Semiparametric Regression Based on Multiple Sources

Benjamin Kedem

Department of Mathematics & Inst. for Systems Research
University of Maryland, College Park

“Give me a place to stand and rest my lever on, and I can move the Earth”,
(Archimedes, 287-212 B.C.)

STAT Seminar, November 1, 2012


   Supplement: DOI:10.1214/12-AOAS552SUPP
Abstract

1. Information is combined from multiple multivariate sources.
2. Density ratio model: Multivariate distributions are “regressed” on a reference distribution.
3. Random covariates.
4. Kernel density estimator from many data sources: more efficient than the traditional single-sample kernel density estimator.
5. Each multivariate distribution and the corresponding conditional expectation (regression) of interest are estimated from the combined data using all sources.
6. Graphical and quantitative diagnostic tools are suggested to assess model validity.
7. The method is applied in quantifying the effect of height and age on weight of germ cell testicular cancer patients.
8. Comparison with multiple regression, GAM, and nonparametric kernel regression.
Main Points

- Density Ratio Examples
  - Semiparametric Statistical Formulation
  - Combined semiparametric density estimators
  - Some Simulation Results
  - Application to Testicular Germ Cell Cancer
  - Summary
Main Points

- Density Ratio Examples
- Semiparametric Statistical Formulation
- Combined semiparametric density estimators
- Some Simulation Results
- Application to Testicular Germ Cell Cancer
- Summary
Main Points

- Density Ratio Examples
- Semiparametric Statistical Formulation
- Combined semiparametric density estimators
- Some Simulation Results
- Application to Testicular Germ Cell Cancer
- Summary
Main Points

- Density Ratio Examples
- Semiparametric Statistical Formulation
- Combined semiparametric density estimators
- Some Simulation Results
- Application to Testicular Germ Cell Cancer
- Summary
Main Points

▶ Density Ratio Examples
▶ Semiparametric Statistical Formulation
▶ Combined semiparametric density estimators
▶ Some Simulation Results
▶ Application to Testicular Germ Cell Cancer
▶ Summary
Main Points

- Density Ratio Examples
- Semiparametric Statistical Formulation
- Combined semiparametric density estimators
- Some Simulation Results
- Application to Testicular Germ Cell Cancer
- Summary
TRMM sensors: Likely distortions of ground truth
Reference: Ground truth
Multiple filtering of a signal

\[ f_1(\omega) = |H_1(\omega)|^2 f(\omega) \]

\[ \cdot \]

\[ \cdot \]

\[ \cdot \]

\[ f_q(\omega) = |H_q(\omega)|^2 f(\omega) \]

That is, \( q \) “distortions” or multiple “tilting” of the same reference spectral density \( f \).
Linear system with Gaussian noise

\[ X_t = \alpha X_{t-1} + \epsilon_t, \quad \epsilon = (\epsilon_1, \ldots, \epsilon_q, \epsilon_{mt})' \]

\[ \epsilon_j \sim g_j \sim N(0, \sigma_j^2), \quad j = 1, \ldots, q, m. \] Choose \( g_m \equiv g \). We get many distortions of the same reference \( g \):

\[ g_1(x) = e^{\alpha_1 + \beta_1 x^2} g(x) \]

\[ \ldots \]

\[ g_q(x) = e^{\alpha_q + \beta_q x^2} g(x) \]
Analysis of variance

Consider the classical one-way ANOVA with \( m = q + 1 \) independent normal random samples:

\[
\begin{align*}
x_{11}, \ldots, x_{1n_1} & \sim g_1(x) \\
\vdots & \\
\vdots & \\
x_{q1}, \ldots, x_{qn_q} & \sim g_q(x) \\
x_{m1}, \ldots, x_{mn_m} & \sim g_m(x)
\end{align*}
\]

\[ g_j(x) \sim N(\mu_j, \sigma^2), \quad j = 1, \ldots, m. \]
Then, holding $g_m(x) \equiv g(x)$ as a reference:

$$g_1(x) = \exp(\alpha_1 + \beta_1 x)g(x)$$

.$$.$$.$$.$$.$$

$$g_q(x) = \exp(\alpha_q + \beta_q x)g(x)$$

$$\alpha_j = \frac{\mu^2_m - \mu^2_j}{2 \sigma^2}, \quad \beta_j = \frac{\mu_j - \mu_m}{\sigma^2}, \quad j = 1, \ldots, q$$
$k$-parameter exponential families

\[ g(x, \theta) = d(\theta)S(x) \exp \left\{ \sum_{i=1}^{k} c_i(\theta) T_i(x) \right\} \]

Let \( g_j(x) \equiv g(x, \theta_j) \), \( g(x) \equiv g(x, \theta_m) \).

