The Annals of Applied Statistics
2012, Vol. 6, No. 3, 1185–1208
DOI: 10.1214/12-AOAS552

SEMIPARAMETRIC REGRESSION IN TESTICULAR GERM CELL DATA

BY ANASTASIA VOULGARAKI[1], BENJAMIN KEDEM[1] AND BARRY I. GRAUBARD

University of Maryland, University of Maryland and National Cancer Institute

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. A kernel density estimator can be constructed from many data sources in conjunction with the semiparametric model. The estimator is shown to be more efficient than the traditional single-sample kernel density estimator, and its optimal bandwidth is discussed in some detail. Each multivariate distribution and the corresponding conditional expectation (regression) of interest are estimated from the combined data using all sources. Graphical and quantitative diagnostic tools are suggested to assess model validity. The method is applied in quantifying the effect of height and age on weight of germ cell testicular cancer patients. Comparisons are made with multiple regression, generalized additive models (GAM) and nonparametric kernel regression.

1. Introduction. This paper addresses the relationship between weight, height and age of germ cell testicular cancer patients. The problem is approached by a nonlinear regression method based on the so-called density ratio model. The method fuses or combines information from multiple sources in order to create an efficient kernel density estimator, which is then used in the direct estimation of the conditional expectation, bypassing linearity and the normal assumption. The choice of bandwidth parameters associated with the density kernel estimates is discussed in some detail.

In Section 2 we present the general multidimensional semiparametric density ratio model, review the procedure for estimating the distributions and parameters of the model, and discuss the asymptotic behavior of the estimators. In Section 3 we introduce the combined (from many samples) semiparametric multivariate kernel density estimator, and show that it is more efficient than the traditional single-sample kernel estimator. Moreover, we discuss the associated problem of bandwidth selection. Section 4 deals with a semiparametric approach to regression

Received September 2011; revised February 2012.
[1]Supported by NSF Grant DMS-10-07647.
Key words and phrases. Multivariate density ratio model, kernel, random covariates, diagnostic, Nadaraya–Watson, GAM.

with random covariates, that is, semiparametric estimation of $E[y|\mathbf{x}]$. The proposed estimator of $E[y|\mathbf{x}]$ may be viewed as a semiparametric extension of the Nadaraya–Watson nonparametric estimator. We also propose various diagnostic measures to check model validity. The method is first illustrated by a simulation study in Section 5 and is then applied in Section 6 to Testicular Germ Cell Tumor (TGCT) data. A comparison with other methods is made in both Sections 5 and 6.

1.1. *Motivation*. The p-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; see Kedem et al. (2009). Increased height has been shown to be associated with increased risk of germ cell testicular cancer in a majority of studies, reflecting exposure to, possibly, early life factors due to genetics, nutrition or endogenous or exogenous hormones; see McGlynn and Cook (2010). Body weight reflects potentially later life exposures such as dietary intake and energy expenditure behavior. A few studies have found that increased body mass (body weight divided by height squared) was associated with a decrease in risk of testicular cancer, but most studies have found no association [McGlynn and Cook (2010)]. This lack of association may be due to inappropriate parametric modeling, usually logistic regression. 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; see McGlynn et al. (2007). 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; see McGlynn and Cook (2010) and Ogden et al. (2004). 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.

1.2. *Background and preliminaries*. Suppose there are $m = q + 1$ data sources, such as q case groups and a control group, each giving a sample of random vectors from an unknown multivariate distribution. In the density ratio model one of these distributions serves as a reference or baseline, and all other distributions are 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; see Fokianos et al. (2001) and Qin and Zhang (1997). In its multivariate form, the model is motivated by classical classification given multivariate normal samples, and generalized logistic regression; see Anderson (1971) and Prentice and Pyke (1979).

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,

$$x_{ij} \sim g_i, \qquad i = 1, \ldots, q, m,\ j = 1, \ldots, n_i, \tag{1}$$

where $g_m \equiv g$ is called the *reference* probability density. Assuming exponential tilts, the g_i satisfy the (exponential) *density ratio model*

$$\frac{g_j(x)}{g(x)} = \exp\bigl(\alpha_j + \boldsymbol{\beta}_j' \mathbf{h}(x)\bigr), \qquad j = 1, \ldots, q. \tag{2}$$

It is assumed that the *distortion function* $\mathbf{h}(x)$ is a known vector-valued function. The objective is to estimate the reference density g, the corresponding cumulative distribution function (CDF) G and the parameters α_j, β_j from the combined data

$$\mathbf{t} = \{(x_{11}, \ldots, x_{1n_1}), \ldots, (x_{q1}, \ldots, x_{qn_q}), (x_{m1}, \ldots, x_{mn_m})\}'. \tag{3}$$

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, Lele and Vardi (1999)], mortality rate prediction [Kedem et al. (2008)], microarrays evaluation [Phue et al. (2007)], 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 asymptotic results for the semiparametric kernel density estimator and the estimation of the conditional expectation of a response given covariate information are formulated under the general multiple sample p-dimensional density ratio model. Specifically, for each of the m data sources, we use the p-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 model validity. A comparison with linear multiple regression, generalized additive models (GAM) and the Nadaraya–Watson kernel nonparametric regression is made using both real and simulated data.

2. Statistical formulation. Suppose we have $m = q + 1$ independent data sets or random samples of p-dimensional vectors $\mathbf{x} = \mathbf{x}_{p\times 1} = (x_1, x_2, \ldots, x_p)'$. Let $g_i(x_1, x_2, \ldots, x_p)$ be the probability function corresponding to the ith sample. Assume that the ith sample size is n_i and $n = \sum_{i=1}^{m} n_i$ is the total sample size. Thus, for $i = 1, \ldots, q, m$, $j = 1, \ldots, n_i$, we have that

$$\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp}) \sim g_i(x_1, \ldots, x_p)$$

and

$$\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i} \overset{\text{i.i.d.}}{\sim} g_i,$$

where $\mathbf{x}_{ij}, \mathbf{x}_{ij'}$ are independent for $j \neq j'$ and $\mathbf{x}_{ij}, \mathbf{x}_{i'k}$ are independent for $i \neq i'$ and all j and k. We choose $\mathbf{x}_{mj}$ as the reference sample. Then $g \equiv g_m(\mathbf{x}) \equiv g_m(x_1, \ldots, x_p)$ is called the reference or baseline probability density function (p.d.f.). We assume that the $g_i(\mathbf{x}), i = 1, \ldots, q$, satisfy the (general) *density ratio model*:

$$\frac{g_i(\mathbf{x})}{g_m(\mathbf{x})} = w(\mathbf{x}, \boldsymbol{\theta}_i) \tag{4}$$

or, equivalently,

$$g_i(\mathbf{x}) = w(\mathbf{x}, \boldsymbol{\theta}_i) g_m(\mathbf{x}), \tag{5}$$

where $g_i(\mathbf{x})$, $g_m(\mathbf{x})$ are not specified, w is a known positive and continuous function, and the $\boldsymbol{\theta}_i$ are unknown d-dimensional vectors of parameters. This construction accommodates both continuous and discrete distributions, and it does not require symmetry, let alone normality in the continuous case.

Let $G(\mathbf{x}) \equiv G_m(\mathbf{x})$ denote the reference cdf and define $p_{ij} = dG(\mathbf{x}_{ij}) = dG_m(\mathbf{x}_{ij})$. Using the method of constrained empirical likelihood, we can estimate g_i and $\boldsymbol{\theta}_i$ from the entire combined data, and not just from the corresponding samples $\mathbf{x}_{ij}$ and $\mathbf{x}_{mj}$; see Fokianos (2004). The empirical likelihood based on the pooled data $\mathbf{x}_{ij}, i = 1, \ldots, m, j = 1, \ldots, n_i$, is

$$\begin{aligned} L(\boldsymbol{\theta}, G_m) &= \left[\prod_{j=1}^{n_1} p_{1j} w(\mathbf{x}_{1j}, \boldsymbol{\theta}_1)\right]\left[\prod_{j=1}^{n_2} p_{1j} w(\mathbf{x}_{2j}, \boldsymbol{\theta}_2)\right] \cdots \left[\prod_{j=1}^{n_m} p_{mj}\right] \\ &= \left[\prod_{i=1}^{m}\prod_{j=1}^{n_i} p_{ij}\right]\left[\prod_{i=1}^{q}\prod_{j=1}^{n_i} w(\mathbf{x}_{ij}, \boldsymbol{\theta}_i)\right]. \end{aligned} \tag{6}$$

