72428         | 90.718 | 77.8726          |
|     |        |        |                   | 77.111 |                 |
| 22  | 175.26 | 77.111 | 75.67005         | 86.183 | 75.8382          |
|     |        |        |                   | 65.771 | 86.183           |
|     |        |        |                   | 79.379 |                 |
|     |        |        |                   | 83.915 |                 |
|     |        |        |                   | 65.771 |                 |
| 25  | 175.26 | 68.039 | 75.74933         | 79.379 | 75.86762         |
|     |        |        |                   | 83.915 | 72.575           |
|     |        |        |                   | 74.843 | 83.915           |
|     |        |        |                   | 83.915 | 74.843           |
|     |        |        |                   | 79.379 | 72.575           |
|     |        |        |                   | 86.183 | 74.843           |
|     |        |        |                   | 74.843 | 61.235           |
|     |        |        |                   | 61.235 | 65.771           |
|     |        |        |                   | 79.379 |                 |
| 26  | 177.80 | 79.379 | 78.7966          | 77.111 | 78.9987          |
|     |        |        |                   | 81.647 | 104.326          |
|     |        |        |                   | 58.967 | 77.111           |
|     |        |        |                   | 81.647 |                 |
|     |        |        |                   | 79.379 |                 |
|     |        |        |                   | 74.843 |                 |
|     |        |        |                   | 88.451 |                 |
|     |        |        |                   | 68.039 |                 |

16
4.1 Comparison With Nadaraya-Watson

The Nadaraya-Watson (NW) kernel estimator of $E(y \mid x)$ is estimated by a weighted average $\sum_i w_i y_i$ where the $w_i$ are large for $x_i$ close to $x$ and small for $x_i$ farther away from $x$. Thus, nearest neighbors are counted but $x_i$ whose distance from $x_i$ is relatively large are discounted (Lee 1996, p. 144). From experience, for sufficiently condensed data the NW estimator and (13) are quite comparable. Thus, it is interesting to compare (13) with the Nadaraya-Watson estimator for troublesome scenarios when $x$ is
an extreme case with few or no neighbors.

For such a comparison, we identified 15 extreme (age, height) case pairs and 15 control pairs, and tried to predict the corresponding weights from the rest (or “middle”) of the TGCT data. Accordingly, the semiparametric model parameters were estimated with \( n_1 = 40 \) values from case and \( n_2 = 50 \) values from control, giving a total of \( n = 90 \) observations.

We encountered a technical difficulty when computing the NW estimator using R, as the default value for the control bandwidth for age was huge, 168,395,326. However, for case the default bandwidths were 5.432816 for age and 4.25757 for height. Thus, to conform to these last bandwidths, we fitted the semiparametric model once with bandwidth equal to 4 and once with bandwidth equal to 5, and report only case results in Table 7. It should be noted that the NW estimator was computed with \( n_1 = 40 \) case observations, whereas (13) was computed from a combined sample with \( n_1 + n_2 = 90 \) case and control observations. From Table 7 it seems that the semiparametric prediction is somewhat more immune to extremes than NW.

| Model                                      | MSE  | MAE  |
|--------------------------------------------|------|------|
| Semiparametric model with bw=4             | 101.543 | 9.222 |
| Semiparametric model with bw=5             | 90.525 | 8.657 |
| Multiple Regression                        | 109.729 | 8.848 |
| Nadaraya-Watson                            | 119.984 | 10.252 |

5 Summary

In this paper we have demonstrated that the \( L \)-dimensional density ratio model is useful in estimating the conditional expectation of a response variable given random covariates when multiple data sources are available. In addition we have suggested overall qualitative (graphical) and quantitative validation measures to assess the suitability of the method. The simulation results, the analysis of the TGCT data, and the comparison with multiple regression and the Nadaraya-Watson estimator point to the merit of the method, at least for a small number of covariates.
The approach offers a way of understanding how multivariate distributions representing many different data sources are related to each other. This leads to a ramification of the notion of regression where the objective is to model relationships between distributions. Relationships between response variables and their covariates, corresponding to the data sources, are byproducts.

Finally, we note that the suggested validation measures may shed light on the selection of \( h(x) \) in \( (2) \). This idea was not dealt with in the present paper.

