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.)

AISC, October 5-7, 2012

1. Cheng, K.F., and Chu, C.K. (2004), “Semiparametric density estimation under a two-sample density ratio model,” *Bernoulli*, 10(4), 583-604.
2. Fokianos, K. (2004). Merging information for semiparametric density estimation. *Journal of the Royal Statistical Society, Series B*, 66, 941-958.
3. Fokianos, K., Kedem, B., Qin, J., Short, D.A. (2001). A semiparametric approach to the one-way layout. *Technometrics*, 43, 56-65.
4. McGlynn K.A., Sakoda1 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.
5. Qin, J., Zhang, B. (1997). A goodness of fit test for logistic regression models based on case-control data. *Biometrika*, 84, 609-618.
6. Qin, J., and Zhang, B. (2005). Density Estimation Under A Two-Sample Semiparametric Model. *Nonparametric Statistics*, 1, 665-683.
7. Kedem, B., Kim, E-Y, Voulgaraki, A., and Graubard, B. (2009). Two-Dimensional Semiparametric Density Ratio Modeling of Testicular Germ Cell Data. *Statistics in Medicine*, 2009, 28, 2147-2159.
8. Voulgaraki, A., Kedem, B., and Graubard, B.(2012). "Semiparametric Regression in Testicular Germ Cell Data", *Annals of Applied Statistics*, 6, 1185-1208
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 cocariates.
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

Outline

1. Density Ratio Examples
2. Semiparametric Statistical Formulation
3. Combined semiparametric density estimators
4. Semiparametric regression
5. Application to Testicular Germ Cell Cancer

Multiple filtering of a signal

$$\begin{aligned} f_1(\omega) &= |H_1(\omega)|^2 f(\omega) \\ &\vdots \\ f_q(\omega) &= |H_q(\omega)|^2 f(\omega) \end{aligned} \tag{1}$$

That is, q "distortions" or multiple "tilting" of the same *reference* spectral density f.

Linear system with Gaussian noise

$$\boldsymbol{X}_t = \boldsymbol{\alpha}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon} = (\epsilon_{1t}, ..., \epsilon_{qt}, \epsilon_{mt})'$$

$\epsilon_{jt} \sim g_j \sim N(0, \sigma_j^2)$, $j = 1, ..., q, m$. Choose $g_m \equiv g$. We get many distortions of the same *reference* g:

$$\begin{aligned} g_1(x) &= e^{\alpha_1 + \beta_1 x^2} g(x) \\ &\cdot \\ &\cdot \\ &\cdot \\ g_q(x) &= e^{\alpha_q + \beta_q x^2} g(x) \end{aligned}$$

Analysis of variance

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

$$
\begin{aligned}
x_{11}, \quad \ldots \quad & , x_{1n_1} \sim g_1(x) \\
& \cdot \\
& \cdot \\
& \cdot \\
x_{q1}, \quad \ldots \quad & , x_{qn_q} \sim g_q(x) \\
x_{m1}, \quad \ldots \quad & , x_{mn_m} \sim g_m(x)
\end{aligned}
$$

$$
g_j(x) \sim \mathrm{N}(\ \mu_{\mathrm{j}},\ \sigma^2), \qquad \mathrm{j} = 1, ..., \mathrm{m}.
$$

Then, holding $g_m(x) \equiv g(x)$ as a reference:

$$\begin{aligned} g_1(x) &= \exp(\alpha_1 + \beta_1 x) g(x) \\ &\vdots \\ g_q(x) &= \exp(\alpha_q + \beta_q x) g(x) \end{aligned}$$

$$\alpha_j = \frac{\mu_m^2 - \mu_j^2}{2\,\sigma^2}, \qquad \beta_j = \frac{\mu_j - \mu_m}{\sigma^2}, \qquad j = 1, ..., q$$

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 $g(x)$:

$$\begin{aligned} g_1(x) &= \exp\{\alpha_1 + \beta_1' \boldsymbol{h}(x)\}g(x) \\ &\vdots \\ g_q(x) &= \exp\{\alpha_q + \beta_q' \boldsymbol{h}(x)\}g(x) \end{aligned}$$

Case-control: Multinomial logistic regression (Prentice and Pyke 1979).