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_q')'$, a vector of dimension of qd. The log-likelihood is given by

$$l = \log L = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log(p_{ij}) + \sum_{i=1}^{q}\sum_{j=1}^{n_i} \log(w(\mathbf{x}_{ij}, \boldsymbol{\theta}_i)) \tag{7}$$

and is subject to the constraints

$$\begin{aligned} &p_{ij} \geq 0, \qquad \sum_{i=1}^{m}\sum_{j=1}^{n_i} p_{ij} = 1, \qquad \sum_{i=1}^{m}\sum_{j=1}^{n_i} p_{ij} w(\mathbf{x}_{ij}, \boldsymbol{\theta}_k) = 1 \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } k = 1, \ldots, q. \end{aligned} \tag{8}$$

Fokianos (2004) and Qin and Lawless (1994) gave conditions guaranteeing that, with probability approaching 1, there is a maximum in a small neighborhood of the true parameter $\boldsymbol{\theta}_0$. Define $\mu_k \equiv \frac{\lambda_k}{n}$, where λ_k are the Lagrange multipliers. Then,

replacing μ_k and $\boldsymbol{\theta}_k$ by their estimators, p_{ij} and $G_m(x)$ are estimated by

$$\hat{p}_{ij} = \frac{1}{n} \frac{1}{1 + \sum_{k=1}^{q} \hat{\mu}_k [w(\mathbf{x}_{ij}, \hat{\boldsymbol{\theta}}_k) - 1]}, \tag{9}$$

$$\begin{aligned} \hat{G}_m(\mathbf{x}) &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{p}_{ij} I(\mathbf{x}_{ij} \le \mathbf{x}) \\ &= \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{I(\mathbf{x}_{ij} \le \mathbf{x})}{1 + \sum_{k=1}^{q} \hat{\mu}_k [w(\mathbf{x}_{ij}, \hat{\boldsymbol{\theta}}_k) - 1]}, \end{aligned} \tag{10}$$

where $I(B)$ is the indicator of the event B, and $I(\mathbf{x}_{ij} \le \mathbf{x})$ is defined componentwise. More generally, for $l = 1, \ldots, m$ and $w(\mathbf{x}_{ij}, \hat{\boldsymbol{\theta}}_m) \equiv 1$,

$$\begin{aligned} \hat{G}_l(\mathbf{x}) &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{p}_{ij} w(\mathbf{x}_{ij}, \hat{\boldsymbol{\theta}}_l) I(\mathbf{x}_{ij} \le \mathbf{x}) \\ &= \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{w(\mathbf{x}_{ij}, \hat{\boldsymbol{\theta}}_l)}{1 + \sum_{k=1}^{q} \hat{\mu}_k [w(\mathbf{x}_{ij}, \hat{\boldsymbol{\theta}}_k) - 1]} I(\mathbf{x}_{ij} \le \mathbf{x}). \end{aligned} \tag{11}$$

Let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$ under model (4). Define the sample size ratios $\rho_i = n_i / n_m$ and set $w(\mathbf{x}, \hat{\boldsymbol{\theta}}_i) = w_i(\mathbf{x})$ for $i = 1, \ldots, m$. Then $\rho_m \equiv 1$, $w_m(\mathbf{x}) \equiv 1$. We assume the ρ_i are positive and finite and remain fixed as $n \to \infty$. Let $\boldsymbol{\zeta}$ denote the true value of $\boldsymbol{\mu}$. Set $\boldsymbol{\zeta}_n = (\zeta_{1n}, \ldots, \zeta_{qn})$ and $\zeta_{ln} = n_l / n$ for $l = 1, \ldots, q$. As $n \to \infty$, assume that $\zeta_{ln} \to \zeta_l$. Then Fokianos (2004) showed that $\boldsymbol{\zeta}_n \to \boldsymbol{\zeta}$ and that under regularity conditions $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\mu}} - \boldsymbol{\zeta}$ are jointly asymptotically normal. The complete statement is Theorem 1 in an Appendix in Voulgaraki, Kedem and Graubard (2012).

3. Combined semiparametric density estimators. Fokianos (2004), Cheng and Chu (2004) and Qin and Zhang (2005) constructed a kernel-based density estimator by smoothing the increments of $\hat{G}_i$, $i = 1, \ldots, m$. Fokianos (2004) studied the statistical properties of the proposed kernel density estimator (mean, variance) and showed that combining data leads to more efficient kernel density estimators under the univariate case of the general model (4). Qin and Zhang (2005) studied semiparametric inference for the univariate version of model (4) with $w(x, \alpha, \beta) = \exp(\alpha + r(x)\beta)$. Cheng and Chu (2004) studied the same special case as Qin and Zhang (2005) but followed a different approach.

In this section we aim to study the corresponding asymptotic theory and convergence properties of the proposed kernel density estimator for the general multivariate multiple-sample case model (4). The estimator is shown to be more efficient than the traditional kernel density estimator. In addition, several methods for calculating the optimal bandwidth are discussed. Precise statements and proofs are given in Voulgaraki, Kedem and Graubard (2012).

The traditional kernel density estimator is a convolution of the jumps in the empirical distribution function obtained from a single sample of size n and a kernel function taken as a symmetric probability density function parametrized by a bandwidth parameter [Parzen (1962)]. Specifically, the traditional kernel density estimator of a probability density $f(\mathbf{x})$ is given by

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_n^p} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right), \tag{12}$$

where h_n is a sequence of bandwidths such that $h_n \to 0$ and $nh_n^p \to \infty$ as $n \to \infty$. The kernel function $K(\mathbf{x})$ is defined for p-dimensional $\mathbf{x}$. It is nonnegative, symmetric around $\mathbf{0}$ and satisfies $\int_{\mathbf{R}^p} K(\mathbf{x})\, d\mathbf{x} = 1$. Under certain conditions, $\hat{f}(\mathbf{x})$ is a consistent estimator of $f(\mathbf{x})$ [Parzen (1962), Shao (2003)]. As such, the traditional kernel density estimator is a "single sample" estimator.

Using a similar idea to (12), we use the the probabilities $\hat{p}_{ij}$ in (9) to form kernel estimates for the probability densities $g_l(\mathbf{x})$,

$$\hat{g}_l(\mathbf{x}) = \frac{1}{h_n^p} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{p}_{ij} \hat{w}_l(\mathbf{x}_{ij}) K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right), \tag{13}$$

where h_n is a sequence of bandwidths such that $h_n \to 0$ and $nh_n^p \to \infty$ as $n \to \infty$, $w_l(\mathbf{x}) \equiv w(\mathbf{x}, \boldsymbol{\theta}_l)$, $\hat{w}_l(\mathbf{x}) \equiv w(\mathbf{x}, \hat{\boldsymbol{\theta}}_l)$, and K is a nonnegative kernel function that satisfies the following requirements:

(1) $\int K(\mathbf{x})\, d\mathbf{x} = 1$ and $\int |K(\mathbf{x})|\, d\mathbf{x} < \infty$;
(2) $\int \mathbf{x} K(\mathbf{x})\, d\mathbf{x} = \mathbf{0}$ and $\int |\mathbf{x} K(\mathbf{x})|\, d\mathbf{x} < \infty$;
(3) $\int \mathbf{x}'\mathbf{x} K(\mathbf{x})\, d\mathbf{x} = k_2$ and $\int |\mathbf{x}'\mathbf{x} K(\mathbf{x})|\, d\mathbf{x} < \infty$.

It is easy to verify that $\hat{g}_l$ is a proper probability function.

3.1. *Asymptotic results for* $\hat{g}_l$. To facilitate the study of $\hat{g}_l$, it is convenient to define first $\tilde{g}_l(\mathbf{x})$:

$$\tilde{g}_l(\mathbf{x}) = \frac{1}{h_n^p} \sum_{i=1}^{m} \sum_{j=1}^{n_i} p_{ij} w_l(\mathbf{x}_{ij}) K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right). \tag{14}$$

With this device, and with the help of Lemmas 1–4 and Theorem 2 in Voulgaraki, Kedem and Graubard (2012), in Corollary 1 in there it is shown that

$$\sqrt{nh_n^p}\left(\hat{g}_l(\mathbf{x}) - g_l(\mathbf{x}) - \frac{1}{2}h_n^2 \int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x}^*)}{\partial \mathbf{x}\, \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u})\, d\mathbf{u}\right) \xrightarrow{D} N(\mathbf{0}, \sigma^2(\mathbf{x}))$$

as $n \to \infty$, where

$$\sigma^2(\mathbf{x}) = \frac{w_l(\mathbf{x}) g_l(\mathbf{x})}{\sum_{k=1}^{m} \zeta_k w_k(\mathbf{x})} \int K^2(\mathbf{u})\, d\mathbf{u}$$

for any fixed $\mathbf{x}$.