Again, \( q \) distortions of a reference \( g(x) \):

\[ g_1(x) = \exp\{\alpha_1 + \beta'_1 h(x)\} g(x) \]

\[ . \]

\[ . \]

\[ . \]

\[ g_q(x) = \exp\{\alpha_q + \beta'_q h(x)\} g(x) \]
Case-control: Multinomial logistic regression (Prentice and Pyke 1979).

- RV $y$ s.t. $P(y = j) = \pi_j$, $\sum_{j=1}^{m} \pi_j = 1$.
- Assume: For $j = 1, \ldots, m$, and any $h(x)$,

$$P(y = j | x) = \frac{\exp(\alpha_j^* + \beta_j h(x))}{1 + \sum_{k=1}^{q} \exp(\alpha_k^* + \beta_k h(x))}$$

- Define: $f(x|y = j) = g_j(x)$, $j = 1, \ldots, m$

Then with $\alpha_j = \alpha_j^* + \log[\pi_m/\pi_j]$, $j = 1, \ldots, q$, and $g_m \equiv g$,

$$g_j(x) = \exp(\alpha_j + \beta_j h(x))g(x), \ j = 1, \ldots, q.$$
Rao (1965) unified the concept of weighted distributions of the form:

$$p_w(x; \theta, \alpha) = \frac{W(x; \alpha)p(x; \theta)}{E[W(X; \alpha)]]}$$

where $W(x; \alpha)$ is known. This is an example of “tilting” of a reference distribution. By changing the weight $W(x; \alpha)$ we obtain different distortions of the same reference $p(x; \theta)$. 
Biased sampling

- **Length-biased Sampling**: Vardi (1982) introduced
  
  $$F(y) = \frac{1}{\mu} \int_0^y x dG(x), \ y \geq 0,$$

  where \(\mu = \int_0^\infty x dG(x) < \infty\). This is a tilt model.

- **Biased Sampling/Selection Bias**: Vardi (1985), and Gill, Vardi, Wellner (1988) considered the more general biased sampling model
  
  $$F(y) = W(G)^{-1} \int_{-\infty}^y w(x) dG(x),$$

  where \(w(x)\) is known and \(W(G) = \int_{-\infty}^\infty w(x) dG(x)\). This is a tilt model. \(F, G\) are obtained by NPMLE.
Biased sampling

- **Length-biased Sampling:** Vardi (1982) introduced
  \[ F(y) = \frac{1}{\mu} \int_0^y x dG(x), \quad y \geq 0, \]
  where \( \mu = \int_0^\infty x dG(x) < \infty \). This is a tilt model.

- **Biased Sampling/Selection Bias:** Vardi (1985), and Gill, Vardi, Wellner (1988) considered the more general biased sampling model
  \[ F(y) = W(G)^{-1} \int_{-\infty}^y w(x) dG(x), \]
  where \( w(x) \) is known and \( W(G) = \int_{-\infty}^\infty w(x) dG(x) \). This is a tilt model. \( F, G \) are obtained by NPMLE.
Comparison Distributions (Parzen 1977,...,2009)

Sequence of cdf’s: \( \{ F_1, ..., F_q \} \ll G \), with cont. densities \( f_1, ..., f_q, g \).