- RV y s.t. $P(y=j) = \pi_j, \quad \sum_{j=1}^{m} \pi_j = 1.$
- Assume: For $j = 1, ..., m$, 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: $f(x|y=j) = g_j(x), \quad j = 1, ..., m$

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

$$g_j(x) = \exp(\alpha_j + \beta_j h(x)) g(x), \; j = 1, ..., q.$$

Weighted Distributions

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

$$F(y) = 1/\mu \int_0^y x dG(x),\ y \geq 0,$$

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

$$F(y) = 1/\mu \int_0^y x dG(x),\ y \geq 0,$$

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.

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:

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

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

$$\begin{aligned} f_1(x) &= d(G(x); G, F_1)g(x) \\ &\cdot \\ &\cdot \\ &\cdot \\ f_q(x) &= d(G(x); G, F_q)g(x) \end{aligned}$$

Outline

1 Density Ratio Examples

2 Semiparametric Statistical Formulation

3 Combined semiparametric density estimators

4 Semiparametric regression

5 Application to Testicular Germ Cell Cancer

$m = q + 1$ independent p-dim multivariate samples:

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

and

$$\mathbf{x}_{i1},\ \mathbf{x}_{i2}, \ldots,\ \mathbf{x}_{in_i} \overset{iid}{\sim} g_i$$

Reference or baseline probability density function (pdf):

$$g \equiv g_m(\mathbf{x}) \equiv g_m(x_1, \ldots, x_p)$$

Density ratio model:

$$\frac{g_i(\mathbf{x})}{g(\mathbf{x})} = w(\mathbf{x}, \theta_i) \tag{2}$$

or equivalently

$$g_i(\mathbf{x}) = w(\mathbf{x}, \theta_i) g(\mathbf{x}) \tag{3}$$

1. $g_i(\mathbf{x})$, $g(\mathbf{x})$ are unknown.
2. w is a known positive and continuous function.
3. θ_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

$$\mathbf{x}_{11}, ..., \mathbf{x}_{1n_1}, ...\mathbf{x}_{m1}, ..., \mathbf{x}_{mn_m}$$

of size $n = n_1 + \cdots + n_m$ to estimate all the parameters θ_i, $i = 1, ..., q$, and all the densities $g_1, ..., g_q$ and $g_m = g$.

$$G(\mathbf{x}) \equiv G_m(\mathbf{x}),\ p_{ij} = dG(\mathbf{x}_{ij}) = dG_m(\mathbf{x}_{ij}).$$

The empirical log-likelihood is given by:

$$l = \log L = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log(p_{ij}) + \sum_{i=1}^{q}\sum_{j=1}^{n_i} \log(w(\mathbf{x}_{ij}, \theta_i)) \qquad (4)$$

subject to the constraints:

$$p_{ij} \geq 0,\ \sum_{i=1}^{m}\sum_{j=1}^{n_i} p_{ij} = 1,\ \sum_{i=1}^{m}\sum_{j=1}^{n_i} p_{ij} w(\mathbf{x}_{ij}, \theta_k) = 1 \text{ for } k = 1, \ldots, q. \qquad (5)$$

$$\hat{p}_{ij} = \frac{1}{n} \frac{1}{1 + \sum_{k=1}^{q} \hat{\mu}_k \left[w(\mathbf{x}_{ij}, \hat{\theta}_k) - 1 \right]}, \tag{6}$$

$$\hat{G}(\mathbf{x}) = \hat{G}_m(\mathbf{x}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{p}_{ij} I(\mathbf{x}_{ij} \leq \mathbf{x}) \tag{7}$$

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

$$\hat{G}_l(\mathbf{x}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{p}_{ij} w(\mathbf{x}_{ij}, \hat{\theta}_l) I(\mathbf{x}_{ij} \leq \mathbf{x}) \tag{8}$$

Outline

1 Density Ratio Examples

2 Semiparametric Statistical Formulation

3 Combined semiparametric density estimators

4 Semiparametric regression

5 Application to Testicular Germ Cell Cancer

Traditional single sample kernel density estimator:

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_n^p} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right) \tag{9}$$

h_n is a sequence of bandwidths such that:

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

Under certain conditions, $\hat{f}(\mathbf{x})$ is a consistent estimator of $f(\mathbf{x})$ (Parzen 1962, Shao 2003).