3.2. *Comparison of $\hat{g}_l$ and the traditional $\hat{f}$.* In Theorem 3 in Voulgaraki, Kedem and Graubard (2012) we show that as $n \to \infty$, $h_n \to 0$, and $nh_n^p \to \infty$,

$$\begin{aligned}\mathrm{MISE}(\hat{g}_l) = {} & \frac{1}{nh_n^p} \int \frac{w_l(\mathbf{x}) g_l(\mathbf{x})}{\sum_{k=1}^m \zeta_k w_k(\mathbf{x})}\, d\mathbf{x} \int K^2(\mathbf{u})\, d\mathbf{u} \\ & + \frac{h_n^4}{4} \int \left(\int \mathbf{u}' \frac{\partial^2 g_l(\mathbf{x})}{\partial \mathbf{x}\, \partial \mathbf{x}'} \mathbf{u} K(\mathbf{u})\, d\mathbf{u} \right)^2 d\mathbf{x} \\ & + o\left(\frac{1}{nh_n^p} \right) + o(h_n^4),\end{aligned}$$

from which we get the optimal bandwidth h_n^* given in formula (4) in Voulgaraki, Kedem and Graubard (2012). In Theorem 4 there it is shown that under mild conditions $\hat{g}_l$ is more efficient (MISE) than the traditional single-sample $\hat{f}$ for every l, as $n \to \infty$, $h_n \to 0$, and $nh_n^p \to \infty$.

3.3. *Bandwidth selection for $\hat{g}_l$.* From Section 3.1 we see that, as is the case with the traditional single-sample estimator, the pooled estimator $\hat{g}_l$ also suffers from a similar bias-variance trade-off problem where a smaller h_n reduces the bias at the expense of the variance, whereas a larger h_n increases the bias but reduces the variance. We discuss next practical ways for choosing bandwidths which are optimal in some sense.

The formula for the asymptotically optimal bandwidth h_n^* given in equation (4) in Voulgaraki, Kedem and Graubard (2012) is not practical since g_l is not known. In the one-dimensional case Silverman (1986) proposes to either use the normal density $N(\mu, \Sigma)$, where μ and Σ are estimated from the data, or $\hat{f}$ to approximate g_l. Following Silverman (1986), Fokianos (2004) and Qin and Zhang (2005), both replace g_l by $\hat{g}_l$. However, in the multidimensional setting the computational burden is heavier and, as Silverman (1986) remarks, it is somewhat hazardous to estimate $\partial^2 g_l(\mathbf{x})/\partial \mathbf{x}\, \partial \mathbf{x}'$ by $\partial^2 \hat{g}_l(\mathbf{x})/\partial \mathbf{x}\, \partial \mathbf{x}'$ unless very large samples are available.

The bandwidth can also be selected via *cross-validation*, which minimizes, with respect to h_n, an estimate for the integrated squared error (ISE):

$$\begin{aligned}\mathrm{ISE}(h_n) &= \int \left(\hat{g}_l(\mathbf{x}) - g_l(\mathbf{x}) \right)^2 d\mathbf{x} \\ &= \int \hat{g}_l^2(\mathbf{x})\, d\mathbf{x} - 2 \int \hat{g}_l(\mathbf{x}) g_l(\mathbf{x})\, d\mathbf{x} + \int g_l^2(\mathbf{x})\, d\mathbf{x}.\end{aligned}$$

The last term does not depend on h_n, so we may drop it in the minimization of ISE. To minimize ISE, we need to rewrite the first and second terms as functions of h_n and the data. Denote by

$$\mathbf{t} = [\mathbf{x}'_{11}, \ldots, \mathbf{x}'_{1n_1}, \ldots, \mathbf{x}'_{m1}, \ldots, \mathbf{x}'_{mn_m}]'_{n \times 1} = (\mathbf{t}'_1, \ldots, \mathbf{t}'_n)'$$

the combined data. So $\mathbf{t}$ has n rows. The first term can be written

$$\begin{aligned}\int \hat{g}_l^2(\mathbf{x})\,d\mathbf{x} &= \int\left[\frac{1}{h_n^p}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{p}_{ij}\hat{w}_l(\mathbf{x}_{ij})K\left(\frac{\mathbf{x}-\mathbf{x}_{ij}}{h_n}\right)\right]^2 d\mathbf{x}\\ &= \frac{1}{h_n^{2p}}\int\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{i'=1}^{m}\sum_{j'=1}^{n_i}\hat{p}_{ij}\hat{w}_l(\mathbf{x}_{ij})K\left(\frac{\mathbf{x}-\mathbf{x}_{ij}}{h_n}\right)\\ &\qquad\times\hat{p}_{i'j'}\hat{w}_l(\mathbf{x}_{i'j'})K\left(\frac{\mathbf{x}-\mathbf{x}_{i'j'}}{h_n}\right)d\mathbf{x}\\ &= h_n^{-p}\sum_{i=1}^{n}\sum_{i'=1}^{n}\hat{p}(\mathbf{t}_i)\hat{w}_l(\mathbf{t}_i)\hat{p}(\mathbf{t}_{i'})\hat{w}_l(\mathbf{t}_{i'})\int K(\mathbf{z})K\left(\mathbf{z}+\frac{\mathbf{t}_i-\mathbf{t}_{i'}}{h_n}\right)d\mathbf{z}.\end{aligned}$$

For the second term notice that $\int \hat{g}_l(\mathbf{x})g_l(\mathbf{x})\,d\mathbf{x} = E\hat{g}_l(\mathbf{x})$. Following Silverman (1986) and Cheng and Chu (2004), we can estimate $E\hat{g}_l(\mathbf{x})$ using the leave one out estimator $\widehat{E\hat{g}_l(\mathbf{x})}$,

$$\widehat{E\hat{g}_l(\mathbf{x})} = \frac{1}{n_l}\sum_{i=n_1+\cdots+n_{l-1}+1}^{n_l}\hat{g}_{l,i}(\mathbf{t}_i),$$

where $\hat{g}_{l,i}(\mathbf{t}_i)$ is $\hat{g}_l(\mathbf{t}_i)$ with $\mathbf{t}_i$ dropped from the combined data. Therefore, a nearly optimal bandwidth h_n is obtained by minimizing

$$\begin{aligned}&h_n^{-p}\sum_{i=1}^{n}\sum_{i'=1}^{n}\hat{p}(\mathbf{t}_i)\hat{w}_l(\mathbf{t}_i)\hat{p}(\mathbf{t}_{i'})\hat{w}_l(\mathbf{t}_{i'})\int K(\mathbf{z})K\left(\mathbf{z}+\frac{\mathbf{t}_i-\mathbf{t}_{i'}}{h_n}\right)d\mathbf{z}\\ &\quad-\frac{2}{n_l}\sum_{i=n_1+\cdots+n_{l-1}+1}^{n_l}\hat{g}_{l,i}(\mathbf{t}_i).\end{aligned}\tag{15}$$

In general, cross-validation using the leave one out estimator is computationally inefficient. However, for sufficiently large samples and $l = 1,\ldots,q,m$, a useful simplification is obtained from the approximation

$$\int \hat{g}_l(\mathbf{x})g_l(\mathbf{x})\,d\mathbf{x}\overset{n\to\infty}{\to}\int \tilde{g}_l(\mathbf{x})g_l(\mathbf{x})\,d\mathbf{x}.$$

Moreover,

$$\begin{aligned}&E\left[\int \tilde{g}_l(\mathbf{x})g_l(\mathbf{x})\,d\mathbf{x}\right]\\ &\quad= E\left[\int\frac{1}{nh_n^p}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{w_l(\mathbf{x}_{ij})}{\sum_{k=1}^{m}\zeta_k w_k(\mathbf{x}_{ij})}K\left(\frac{\mathbf{x}-\mathbf{x}_{ij}}{h_n}\right)g_l(\mathbf{x})\,d\mathbf{x}\right]\\ &\quad= h_n^{-p}\int\int K\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right)g_l(\mathbf{x})g_l(\mathbf{y})\,d\mathbf{x}\,d\mathbf{y}\end{aligned}$$

$$= E\left[h_n^{-p} K\left(\frac{\mathbf{x}-\mathbf{y}}{h_n}\right)\right]$$
$$= E\left[\frac{1}{n_l(n_l-1)h_n^p}\sum_{i\neq j} K\left(\frac{\mathbf{x}_{li}-\mathbf{x}_{lj}}{h_n}\right)\right].$$