*Comparison Distributions* are defined as:

\[
D_j(u; G, F_j) = F_j(G^{-1}(u)), \quad 0 < u < 1, \quad j = 1, ..., q
\]

Then by differentiation, with \( x = G^{-1}(u) \):

\[
f_1(x) = d(G(x); G, F_1)g(x)
\]

\[
\cdot
\]

\[
\cdot
\]

\[
f_q(x) = d(G(x); G, F_q)g(x)
\]
A general structure emerges of a reference behavior (distribution) and its many distortions:

\[ g_1 = w_1 g \]

\[ \ldots \]

\[ g_q = w_q g \]

How can we take advantage of this?
Can the distorted information be of use?
Can the distorted and reference information be fused as to improve the reference information?
In other words, can the “bad” improve the “good”? 
A general structure emerges of a reference behavior (distribution) and its many distortions:

\[ g_1 = w_1 g \]
\[ \ldots \]
\[ \ldots \]
\[ g_q = w_q g \]

How can we take advantage of this?

- Can the distorted information be of use?
- Can the distorted and reference information be fused as to improve the reference information?
- In other words, can the “bad” improve the “good”?
A general structure emerges of a reference behavior (distribution) and its many distortions:

\[ g_1 = w_1 g \]
\[ . \]
\[ . \]
\[ . \]
\[ g_q = w_q g \]

How can we take advantage of this?

Can the distorted information be of use?

Can the distorted and reference information be fused as to improve the reference information?

In other words, can the “bad” improve the “good”? 
A general structure emerges of a reference behavior (distribution) and its many distortions:

\[ g_1 = w_1 g \]

\[ \quad \]

\[ \quad \]

\[ g_q = w_q g \]

How can we take advantage of this?

Can the distorted information be of use?

Can the distorted and reference information be fused as to improve the reference information?

In other words, can the “bad” improve the “good”?
A general structure emerges of a reference behavior (distribution) and its many distortions:

\[ g_1 = w_1 g \]
\[ \ldots \]
\[ \ldots \]

\[ g_q = w_q g \]

How can we take advantage of this?

Can the distorted information be of use?

Can the distorted and reference information be fused as to improve the reference information?

In other words, can the “bad” improve the “good”?
The relationship between a reference and its distortions or tilts opens the door to a useful general statistical approach based on *fused* or *combined* information from many sources. A case in point is the estimation of the conditional expectation from fused multivariate samples.
Outline
$m = q + 1$ independent $p$-dim multivariate samples:

$x_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp}) \sim g_i(x_1, \ldots, x_p), \ i = 1, \ldots, q, \ m, \ j = 1, \ldots, n_i$

and

$x_{i1}, x_{i2}, \ldots, x_{in_i} \overset{iid}{\sim} g_i$

Reference or baseline probability density function (pdf):

$$g \equiv g_m(x) \equiv g_m(x_1, \ldots, x_p)$$
Density ratio model:

\[
\frac{g_i(x)}{g(x)} = w(x, \theta_i)
\]  \hspace{1cm} (2)

or equivalently

\[
g_i(x) = w(x, \theta_i)g(x)
\]  \hspace{1cm} (3)

1. \(g_i(x), g(x)\) are unknown.
2. \(w\) is a known positive and continuous function.
3. \(\theta_i\) are unknown \(d\)-dimensional parameters.
4. pdf’s may be continuous or discrete.
5. Normal assumption is not made.
Problem:
Use the combined data from all the samples
\[ \mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1}, \ldots, \mathbf{x}_{m1}, \ldots, \mathbf{x}_{mn_m} \]

of size \( n = n_1 + \cdots + n_m \) to estimate all the parameters \( \theta_i \), \( i = 1, \ldots, q \), and all the densities \( g_1, \ldots, g_q \) and \( g_m = g \).
\[ G(\mathbf{x}) \equiv G_m(\mathbf{x}), \quad p_{ij} = dG(\mathbf{x}_{ij}) = dG_m(\mathbf{x}_{ij}). \]

The empirical log-likelihood is given by:

\[
I = \log L = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log(p_{ij}) + \sum_{i=1}^{q} \sum_{j=1}^{n_i} \log(w(\mathbf{x}_{ij}, \theta_i)) \tag{4}
\]

subject to the constraints:

\[
p_{ij} \geq 0, \quad \sum_{i=1}^{m} \sum_{j=1}^{n_i} p_{ij} = 1, \quad \sum_{i=1}^{m} \sum_{j=1}^{n_i} p_{ij}w(\mathbf{x}_{ij}, \theta_k) = 1 \text{ for } k = 1, \ldots, q. \tag{5}
\]
\[\hat{p}_{ij} = \frac{1}{n_1 + \sum_{k=1}^{q} \hat{\mu}_k \left[ w(x_{ij}, \hat{\theta}_k) - 1 \right]}, \]  