**Appendix: Asymptotic Results**

**Computing S, V**

Define

\[
\nabla \equiv \left( \frac{\partial}{\partial \alpha_1}, \ldots, \frac{\partial}{\partial \alpha_q}, \frac{\partial}{\partial \beta_1}, \ldots, \frac{\partial}{\partial \beta_q} \right)'
\]

Then \( E[\nabla l(\theta)] = E[\nabla l(\alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_q)] = 0 \). Let

\[
E_j(t) = \int t w_j(t) dG(t)
\]

\[
A_0(j, r) = \int \frac{w_j(t)w_r(t)}{1 + \sum_{k=1}^q \rho_k w_k(t)} dG(t)
\]

\[
A_1(j, j') = \int \frac{w_j(t)w_{j'}(t) t}{1 + \sum_{k=1}^q \rho_k w_k(t)} dG(t)
\]

\[
A_2(j, j') = \int \frac{w_j(t)w_{j'}(t) t t'}{1 + \sum_{k=1}^q \rho_k w_k(t)} dG(t)
\]

for \( j, j' = 1, \ldots, q \). The entries in

\[
V \equiv Var \left[ \frac{1}{\sqrt{n}} \nabla l(\alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_q) \right] \tag{19}
\]

are

\[
\frac{1}{n} Var(\frac{\partial l}{\partial \alpha_j}) = \frac{\rho_j^2}{1 + \sum_{k=1}^q \rho_k} [A_0(j, j) - \sum_{r=1}^m \rho_r A_0^2(j, r)]
\]

\[
\frac{1}{n} Cov(\frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \alpha_{j'}}) = \frac{\rho_j \rho_{j'}}{1 + \sum_{k=1}^q \rho_k} [A_0(j, j') - \sum_{r=1}^m \rho_r A_0(j, r) A_0(j', r)]
\]
where

\[ \frac{1}{n} \text{Cov} \left( \frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \beta_j} \right) = \frac{\rho_j^2}{1 + \sum_{k=1}^q \rho_k} [A_0(j,j) E_j(t') - \sum_{r=1}^m \rho_r A_0(j,r) A'_1(j,r)] \]

\[ \frac{1}{n} \text{Cov} \left( \frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \beta_{j'}} \right) = \frac{\rho_j \rho_{j'}}{1 + \sum_{k=1}^q \rho_k} [A_0(j,j') E_{j'}(t') - \sum_{r=1}^m \rho_r A_0(j,r) A'_1(j',r)] \]

\[ \frac{1}{n} \text{Cov} \left( \frac{\partial l}{\partial \beta_j}, \frac{\partial l}{\partial \beta_{j'}} \right) = \frac{\rho_j \rho_{j'}}{1 + \sum_{k=1}^q \rho_k} [-A_2(j,j') + E_j(t) A'_1(j,j')] \]

\[ + E_{j'}(t) A_1(j,j') - \sum_{r=1}^m \rho_r A_1(j,r) A'_1(j',r) \]

\[ + \frac{1}{n} \sum_{i=1}^{n_j} \sum_{k=1}^{n_{j'}} \text{Cov}[(x_{ji1}, \ldots, y_{ji}), (x_{j'k1}, \ldots, y_{j'k})] \]

The last term is zero for \( j \neq j' \).

As \( n \to \infty \),

\[ - \frac{1}{n} \nabla \nabla' l(\alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_q) \to S \quad (20) \]

where \( S \) is a \( q(1 + L) \times q(1 + L) \) matrix with entries corresponding to \( j, j' = 1, \ldots, q \).
Consistency of the Semiparametric Regression

Let $x$ be a vector of size $k = L - 1$ of bounded covariates, and $y$ a bounded response. We wish to prove $\hat{E}(y|x) \xrightarrow{p} E(y|x)$. We have:

$$\hat{E}(y|x) - E(y|x) = \frac{1}{n} \sum_{i=1}^{n} y_i \hat{g}(x, y_i) - \frac{1}{n} \sum_{i=1}^{n} \hat{g}(x, y_i)$$

Or for sufficiently large $n$

$$\hat{E}(y|x) - E(y|x) = \frac{\int y \hat{g}(x, y)dy}{\hat{g}(x)} - \frac{\int yg(x, y)dy}{g(x)} \sim \frac{\int y[\hat{g}(x, y) - g(x, y)]dy}{g(x)}$$

Thus by Cauchy-Schwarz,

$$\left[ \int |\hat{E}(y|x) - E(y|x)| g(x) dx \right]^2 \sim \left[ \int \int |y[\hat{g}(x, y) - g(x, y)]dy | dx \right]^2 \leq \int \int |y||[\hat{g}(x, y) - g(x, y)]|dydx$$

$$\leq \int \int |y|^2 dydx \int \int |[\hat{g}(x, y) - g(x, y)]|^2 dydx$$

$$\leq C \int \int |[\hat{g}(x, y) - g(x, y)]|^2 dydx$$

From Fokianos (2004) and Qin and Zhang (2005) the MISE converges to 0. That is

$$E \left\{ \int \int |[\hat{g}(x, y) - g(x, y)]|^2 dydx \right\} \to 0$$

as $nh^k \to \infty$ and $h \to 0$. It follows that

$$\left[ \int (E|\hat{E}(y|x) - E(y|x)|)g(x) dx \right]^2 \leq E \left[ \int |\hat{E}(y|x) - E(y|x)|g(x) dx \right]^2 \to 0$$

and we have the convergence in mean $E|\hat{E}(y|x) - E(y|x)| \to 0$ which implies convergence in probability.