Thus, for sufficient large n, an unbiased estimator for $\int \tilde{g}_l(\mathbf{x})g_l(\mathbf{x})\,d\mathbf{x}$ is

$$\frac{1}{n_l(n_l-1)h_n^p}\sum_{i\neq j} K\left(\frac{\mathbf{x}_{li}-\mathbf{x}_{lj}}{h_n}\right).$$

Therefore, an alternative way to find h_n is by minimizing

$$h_n^{-p}\sum_{i=1}^{n}\sum_{i'=1}^{n}\hat{p}(\mathbf{t}_i)\hat{w}_l(\mathbf{t}_i)\hat{p}(\mathbf{t}_{i'})\hat{w}_l(\mathbf{t}_{i'})\int K(\mathbf{z})K\left(\mathbf{z}+\frac{\mathbf{t}_i-\mathbf{t}_{i'}}{h_n}\right)d\mathbf{z}$$
$$(16)\qquad -\frac{2}{n_l(n_l-1)h_n^p}\sum_{i\neq j} K\left(\frac{\mathbf{x}_{li}-\mathbf{x}_{lj}}{h_n}\right).$$

Cross-validation has the advantage that (15) and (16) can easily be modified if we wish to use different bandwidths $h_1,\ldots,h_p$ to smooth each term, respectively.

4. Semiparametric regression. Suppose we have $m = q+1$ data sets or samples of p-dimensional vectors, where each vector consists of $p-1$ covariates and one response, and assume that the ith sample size is n_i. Thus, for $i = 1,\ldots,q,m$, $j = 1,\ldots,n_i$, we have

$$(x_{ij1}, x_{ij2},\ldots,x_{ij(p-1)}, y_{ij}) \sim g_i(x_1,\ldots,x_{(p-1)}, y).$$

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

$$(17)\qquad \frac{g_i(\mathbf{x})}{g(\mathbf{x})} = \exp(\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}), \qquad i = 1,\ldots,q,$$

where $\mathbf{x} = (x_1,\ldots,x_{(p-1)}, y)'$ and $\boldsymbol{\beta}_i = (\beta_{i1},\ldots,\beta_{ip})'$. Since the $g_i(\mathbf{x})$, $i = 1,\ldots,q,m$, are probability densities, $\boldsymbol{\beta}_i = \mathbf{0}$ implies $\alpha_i = 0$, $j = 1,\ldots,q$. It follows that the hypothesis $H_0 : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_q = \mathbf{0}$ implies equidistribution: all the g_i are equal.

REMARK 1. Model (17) is motivated from the ratio of two multivariate normal densities assuming the same covariance matrices [Anderson (1971), Kedem et al. (2009)]. It is a special case of model (4) with $w(\mathbf{x},\boldsymbol{\theta}_i) = w(\mathbf{x},\alpha_i,\boldsymbol{\beta}_i) \equiv \exp(\alpha_i + \boldsymbol{\beta}_i'\mathbf{x})$.

Let $\mathbf{t}$ denote the vector of combined data of length $n = n_1 + n_2 + \cdots + n_m$. Following the method of constrained empirical likelihood, we obtain score equations for $\hat{\alpha}_j$ and $\hat{\boldsymbol{\beta}}_j$:

$$\frac{\partial l}{\partial \alpha_j} = -\sum_{i=1}^{n} \frac{\rho_j w_j(\mathbf{t}_i)}{1 + \rho_1 w_1(\mathbf{t}_i) + \cdots + \rho_q w_q(\mathbf{t}_i)} + n_j = 0, \tag{18}$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}_j} = -\sum_{i=1}^{n} \frac{\rho_j w_j(\mathbf{t}_i)\mathbf{t}_i}{1 + \rho_1 w_1(\mathbf{t}_i) + \cdots + \rho_q w_q(\mathbf{t}_i)} + \sum_{i=1}^{n_j} (x_{ji1}, \ldots, y_{ji})' = 0 \tag{19}$$

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

$$\hat{p}_i = \frac{1}{n_m} \cdot \frac{1}{1 + \rho_1 \hat{w}_1(\mathbf{t}_i) + \cdots + \rho_q \hat{w}_q(\mathbf{t}_i)}, \tag{20}$$

$$\hat{G}(\mathbf{t}) = \frac{1}{n_m} \cdot \sum_{i=1}^{n} \frac{I(\mathbf{t}_i \le \mathbf{t})}{1 + \rho_1 \hat{w}_1(\mathbf{t}_i) + \cdots + \rho_q \hat{w}_q(\mathbf{t}_i)}, \tag{21}$$

where $(\mathbf{t}_i \le \mathbf{t})$ is defined componentwise, $\hat{w}_j(\mathbf{t}_i) = \exp(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}_j' \mathbf{t}_i)$, and $I(B)$ is the indicator of the event B. Following Lu (2007), we can show that as $n \to \infty$ the estimators $\hat{\boldsymbol{\theta}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_q, \hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_q)'$ are asymptotically normal.

4.1. *Computing $E[y|\mathbf{x}]$ using the density ratio model.* Under the p-dimensional density ratio model we can predict the response y given the covariate information $x_1, x_2, \ldots, x_{(p-1)}$ for any of the m data sets as follows:

$$\hat{E}_j(y|x_1, \ldots, x_{(p-1)}) = \sum_{i}^{n_j} y_i \frac{\hat{g}_j(x_1, \ldots, x_{(p-1)}, y_i)}{\sum_{y_i} \hat{g}_j(x_1, \ldots, x_{(p-1)}, y_i)}, \quad j = 1, \ldots, q, m. \tag{22}$$

The $\hat{g}_j$ in (22) are the semiparametric kernel density estimates. Theorem 5 in Voulgaraki, Kedem and Graubard (2012) establishes the consistency of (22) under some conditions.

It is interesting to compare the semiparametric estimator for $E[y|\mathbf{x}]$ against the Nadaraya–Watson estimator [Nadaraya (1964), Watson (1964)] and the estimators obtained from linear regression [Rencher (2000)], and GAM [Hastie and Tibshirani (1990), Wood (2006)].

4.2. *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 (22). 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 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 p-dimensional points as to obtain two-dimensional plots. Figures 1 and 2 in the next section are examples of this. We refer to these plots as diagnostic plots.

We found the following measure of goodness of fit useful. Consider the ith sample of size n_i. Let x_α be the number of times the estimated semiparametric cdf falls in the estimated $1 - \alpha$ confidence interval obtained from the corresponding empirical cdf, both evaluated at the sample points. Define

$$R^2_{\alpha,k} = 1 - \exp\left\{-\left(\frac{x_\alpha}{n_i - x_\alpha}\right)^k\right\}, \tag{23}$$

where $k > 0$, and k and α are free parameters, which can be set by the user. Observe that:

- $R^2_{\alpha,k}$ takes values between 0 and 1, being close to 1 when x_α approaches n_i and close to 0 when x_α is close to 0.
- $R^2_{\alpha,k}$ is a flexible criterion that can be adjusted by changing the parameters α and k. Larger α means smaller confidence interval bounds.
- Computing $R^2_{\alpha,k}$ is both simple and fast.

We now describe three natural alternatives to $R^2_{\alpha,k}$. First, as in multiple regression, goodness of fit may be approached by residual analysis. In this vein, we define "R^2" as in linear regression:

$$R^2_1 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}. \tag{24}$$

Next, define

$$R^2_2 = \operatorname{corr}(y, \hat{y})^2. \tag{25}$$

Notice that R^2_1 and R^2_2 depend on $\hat{y}$, and the process of calculating $\hat{y}$ involves selecting the bandwidth, making the process of calculating them complicated. In addition, some early simulation results suggested that they are misleading as measures of goodness of fit, and, thus, they were rejected.

Next, following Qin and Zhang (1997), define

$$R^2_3 = \exp(-\sqrt{n} \cdot \max|\tilde{G}_i - \hat{G}_i|). \tag{26}$$

Clearly, R^2_3 takes values between 0 and 1. Alternatives to R^2_3 are $\exp(-\sqrt{n} \cdot \text{median}|\tilde{G}_i - \hat{G}_i|)$ or $\exp(-\frac{1}{n}\sum|\tilde{G}_i - \hat{G}_i|^2)$.