(6)

\[\hat{G}(x) = \hat{G}_m(x) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{p}_{ij} l(x_{ij} \leq x)\]  

(7)

More generally, for \( l = 1, \ldots, m \) and \( w(x_{ij}, \hat{\theta}_m) \equiv 1 \),

\[\hat{G}_l(x) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{p}_{ij} w(x_{ij}, \hat{\theta}_l) l(x_{ij} \leq x)\]  

(8)
Outline
Traditional single sample kernel density estimator:

\[ \hat{f}(x) = \frac{1}{nh_n^p} \sum_{i=1}^{n} K \left( \frac{x - x_i}{h_n} \right) \]  

(9)

\( h_n \) is a sequence of bandwidths such that:

1. \( h_n \to 0, \ nh_n^p \to \infty \) as \( n \to \infty \).
2. \( K(x) \) is defined for \( p \)-dimensional \( x \).
3. \( K(x) \geq 0 \), symmetric around \( 0 \).
4. \( \int_{\mathbb{R}^p} K(x) dx = 1 \).

Under certain conditions, \( \hat{f}(x) \) is a consistent estimator of \( f(x) \) (Parzen 1962, Shao 2003).
Fokianos (2004), Cheng and Chu (2004), Qin and Zhang (2005), Voulgaraki, Kedem, Graubard (VKG) (2012), use the probabilities $\hat{p}_{ij}$ instead of $1/n$:

$$
\hat{g}_l(x) = \frac{1}{h^n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{p}_{ij} \hat{w}_l(x_{ij}) K \left( \frac{x - x_{ij}}{h_n} \right)
$$  \hspace{1cm} (10)
VKG 2012: Assume that $K(\cdot)$ is a nonnegative bounded symmetric function with $\int K(x)dx = 1$, $\int x'xK(x)dx = k_2 > 0$. Assume that $g_l$ is continuous at $x$ and twice differentiable in a neighborhood of $x$. If, as $n \to \infty$, $h_n = O(n^{-\frac{1}{4+p}})$, then

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

where

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

for any fixed $x$. 
The mean integrated square error (MISE) is defined as:

$$\text{MISE}(\hat{g}_l(x)) = E \left( \int |\hat{g}_l(x) - g_l(x)|^2 \, dx \right)$$  \hspace{1cm} (11)

Minimizing the MISE with respect to $h_n$, the optimal bandwidth is:

$$h_n^* = \left( \frac{p/n}{\int w_l(x)g_l(x)/\left[ \sum_{k=1}^{m} \zeta_k w_k(x) \right] \, dx \int K^2(u) \, du} \right)^{\frac{1}{4+p}}$$

$$\left[ \int \left( \int u'(\partial^2 g_l(x)/\partial x \partial x')uK(u) \, du \right)^2 \, dx \right]$$  \hspace{1cm} (12)
For sufficient large $n$, an alternative way to find $h_n$ is by minimizing

$$h_n^{-p} \sum_{i=1}^{n} \sum_{i'=1}^{n} \hat{p}(t_i) \hat{w}_l(t_i) \hat{p}(t_{i'}) \hat{w}_l(t_{i'}) \int K(z)K \left( z + \frac{t_i - t_{i'}}{h_n} \right) dz$$

$$- \frac{2}{n_l(n_l - 1)h_n^p} \sum_{i \neq j} K \left( \frac{x_{li} - x_{lj}}{h_n} \right). \quad (13)$$
VKG 2012: If $\hat{f}$ is the classical single-sample multivariate kernel density estimator of $g_l$, then As $n \to \infty$, $h_n \to 0$ and $nh_n^p \to \infty$,