Fokianos (2004), Cheng and Chu (2004), Qin and Zhang (2005), Voulgaraki, Kedem, Graubard (VKG) (2012), use the the probabilities $\hat{p}_{ij}$ instead of $1/n$:

$$\hat{g}_l(\mathbf{x}) = \frac{1}{h_n^p} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{p}_{ij} \hat{w}_l(\mathbf{x}_{ij}) K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \tag{10}$$

VKG 2012: Assume that $K(\cdot)$ is a nonnegative bounded symmetric function with $\int K(\mathbf{x})d\mathbf{x} = 1$, $\int \mathbf{x}'\mathbf{x}K(\mathbf{x})d\mathbf{x} = k_2 > 0$. Assume that g_l is continuous at $\mathbf{x}$ and twice differentiable in a neighborhood of $\mathbf{x}$. If, as $n \to \infty$, $h_n = O(n^{-\frac{1}{4+p}})$, then

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

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 $\mathbf{x}$.

The mean integrated square error (MISE) is defined as:

$$MISE(\hat{g}_l(\mathbf{x})) = E\left(\int \left|\hat{g}_l(\mathbf{x}) - g_l(\mathbf{x})\right|^2 d\mathbf{x}\right) \tag{11}$$

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

$$h_n^* = \left(\frac{(p/n)\int w_l(\mathbf{x}) g_l(\mathbf{x}) / \left[\sum_{k=1}^m \zeta_k w_k(\mathbf{x})\right] d\mathbf{x} \int K^2(\mathbf{u}) d\mathbf{u}}{\int \left(\int \mathbf{u}' (\partial^2 g_l(\mathbf{x}) / \partial \mathbf{x} \partial \mathbf{x}') \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x}} \right)^{\frac{1}{4+p}} \tag{12}$$

For sufficient large n, an alternative way to find h_n is by minimizing

$$h_n^{-p} \sum_{i=1}^{n} \sum_{i'=1}^{n} \hat{p}(\mathbf{t}_i) \hat{w}_l(\mathbf{t}_i) \hat{p}(\mathbf{t}_{i'}) \hat{w}_l(\mathbf{t}_{i'}) \int K(\mathbf{z}) K\left(\mathbf{z} + \frac{\mathbf{t}_i - \mathbf{t}_{i'}}{h_n}\right) d\mathbf{z}$$
$$- \frac{2}{n_l(n_l - 1)h_n^p} \sum_{i \neq j} K\left(\frac{\mathbf{x}_{li} - \mathbf{x}_{lj}}{h_n}\right). \quad (13)$$

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 $\hat{g}_l(\mathbf{x})$ is more efficient than $\hat{f}(\mathbf{x})$, i.e for every l

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

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

Diagnostic Plots and Goodness of Fit

Define

$$R^2_{\alpha,k} = 1 - \exp\left\{ - \left(\frac{x_\alpha}{n_i - x_\alpha} \right)^k \right\} \tag{14}$$

- n_i: Sample size of ith sample.
- x_α: 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.
- $k > 0$ abd α chosen by the user.
- $R^2_{\alpha,k}$ takes values between 0 and 1.
- Computing $R^2_{\alpha,k}$ is both simple and fast.