The following simulation study compares $R^2_{\alpha,k}$ and R^2_3. The simulation suggests that $R^2_{\alpha,k}$ is a useful indicator of goodness of fit.

5. Some simulation results. In the present simulation study $m = 2$, g_2 denotes the reference distribution, and the results were obtained from 100 runs (repetitions) of the following four bivariate cases:

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

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

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

(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 and 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 (21)] versus the empirical cdf $\tilde{G}_1$ and $\tilde{G}_2$, respectively, all obtained from one run of the simulated case-control data, and evaluated at selected points in $\mathbb{R}^2$. As expected, in cases (1) and (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).

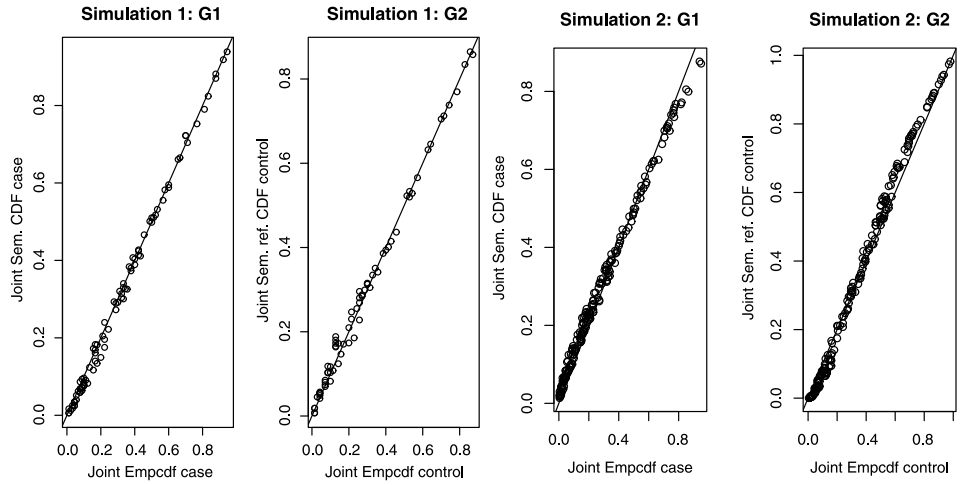

FIG. 1. *Case-control plots of* $\hat{G}_i$ *vs.* $\tilde{G}_i$, $i = 1, 2$, *simulations* (1) *and* (2).

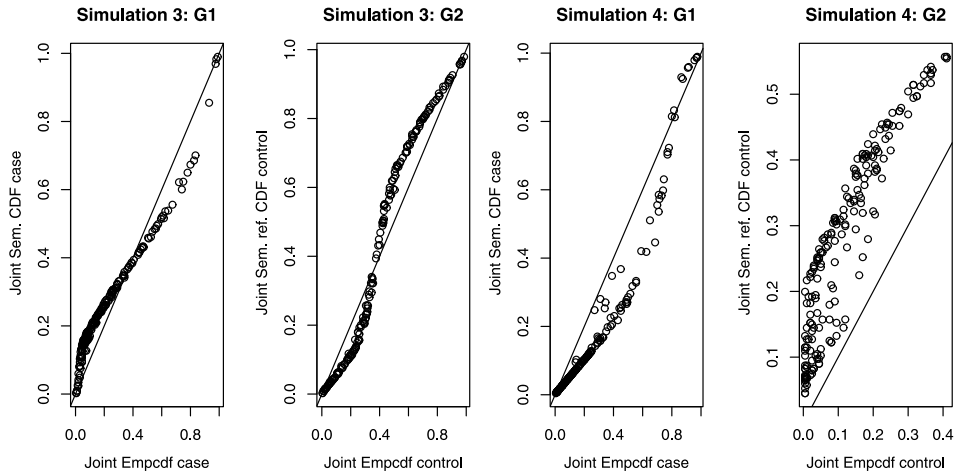

FIG. 2. *Case-control plots of* $\hat{G}_i$ *vs.* $\tilde{G}_i$, $i = 1, 2$, *simulations* (3) *and* (4).

A comparison of R_3^2 and $R_{\alpha,k}^2$ obtained from 100 runs is given in Table 1. It seems that R_3^2 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 $R_{\alpha,k}^2$ classifies correctly the four cases, giving high values for simulations (1) and (2) and low values for (3) and (4). The values of $R_{\alpha,k}^2$ in Table 1 were calculated with $k = 2$ and $1 - \alpha = 95\%$. We observed that, by lowering $1 - \alpha$, $R_{\alpha,k}^2$ gets closer to 0 for cases (3) and (4).

As noted earlier, calculating the semiparametric $\hat{E}[Y|X]$ for cases (1) and (2) entails bandwidth selection, which can be done either via the asymptotically optimal formula (4) in Voulgaraki, Kedem and Graubard (2012), replacing g_l with $N(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ (parameters estimated from the data), or via cross-validation and min-

TABLE 1
Comparison of R_3^2 *and* $R_{0.05,2}^2$ *for* 100 *repetitions of case and control*

Run	Group	R_3^2	$R_{0.05,2}^2$
(1)	Case	0.6307	1
	Control	0.5976	1
(2)	Case	0.3912	0.9353
	Control	0.3766	0.9718
(3)	Case	0.1080	0.3342
	Control	0.1129	0.3324
(4)	Case	0.0507	0.3361
	Control	0.0495	0.0033

TABLE 2
Bandwidth (BW) selection using formula (4) in Voulgaraki, Kedem and Graubard (2012)

	Case BW	Control BW
Simulation 1	0.46	0.47
Simulation 2	0.33	0.51

imize either (15) or (16) (which also allows different bandwidths $h_1, \ldots, h_p$ to smooth the different terms). Tables 2–4 summarize the results for the estimated bandwidths for one run of the simulations, using equations (4) in Voulgaraki, Kedem and Graubard (2012), (15) and (16). The integrals in (4) in Voulgaraki, Kedem and Graubard (2012) were calculated using *Mathematica*. There were no significant differences in the regression results using single or multiple bandwidths.

Using the semiparametric model, the standard normal distribution for the kernel and (22), we estimated $E[Y|X]$ for a single predictor. Table 5 provides MSE and MAE comparisons between the different methods for the first two simulations. In the table SP stands for semiparametric regression, MR for multiple regression, GAM for generalized additive model and NW for Nadaraya–Watson. We did not estimate $E[Y|X]$ for simulations 3 and 4 because the semiparametric model is not applicable in these cases (and was rejected as we saw from the R^2 comparisons). In simulations 1–2, for both case and control, we fitted a thin plate regression spline GAM assuming the normal distribution and identity link. The results for tensor product were almost identical. In simulation 1 the GAM line was almost identical to the multiple regression line. We see that the semiparametric regression performs comparably with the other methods in terms of MSE and MAE.

6. 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

TABLE 3
Bandwidth (BW) selection using the cross-validation method (15)

	Case			Control		
	Same BW	Diff. BWs		Same BW	Diff. BWs	
	h	h_1	h_2	h	h_1	h_2
Simulation 1	0.61	0.90	0.40	0.59	0.31	0.61
Simulation 2	0.38	0.50	0.20	0.61	0.36	0.71

TABLE 4
Bandwidth (BW) selection using the cross-validation method (16)

	Case			Control		
	Same BW	Diff. BWs		Same BW	Diff. BWs	
	h	h_1	h_2	h	h_1	h_2
Simulation 1	0.64	0.90	0.50	0.63	0.21	0.71
Simulation 2	0.30	0.40	0.20	0.74	0.11	0.96

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. We considered two cases: the 2D TGCT data set with variables height and weight and the 3D TGCT data set with variables height, weight and age. In both cases the control distribution was the reference distribution.

Equation (4) in Voulgaraki, Kedem and Graubard (2012), (15), (16) with kernel $K = N(\mathbf{0}, \mathbf{1})$ and $w(\mathbf{x}, \boldsymbol{\theta}_i) \equiv \exp(\alpha_i + \boldsymbol{\beta}_i'\mathbf{x})$ were used to calculate the different bandwidths. The three methods gave similar results. For the 2D TGCT data set, the case bandwidths were 1.01 and 3.51 for height and weight, respectively, whereas, for control, we used 2.02 and 1.01. For the 3D TGCT the bandwidths were 2.24 for control and 2.5 for case.