(a)  

$$AMISE(\hat{g}_l) \leq AMISE(\hat{f})$$

(b) Using optimal bandwidths, the proposed semiparametric density estimator $\hat{g}_l(x)$ is more efficient than $\hat{f}(x)$, i.e for every $l$

$$\text{eff}(\hat{f}, \hat{g}_l) \equiv \frac{AMISE^*(\hat{g}_l)}{AMISE^*(\hat{f})} \leq 1$$

where $AMISE^*$ is the optimal $AMISE$. 
Define

\[ R^{2}_{\alpha,k} = 1 - \exp \left\{ - \left( \frac{x_{\alpha}}{n_{i} - x_{\alpha}} \right)^{k} \right\} \]  

\[ (14) \]

- \( n_{i} \): Sample size of \( i \)th sample.
- \( x_{\alpha} \): 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.
- \( k > 0 \) and \( \alpha \) chosen by the user.
- \( R^{2}_{\alpha,k} \) takes values between 0 and 1.
- Computing \( R^{2}_{\alpha,k} \) is both simple and fast.
Qin and Zhang (1997):

\[ R_3^2 = \exp(-\sqrt{n} \cdot \max |\tilde{G}_i - \hat{G}_i|) \]  \hspace{1cm} (15)

Clearly, \( R_3^2 \) takes values between 0 and 1.
Simulation Results: $\hat{G}_i$ vs. $\tilde{G}_i$, $i = 1, 2$

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

Simulation 1: G1

Simulation 1: G2
Simulation 2: \( g_1 \sim N((0, 0)', \Sigma) \) (case), \( g_2 \sim N((1, 1)', \Sigma) \) (control) with
\[
\Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix},
\]
n_1 = 200, n_2 = 200.
Simulation 3: \( g_1 \) from standard two dimensional Multivariate Cauchy (case) and \( g_2 \) from two dimensional Multivariate Cauchy (control) with \( \mu = (1, 1)' \), \( V = \begin{pmatrix} 5 & 5 \\ 5 & 10 \end{pmatrix} \), \( n_1 = 200, n_2 = 200 \).
Simulation 4: \( g_1 \) from standard two dimensional Multivariate Cauchy (case) and \( g_2 \) from uniform distribution on the triangle \((0, 0), (6, 0), (-3, 4)\) (control), and \( n_1 = 200, n_2 = 200 \).
**Table:** Comparison of $R^2_3$ and $R^2_{0.05,2}$ for 100 repetitions of case and control.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Group</th>
<th>$R^2_3$</th>
<th>$R^2_{0.05,2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Case</td>
<td>0.6307</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>0.5976</td>
<td>1</td>
</tr>
<tr>
<td>(2)</td>
<td>Case</td>
<td>0.3912</td>
<td>0.9353</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>0.3766</td>
<td>0.9718</td>
</tr>
<tr>
<td>(3)</td>
<td>Case</td>
<td>0.1080</td>
<td>0.3342</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>0.1129</td>
<td>0.3324</td>
</tr>
<tr>
<td>(4)</td>
<td>Case</td>
<td>0.0507</td>
<td>0.3361</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>0.0495</td>
<td>0.0033</td>
</tr>
</tbody>
</table>
Bandwidth (BW) selection using cross validation.

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Same BW</td>
<td>Diff. BW's</td>
</tr>
<tr>
<td></td>
<td>$h$</td>
<td>$h_1$</td>
</tr>
<tr>
<td>Leave One Out</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulation 1</td>
<td>0.61</td>
<td>0.90</td>
</tr>
<tr>
<td>Simulation 2</td>
<td>0.38</td>
<td>0.50</td>
</tr>
<tr>
<td>Alternative (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulation 1</td>
<td>0.64</td>
<td>0.90</td>
</tr>
<tr>
<td>Simulation 2</td>
<td>0.30</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Outline
Suppose we have $m$ multivariate samples of $p$-dim vectors:

$$(\text{Random Covariates, Response}) = (x_1, \ldots, x_{(p-1)}, y)$$

We wish to estimate $m$ conditional expectations:

$$E(y | x_1, \ldots, x_{(p-1)})$$
Data:
\[ i = 1, \ldots, q, m, \ j = 1, \ldots, n_i, \]
\[ (x_{ij1}, x_{ij2}, \ldots, x_{ij(p-1)}, y_{ij}) \sim g_i(x_1, \ldots, x_{(p-1)}, y). \]

Reference:
\[ g \equiv g_m(x_1, \ldots, x_{(p-1)}, y) \]