References

1. Anderson, T.W. (1971). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

2. Bondell, H.D. (2007). Testing goodness-of-fit in logistic case-control studies. *Biometrika*, 94, 487-495.
3. Cheng, K.F. and Chu, C.K. (2004). Semiparametric Density Estimation Under A Two-Sample Density Ratio Model. *Bernoulli*, 10, 583-604.

4. Fokianos, K., Kedem, B., Qin, J., Short, D.A. (2001). A semiparametric approach to the one-way layout. *Technometrics*, 43, 56-65.

5. Fokianos, K. (2004). Merging information for semiparametric density estimation. *Journal of the Royal Statistical Society, Series B*, 66, 941-958.

6. Gilbert, P.B., Lele, S.R., Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*, 86, 27-43.

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. (2004). Goodness-of-fit tests for semiparametric biased sampling models. *J. Statist. Plan. Infer.*, 118, 51-81.

9. Kedem, B., Lu, G., Wei, R., Williams, D. (2008). Forecasting mortality rates via density ratio modeling. *Canadian Jour. of Statistics*, 36, 193-206.

10. Kedem, B., Kim, E., Voulgaraki, A., Graubard, B.I. (2009). Two-dimensional semiparametric density ratio modeling of testicular germ cell data. *Statistics in Medicine*, 28, 2147-2159.

11. Lee, M-J. (1996). *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models*. Springer, New York.

12. Lu, G. (2007). *Asymptotic Theory for Multiple-sample Semiparametric Density Ratio Models and its Application to Mortality Forecasting*. Ph.D. Dissertation, University of Maryland, College Park, 2007.

13. McGlynn, K.A., Devesa, S.S., Sigurdson, A.J., Brown, L.M., Tsao, L., Tarone, R.E. (2003). Trends in the incidence of testicular germ cell tumors in the United States. *Cancer*, 97, 63-70.

14. McGlynn K.A., Sakoda L.C., Rubertone M.V., Sesterhenn I.A., Lyu C., Graubard B.I., Erickson R.L. (2007). Body size, dairy consumption, puberty, and risk of testicular germ cell tumors. *American Journal of Epidemiology*, 165, 355-363.
15. McGlynn, K.A., Cook, M.B. (2010). The Epidemiology of testicular cancer. In: *Male Reproductive Cancers: Epidemiology, Pathology and Genetics*, W.D. Foulkes and K.A. Cooney eds. Springer, New York, 2010, 51-83.

16. Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9, 141-142.

17. Nagelkerke, N. (1991). A Note on a General Definition of the Coefficient of Determination, *Biometrika*, 78, 691-692.

18. Ogden, C.L., Fryar, C.D., Carroll, M.D., Flegal, K.M. (2004). Mean body weight, height, and body mass index, United States 1960-2002. *Advance Data*, No. 347, 1-17.

19. Phue, J.N., Kedem, B., Jaluria, P., 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.

20. Prentice, R.L., Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.

21. Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85, 619-630.

22. Qin, J., Zhang, B. (1997). A goodness of fit test for logistic regression models based on case-control data. *Biometrika*, 84, 609-618.

23. Qin, J., and Zhang, B. (2005). Density Estimation Under A Two-Sample Semiparametric Model. *Nonparametric Statistics*, 1, 665-683.

24. Vardi, Y. (1982). Nonparametric estimation in the presence of length bias, *Annals of Statistics*, 10, 616-20.

25. Vardi, Y. (1985). Empirical distribution in selection bias models, *Annals of Statistics*, 13, 178-203.

Watson, G.S. (1964). Smooth regression analysis. *Sankhya A*, 26, 359-372.

26. Wen, S., and Kedem, B. (2009). A semiparametric cluster detection method - a comprehensive power comparison with Kulldorff’s method. *International Journal of Health Geographics*, 8:73 (31 December 2009). Online journal without page numbers.
27. Zhang, B. (1999). A chi-squared goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, **86**, 531-539.

28. Zhang, B. (2000). A goodness of fit test for multiplicative-intercept risk models based on case-control data. *Statistica Sinica*, **10**, 839-866.

29. Zhang B. (2001) An information matrix for logistic regression models based on case-control data. *Biometrika*, **88**, 921-932.

30. Zhang, B. (2002) Assessing Goodness-of-Fit of Generalized Logit Models Based on Case-Control Data. *Journal of Multivariate Analysis*, **82**, 17-38.