Before applying the three-dimensional density ratio model to the TGCT data, it is interesting to apply the two-dimensional model to get a prediction of weight given height only. As Figure 3 shows, the density ratio model is a suitable model for the TGCT data: there is almost a perfect agreement between the plots of the semiparametric $\hat{G}_i$ and the corresponding empirical $\tilde{G}_i$, $i = 1, 2$. The value of $R^2_{0.20,1}$ is 1 for both case and control. Figure 4 shows the estimated $E[Y|X]$ using

TABLE 5
MAE and MSE comparison between regression methods, for simulations 1 and 2. G_1, G_2 signify case and control, respectively

		MSE				MAE			
		SP	MR	GAM	NW	SP	MR	GAM	NW
Simulation 1	G_1	0.913	0.834	0.834	0.851	0.752	0.741	0.741	0.736
	G_2	0.856	0.892	0.892	0.849	0.750	0.786	0.786	0.740
Simulation 2	G_1	0.820	0.841	0.799	0.792	0.723	0.730	0.709	0.704
	G_2	1.740	1.482	1.429	1.388	1.001	0.992	0.958	0.946

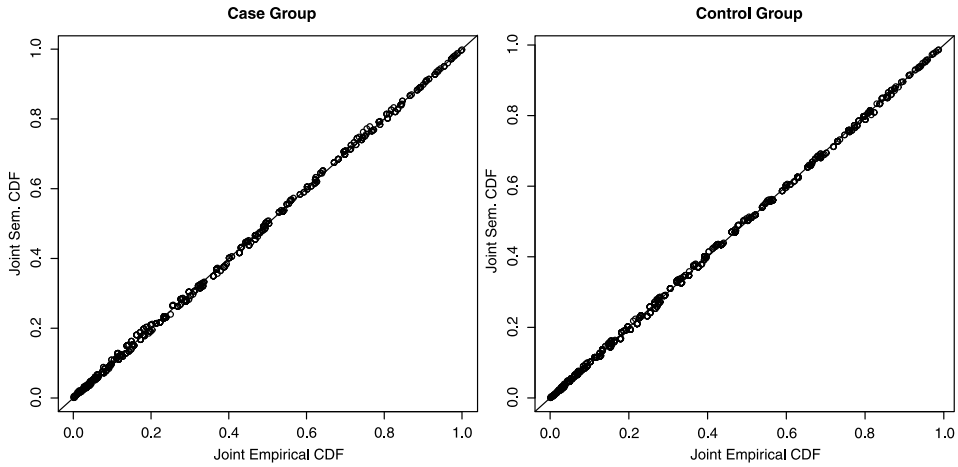

FIG. 3. *2D problem: diagnostic plots of* $\hat{G}_i$ *versus* $\tilde{G}_i$, $i = 1, 2$, *evaluated at* (*height*, *weight*) *pairs for the case and control groups from the TGCT data.*

(22) for the case and control groups, where in the 2D TGCT data set Y is weight and X is height. Superimposed are the regression lines obtained from linear regression under the normal assumption, GAM and the Nadaraya–Watson regression. For the 2D TGCT data, assuming normal distribution and identity link, we fitted a tensor product GAM; there were essentially no differences between the different kinds of splines. We notice that all models give similar results. The residual plots for the semiparametric model in Figure 5 are centered around zero.

Next we fitted the 3D TGCT data with variables age, height and weight. The semiparametric model is an appropriate model for this data set as Figure 6 shows.

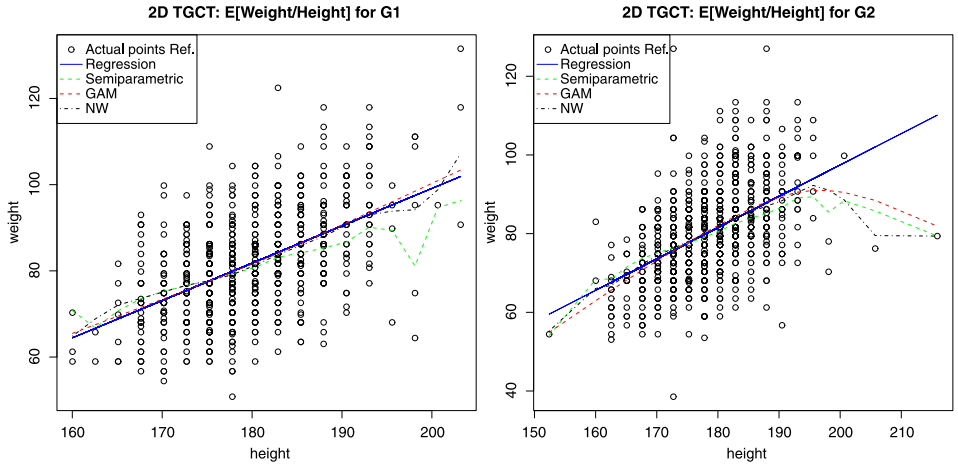

FIG. 4. *Comparison of* $\hat{E}$[*weight*|*height*] *for the 2D TGCT data set.*

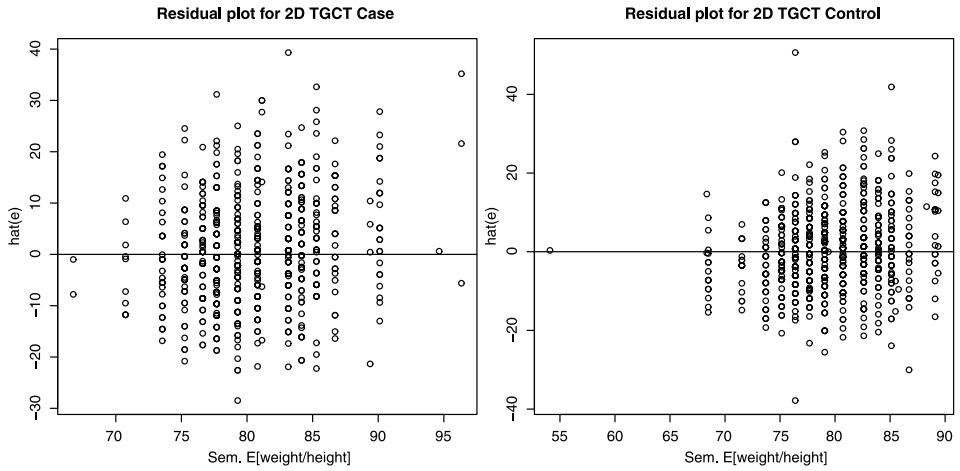

FIG. 5. *Residual plots for the semiparametric model in the* 2*D TGCT data set.*

The value of $R^2_{0.20,1}$ is 1 for both case and control. *An advantage of the method is that it gives estimates for the joint probabilities of age, height and weight in both case and control groups as in Table* 6. The table shows the two groups are moderately different.

In order to calculate $\hat{E}[Y|\mathbf{X}]$ for the case and control groups, we used (22), where in the 3D TGCT data set Y is weight and $\mathbf{X}$ represents jointly height and age. Figure 7 shows the residual plots for the semiparametric model. Table 7 gives the MSE and MAE comparison between the different regression methods for the 2D and the 3D TGCT data. For the 3D TGCT data, assuming normal distribution

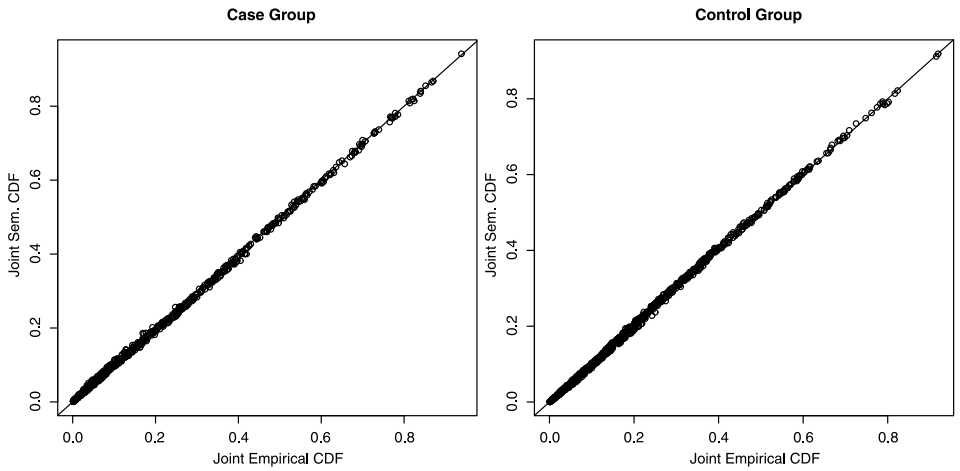

FIG. 6. *Case-control diagnostic plots of* $\hat{G}_i$ *versus* $\tilde{G}_i, i = 1, 2$, *for the* 3*D TGCT problem*: *the* $\hat{G}_i, \tilde{G}_i$ *are evaluated at selected* (*age*, *height*, *weight*) *triplets*.