Distortions:
\[ g_i(x_1, \ldots, x_{(p-1)}, y), \ i = 1, \ldots, q \]
Model:

As in ratio of pdf’s from $N(\mu_1, \Sigma), N(\mu_2, \Sigma)$ assume

$$\frac{g_i(x)}{g(x)} = w(x, \alpha_i, \beta_i) \equiv \exp(\alpha_i + \beta_i'x), \quad i = 1, \ldots, q \quad (16)$$

where $x = (x_1, \ldots, x_{(p-1)}, y)'$ and $\beta_i = (\beta_{i1}, \ldots, \beta_{ip})'$. 
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 \( \hat{\beta}_j \):

\[
\frac{\partial l}{\partial \alpha_j} = -\sum_{i=1}^{n} \frac{\rho_j w_j(t_i)}{1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)} + n_j = 0 \quad (17)
\]

\[
\frac{\partial l}{\partial \beta_j} = -\sum_{i=1}^{n} \frac{\rho_j w_j(t_i) t_i}{1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)} + \sum_{i=1}^{n_j} (x_{ji1}, \ldots, y_{ji})' (18)
\]

\( j = 1, \ldots, q \) and

\[
\rho_j = n_j/n_m
\]

\[
w_j(t_i) = \exp(\alpha_j + \beta_j' t_i)
\]
\[ \hat{\rho}_i = \frac{1}{n_m} \cdot \frac{1}{1 + \rho_1 \hat{w}_1(t_i) + \cdots + \rho_q \hat{w}_q(t_i)} \]  

(19)

\[ \hat{G}(t) = \frac{1}{n_m} \cdot \sum_{i=1}^{n} \frac{l(t_i \leq t)}{1 + \rho_1 \hat{w}_1(t_i) + \cdots + \rho_q \hat{w}_q(t_i)} \]  

(20)

where

\[ \hat{w}_j(t_i) = \exp(\hat{\alpha}_j + \hat{\beta}'_j t_i) \]

As \( n \to \infty \), the estimators \( \hat{\theta} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_q, \hat{\beta}_1, \cdots, \hat{\beta}_q)' \) are asymptotically normal (Qin and Zhang 1997, Lu 2007).
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:

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

(21)

The \( \hat{g}_j \) in (21) are the semiparametric kernel density estimates.
Assume that the data are bounded. Then:

(a) As $n \to \infty$, $h \to 0$ and $nh^p \to \infty$, 

$$\int |\hat{E}(y|x) - E(y|x)| g(x) dx \to 0$$

in the mean square sense.

(b) If, in addition $0 < A < g(x)$, then

$$\int |\hat{E}(y|x) - E(y|x)| dx \to 0$$

in the mean square sense.
Outline
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).

Increased risk is significantly related to height, whereas body mass index was not found significant (McGlynn et al 2007).

Using a two dimensional semiparametric model, it was shown that jointly height and weight are significant risk factors (Kedem et al 2009).

We use TGCT data with 3 variables: age, height (cm), weight (kg).

Case: \(n_1 = 763\). Control: \(n_2 = 928\). \(n = 1691\) individuals.

We consider two cases:
2D with variables height and weight
3D with variables height, weight and age.
In both cases the control distribution is the reference distribution.
2D Problem: $E(\text{Weight} \mid \text{Height})$

Diagnostic plots of $\hat{G}_i$ versus $\tilde{G}_i$, $i = 1, 2$ evaluated at (height, weight) pairs.
2D Case Regression

2D TGCT: E[Weight/Height] for G1

Residual plot for 2D TGCT Case
2D Control Regression

2D TGCT: $E[\text{Weight/Height}]$ for G2

Residual plot for 2D TGCT Control

Sem. $E[\text{weight/height}]$
3D Problem: $E(\text{Weight} \mid \text{Height, Age})$

Diagnostic plots of $\hat{G}_i$ versus $\tilde{G}_i$, $i = 1, 2$ evaluated at (age, height, weight) triplets.
3D Residuals

Residual plots for the semiparametric model in the 3D TGCT data set.
Some joint probabilities of age, height and weight in the case and control groups.