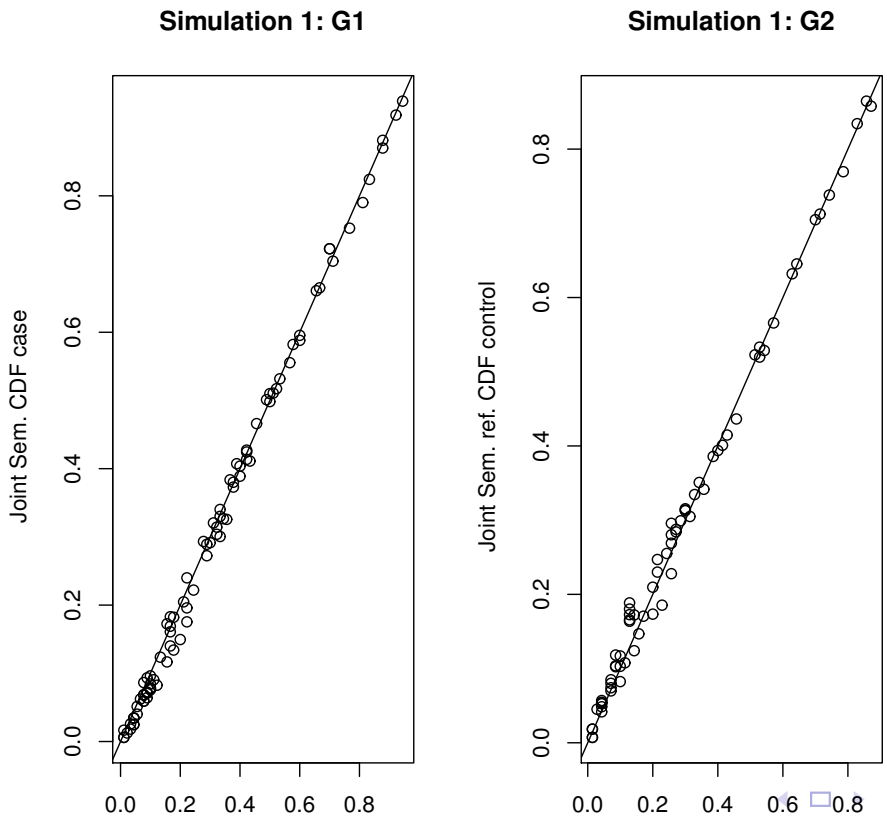
Qin and Zhang (1997):

$$R_3^2 = \exp(-\sqrt{n} \cdot \max |\tilde{G}_i - \hat{G}_i|) \tag{15}$$

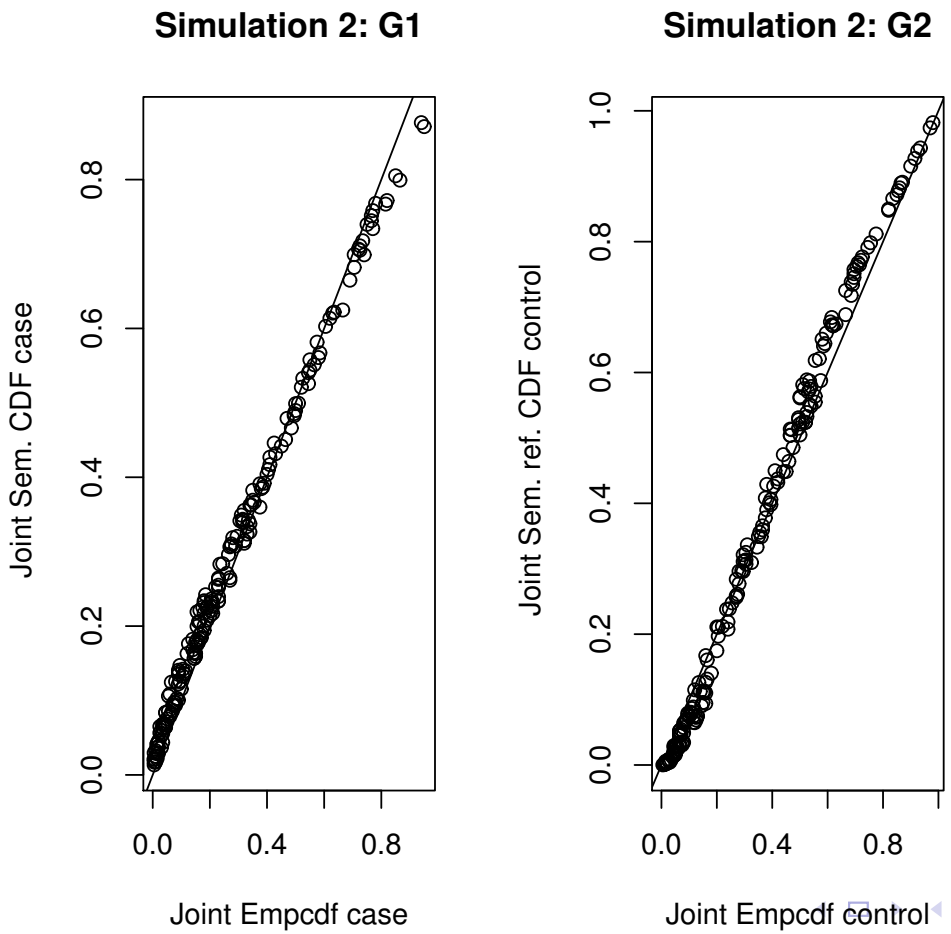
Clearly, R_3^2 takes values between 0 and 1.