TABLE 6
Some joint probabilities of age, height and weight in the case and control groups

Probability	**Case**	**Control**
$\Pr(A \leq 45, H \leq 152.40, W \leq 58.967)$	0.000378	0.000767
$\Pr(A \leq 26, H \leq 165.10, W \leq 58.967)$	0.004502	0.007074
$\Pr(A \leq 29, H \leq 177.80, W \leq 65.317)$	0.042723	0.054313
$\Pr(A \leq 33, H \leq 185.42, W \leq 70.307)$	0.157968	0.184774
$\Pr(A \leq 34, H \leq 180.34, W \leq 79.832)$	0.316077	0.362967
$\Pr(A \leq 37, H \leq 180.34, W \leq 89.811)$	0.513664	0.575512
$\Pr(A \leq 40, H \leq 187.96, W \leq 94.801)$	0.797157	0.833803
$\Pr(A \leq 43, H \leq 200.66, W \leq 99.790)$	0.943058	0.956300
$\Pr(A \leq 45, H \leq 203.20, W \leq 117.934)$	0.995010	0.996560

and identity link, we fitted a thin plate regression spline GAM because it produced better looking residual and Q–Q plots. The semiparametric regression performs comparably with the other estimators, although it has a somewhat higher MSE. These results can be explained by the fact that our method consists of an extra step of density estimation. However, we have the extra advantage that we also obtain joint probabilities of the variables, unlike multiple regression and GAM.

Tables 8 and 9 give some predicted values for weight given age and height for the two groups. The results from the different methods are not much different.

We end this section by providing $\hat{E}(y|\mathbf{x})$ in (22) to help the reader interpret the results of the semiparametric analysis. Tables 10 and 11 give the case-control weight predictions (22) and the actual weights. From the tables, as expected, $\hat{E}(y|\mathbf{x})$ in (22) tends to be close to the average of y's which correspond to the

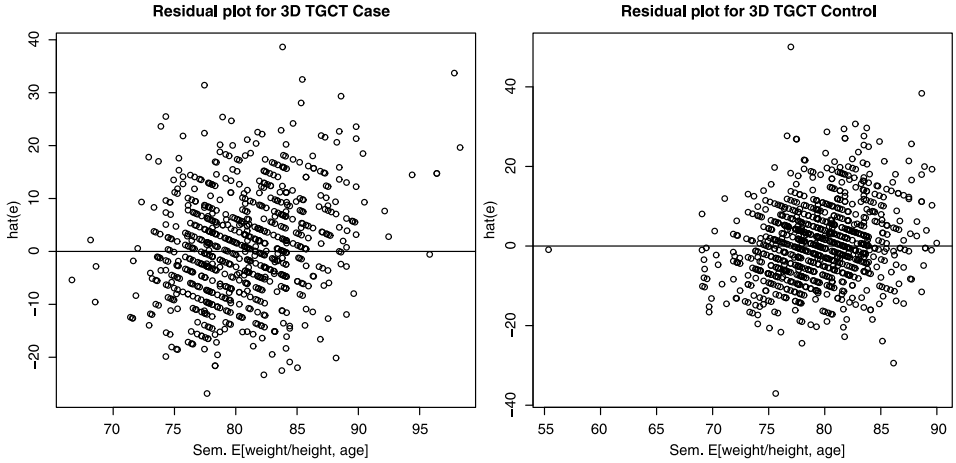

FIG. 7. *Residual plots for the semiparametric model in the 3D TGCT data set.*

TABLE 7
MAE and MSE comparison of the semiparametric, multiple, GAM and Nadaraya–Watson regression methods for 2D and 3D TGCT data

		MSE				MAE			
		SP	**MR**	**GAM**	**NW**	**SP**	**MR**	**GAM**	**NW**
2D TGCT	G_1	104.003	99.510	99.250	98.648	7.947	7.784	7.770	7.774
	G_2	93.010	92.264	90.284	90.332	7.347	7.296	7.246	7.241
3D TGCT	G_1	98.283	96.367	96.091	89.124	7.770	7.679	7.672	7.390
	G_2	91.643	90.291	88.147	86.932	7.280	7.244	7.173	7.139

same **x**. 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 point." Thus, for example, the control-weights corresponding to age 22 and height 175.26 average to 74.3894 and the prediction is 76.62195. Across different ages, except for heights less than 167.64 cm, the estimated conditional expectation in cases consistently has greater body weights than controls, indicating that later life exposures such as increased caloric diet intake and/or reduced energy expenditure and lack of physical exercise may increase the risk of testicular cancer.

7. Summary. In this paper we have shown that using our proposed semiparametric regression method we can detect an important increased risk of germ cell testicular cancer with greater body weight after adjusting for age and height that was not found with these same data using standard logistic regression modeling. This is an important finding because body weight is likely a later life exposure involving dietary caloric intake and/or energy expenditure from physical activity. This is in contrast to height that is influenced by early life factors such as genet-

TABLE 8
Predicted control values of weight given height and age

			Case			
Age	**Height**	**Weight**	**SP**	**MR**	**GAM**	**NW**
26	193.04	102.058	89.81775	92.47554	92.80697	95.96000
24	167.64	72.575	73.59282	70.00329	70.68805	71.90371
29	180.34	65.771	81.41551	82.42360	82.17237	81.60395
38	185.42	81.647	86.29762	89.46406	89.50287	89.70666
34	195.58	89.811	89.03635	97.03194	98.08814	92.45555
27	162.56	58.967	68.53652	66.51540	67.76775	65.18988

TABLE 9
Predicted case values of weight given height and age

Control						
Age	**Height**	**Weight**	**SP**	**MR**	**GAM**	**NW**
29	180.34	90.718	81.11841	82.06293	83.06542	82.35544
39	175.26	77.111	79.40282	80.36549	79.78087	80.05940
19	172.72	63.503	74.76493	73.58821	72.76199	73.40060
33	177.80	83.915	80.51759	80.97707	81.4916	81.14195
31	190.50	102.058	86.0598	90.67494	90.69862	87.47080
25	165.10	58.967	72.08147	68.90777	68.0279	69.49050

ics, early life nutrition or endogenous or exogenous hormones. The possibility of intervening to reduce body weight among young men could help to stem the rise in incidence of testicular cancer.

The semiparametric regression approach taken in this paper requires first efficient estimation of multivariate distributions. This can be achieved under the multidimensional density ratio model, given multiple data sources of multivariate data, and known tilts up to unknown parameters. Subject to this construct, the method produces more efficient kernel density estimators than the traditional single-sample kernel density estimator. This is so since all the finite and infinite-dimensional parameters are estimated from the entire combined data from all sources, and not just from single sources. As in the traditional kernel estimation, our kernel estimates require bandwidths and we have discussed ways for obtaining optimal and nearly optimal kernel bandwidths. The process of fitting the density ratio model and obtaining estimates is quite straightforward and quick. In this regard, several diagnostic measures have been suggested.

Going a step further, the estimated distributions can be used in estimating joint probabilities, in ANOVA-like problems of determining differences between groups, and in estimating the conditional expectation of a response variable given random covariates, provided that multiple data sources are available. An application to predicting weight from height and age in a case-control problem shows the method competes well with several well-known regression methods, and at the same time it provides estimates of joint probabilities. Our experience suggests that the method is effective for a small number of covariates. Computational problems can arise as the number of variables increases.