<table>
<thead>
<tr>
<th>Probability</th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Pr(A \leq 45, H \leq 152.40, W \leq 58.967)$</td>
<td>0.000378</td>
<td>0.000767</td>
</tr>
<tr>
<td>$\Pr(A \leq 26, H \leq 165.10, W \leq 58.967)$</td>
<td>0.004502</td>
<td>0.007074</td>
</tr>
<tr>
<td>$\Pr(A \leq 29, H \leq 177.80, W \leq 65.317)$</td>
<td>0.042723</td>
<td>0.054313</td>
</tr>
<tr>
<td>$\Pr(A \leq 33, H \leq 185.42, W \leq 70.307)$</td>
<td>0.157968</td>
<td>0.184774</td>
</tr>
<tr>
<td>$\Pr(A \leq 34, H \leq 180.34, W \leq 79.832)$</td>
<td>0.316077</td>
<td>0.362967</td>
</tr>
<tr>
<td>$\Pr(A \leq 37, H \leq 180.34, W \leq 89.811)$</td>
<td>0.513664</td>
<td>0.575512</td>
</tr>
<tr>
<td>$\Pr(A \leq 40, H \leq 187.96, W \leq 94.801)$</td>
<td>0.797157</td>
<td>0.833803</td>
</tr>
<tr>
<td>$\Pr(A \leq 43, H \leq 200.66, W \leq 99.790)$</td>
<td>0.943058</td>
<td>0.956300</td>
</tr>
<tr>
<td>$\Pr(A \leq 45, H \leq 203.20, W \leq 117.934)$</td>
<td>0.995010</td>
<td>0.996560</td>
</tr>
</tbody>
</table>
Case-control weight and $\hat{E}[\text{weight}| \text{height, age}]$. Empty entries in the table correspond to subjects with the same height and age, but possibly different weights.

<table>
<thead>
<tr>
<th>Age</th>
<th>Height</th>
<th>Weight</th>
<th>$\hat{E}[W \mid H, A]$</th>
<th>Weight</th>
<th>$\hat{E}[W \mid H, A]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>162.56</td>
<td>58.967</td>
<td>69.08335</td>
<td>58.967</td>
<td>68.53652</td>
</tr>
<tr>
<td>28</td>
<td>162.56</td>
<td>77.111</td>
<td>69.05132</td>
<td>65.771</td>
<td>68.59858</td>
</tr>
<tr>
<td></td>
<td></td>
<td>68.039</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>165.10</td>
<td>68.039</td>
<td>72.20524</td>
<td>72.575</td>
<td>72.0028</td>
</tr>
<tr>
<td>37</td>
<td>165.10</td>
<td>69.40</td>
<td>72.42138</td>
<td>63.503</td>
<td>71.8504</td>
</tr>
<tr>
<td>25</td>
<td>167.64</td>
<td>86.183</td>
<td>73.68129</td>
<td>72.575</td>
<td>73.69978</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>90.718</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>63.503</td>
<td></td>
</tr>
</tbody>
</table>
Case-control weight and $\hat{E}[\text{weight}|\text{height, age}]$. Empty entries in the table correspond to subjects with the same height and age, but possibly different weights.

| Age | Height | Weight | $\hat{E}[W|H, A]$ | Case | $\hat{E}[W|H, A]$ |
|-----|--------|--------|-------------------|------|-------------------|
| 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          |
| 37  | 172.72 | 74.843 | 77.88598          | 88.451| 77.9417           |
| 40  | 172.72 | 70.307 | 77.97789          | 90.718| 78.0441           |
| 22  | 175.26 | 77.111 | 76.62195          | 86.183| 76.70862          |
| 25  | 175.26 | 68.039 | 77.14234          | 79.379| 77.21755          |
Case-control weight and $\hat{E}[\text{weight}|\text{height, age}]$. Empty entries in the table correspond to subjects with the same height and age, but possibly different weights.

| Age | Height | Weight | $\hat{E}[W | H, A]$ | Control | Weight | $\hat{E}[W | H, A]$ | Case |
|-----|--------|--------|---------------------|---------|--------|---------------------|------|
| 21  | 185.42 | 86.183 | 82.46773            | 79.379  | 82.78140         |
|     |        | 72.575 |                     | 77.111  | 85.64845         |
|     |        | 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        |