We have made some numerical comparisons with GAM, but a general comparison is not our focus or intention in the present paper. Still, a few points are in order. From our limited comparison it seems the two methods produce similar regression estimates, and both methods are more complex than multiple regression. The complexity of GAM stems from their iterative nature, which is reminiscent of

TABLE 10

Case-control weight and $\hat{E}$[weight|height, age]. Empty entries in the table correspond to subjects with the same height and age, but possibly different weights

		Control		Case	
Age	**Height**	**Weight**	$\hat{\boldsymbol{E}}[\boldsymbol{W}\|\boldsymbol{H}, \boldsymbol{A}]$	**Weight**	$\hat{\boldsymbol{E}}[\boldsymbol{W}\|\boldsymbol{H}, \boldsymbol{A}]$
27	162.56	58.967	69.08335	58.967	68.53652
28	162.56	77.111	69.05132	65.771	68.59858
		68.039			
30	165.10	68.039	72.20524	72.575	72.0028
37	165.10	69.40	72.42138	63.503	71.8504
25	167.64	86.183	73.68129	72.575	73.69978
				90.718	
				63.503	
30	167.64	72.575	74.81333	88.451	74.93543
18	170.18	61.235	73.67032	72.575	73.67518
32	170.18	70.307	76.53351	81.647	76.64543
		63.503			
37	172.72	74.843	77.88598	88.451	77.9417
40	172.72	70.307	77.97789	90.718	78.0441
		77.111			
22	175.26	77.111	76.62195	86.183	76.70862
		65.771		65.771	
		79.379		86.183	
		83.915			
		65.771			
25	175.26	68.039	77.14234	79.379	77.21755
		83.915		72.575	
		74.843		83.915	
		83.915		74.843	
		79.379		72.575	
		86.183		74.843	
				61.235	
				61.235	
				65.771	
				79.379	
26	177.80	79.379	78.74752	77.111	78.92705
		81.647		104.326	
		58.967		77.111	
		81.647			
		79.379			
		74.843			
		88.451			
		68.039			
42	177.80	70.307	80.50100	91.626	80.67493

TABLE 11
Case-control weight and $\hat{E}$[weight|height, age] continued. Empty entries in the table correspond to subjects with the same height and age, but possibly different weights

		Control		Case	
Age	**Height**	**Weight**	$\hat{E}[W\|H, A]$	**Weight**	$\hat{E}[W\|H, A]$
20	180.34	79.832	79.17623	84.368	79.35688
		65.771		68.039	
		77.111		79.379	
		79.379		81.647	
				72.575	
33	180.34	79.379	81.92536	77.111	82.17689
				81.647	
18	182.88	77.111	80.23013	68.039	80.29011
41	182.88	79.379	83.65558	86.183	84.06475
19	185.42	63.503	81.45580	68.039	82.09186
				94.347	
				68.039	
21	185.42	86.183	82.46773	79.379	82.78140
		72.575		77.111	
		102.058		97.522	
22	190.50	97.522	85.23493	86.183	85.64845
		95.254		71.668	
31	190.50	102.058	86.05980	104.326	86.27744
				74.843	
22	193.04	86.183	86.73352	102.058	87.18440
				80.739	
24	193.04	99.337	87.50020	108.862	88.23938
		86.183			
		99.790			
		108.862			
34	193.04	113.398	87.72937	88.451	88.58960
				117.934	
34	195.58	83.915	88.81524	89.811	89.036535

fixed point problems in repeated parametric filtering where estimates are evaluated at estimates iteratively, and this may affect the interpretability of the results [Li and Song (2002)]. It seems to us that the semiparametric approach, on the other hand, is somewhat more straightforward. We have illustrated in the TGCT data analysis that the resulting semiparametric regression estimate is indeed close to the average of the response conditional on fixed covariates, as one would expect. This property is shared by GAM as well. GAM assume additivity. On the other hand, the density ratio approach requires an assumption about the tilts. The suggested diagnostic tools shed light, albeit indirectly, on the appropriateness of the tilts.

APPENDIX

This Appendix contains supplemental material described in Voulgaraki, Kedem and Graubard (2012). It provides formal statements and indication of proofs of the results described in Sections 3.1 and 3.2.

Acknowledgments. The authors wish to thank the referees and the Area Editor for their dedication, effort and important suggestions.

SUPPLEMENTARY MATERIAL

Supplement to "Semiparametric regression in testicular germ cell data" (DOI: 10.1214/12-AOAS552SUPP; .pdf). The supplementary material contains all the mathematical proofs of the lemmas, corrolaries and theorems supporting the statements and results, including some additional references.

REFERENCES

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

BONDELL, H. D. (2007). Testing goodness-of-fit in logistic case-control studies. *Biometrika* **94** 487–495. MR2380573

CHENG, K. F. and CHU, C. K. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli* **10** 583–604. MR2076064

FOKIANOS, K. (2004). Merging information for semiparametric density estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 941–958. MR2102474

FOKIANOS, K., KEDEM, B., QIN, J. and SHORT, D. A. (2001). A semiparametric approach to the one-way layout. *Technometrics* **43** 56–65. MR1819908

GILBERT, P. B. (2004). Goodness-of-fit tests for semiparametric biased sampling models. *J. Statist. Plann. Inference* **118** 51–81. MR2015221

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. MR1688069

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. Chapman and Hall, London. MR1082147

KEDEM, B., LU, G., WEI, R. and WILLIAMS, P. D. (2008). Forecasting mortality rates via density ratio modeling. *Canad. J. Statist.* **36** 193–206. MR2522160

KEDEM, B., KIM, E.-Y., VOULGARAKI, A. and GRAUBARD, B. I. (2009). Two-dimensional semiparametric density ratio modeling of testicular germ cell data. *Stat. Med.* **28** 2147–2159. MR2751511

LI, T.-H. and SONG, K.-S. (2002). Asymptotic analysis of a fast algorithm for efficient multiple frequency estimation. *IEEE Trans. Inform. Theory* **48** 2709–2720. MR1930338

LU, G. (2007). Asymptotic theory for multiple-sample semiparpametric density ratio models and its application to mortality forecasting. Ph.D. dissertation, Univ. Maryland, College Park, MD. MR2711353

MCGLYNN, K. A. and 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.) 51–83. Springer, New York.

MCGLYNN, K. A., DEVESA, S. S., SIGURDSON, A. J., BROWN, L. M., TSAO, L. and TARONE, R. E. (2003). Trends in the incidence of testicular germ cell tumors in the United States. *Cancer* **97** 63–70.

McGLYNN, K. A., SAKODA1, 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.

NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl*. **9** 141–142.

OGDEN, C. L., FRYAR, C. D., CARROLL, M. D. and FLEGAL, K. M. (2004). Mean body weight, height, and body mass index, United States 1960–2002. *Adv. Data* **347** 1–17.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist*. **33** 1065–1076. MR0143282

PHUE, J.-N., KEDEM, B., JALURIA, P. and SHILOACH, J. (2007). Evaluating microarrays using a semiparametric approach: Application to the central carbon metabolism of Escherichia coli BL21 and JM109. *Genomics* **89** 300–305.

PRENTICE, R. L. and PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–411. MR0556730

QIN, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* **85** 619–630. MR1665814

QIN, J. and LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist*. **22** 300–325. MR1272085

QIN, J. and ZHANG, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **84** 609–618. MR1603924

QIN, J. and ZHANG, B. (2005). Density estimation under a two-sample semiparametric model. *J. Nonparametr. Stat*. **17** 665–683. MR2165095

RENCHER, A. C. (2000). *Linear Models in Statistics*. Wiley, New York. MR1719141

SHAO, J. (2003). *Mathematical Statistics*, 2nd ed. Springer, New York. MR2002723

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London. MR0848134

VOULGARAKI, A., KEDEM, B. and GRAUBARD, B. I. (2012). Supplement to "Semiparametric regression in testicular germ cell data." DOI:10.1214/12-AOAS552SUPP.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372. MR0185765

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**. On-line journal without page numbers.

WOOD, S. N. (2006). *Generalized Additive Models*: *An Introduction With R*. Chapman & Hall/CRC, Boca Raton, FL. MR2206355

ZHANG, B. (1999). A chi-squared goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **86** 531–539. MR1723776

ZHANG, B. (2000). A goodness of fit test for multiplicative-intercept risk models based on case-control data. *Statist. Sinica* **10** 839–865. MR1787782

ZHANG, B. (2001). An information matrix test for logistic regression models based on case-control data. *Biometrika* **88** 921–932. MR1872210

ZHANG, B. (2002). Assessing goodness-of-fit of generalized logit models based on case-control data. *J. Multivariate Anal*. **82** 17–38. MR1918613

A. VOULGARAKI
B. KEDEM
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND 20742
USA
E-MAIL: avoulgar@math.umd.edu
bnk@math.umd.edu

B. I. GRAUBARD
NATIONAL CANCER INSTITUTE
ROCKVILLE, MARYLAND 20852
USA
E-MAIL: graubarb@exchange.nih.gov