Simulation Results: $\hat{G}_i$ vs. $\tilde{G}_i$, $i = 1, 2$

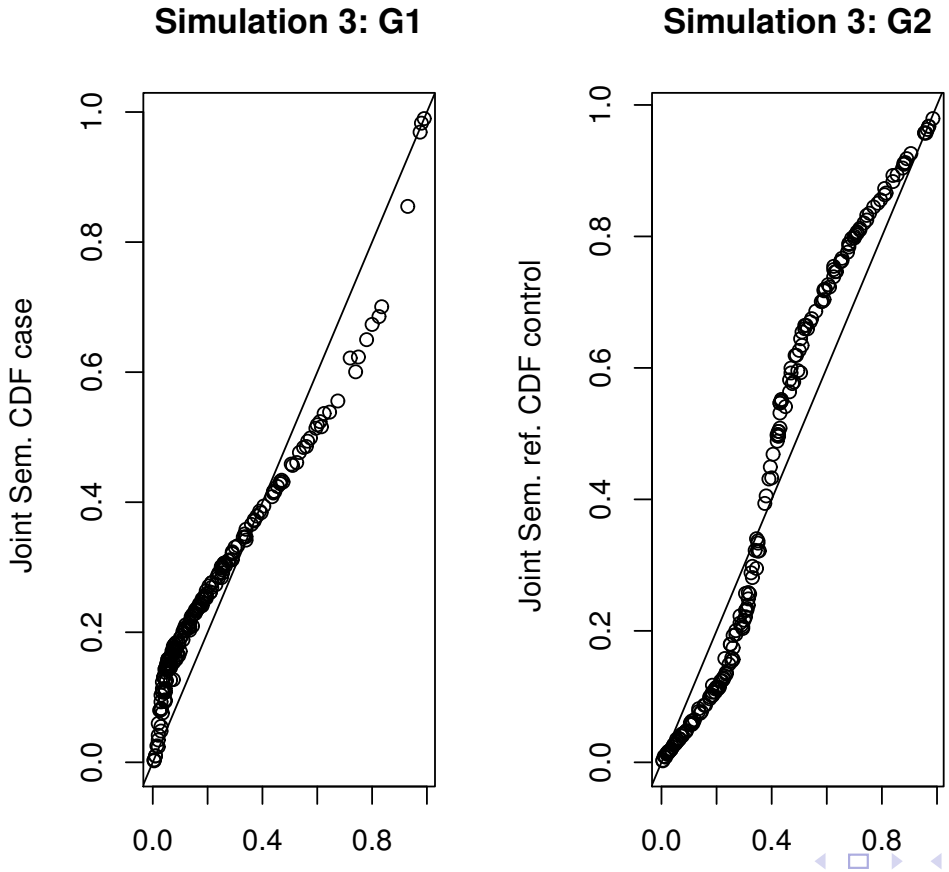
Simulation 1: $g_1 \sim N((0,0)', \mathbf{\Sigma}))$ (case), $g_2 \sim N((0,0)', \mathbf{\Sigma}))$ (control),
$\mathbf{\Sigma} = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$, $n_1 = 40$, $n_2 = 30$.



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



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



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 $n_1 = 200$, $n_2 = 200$.

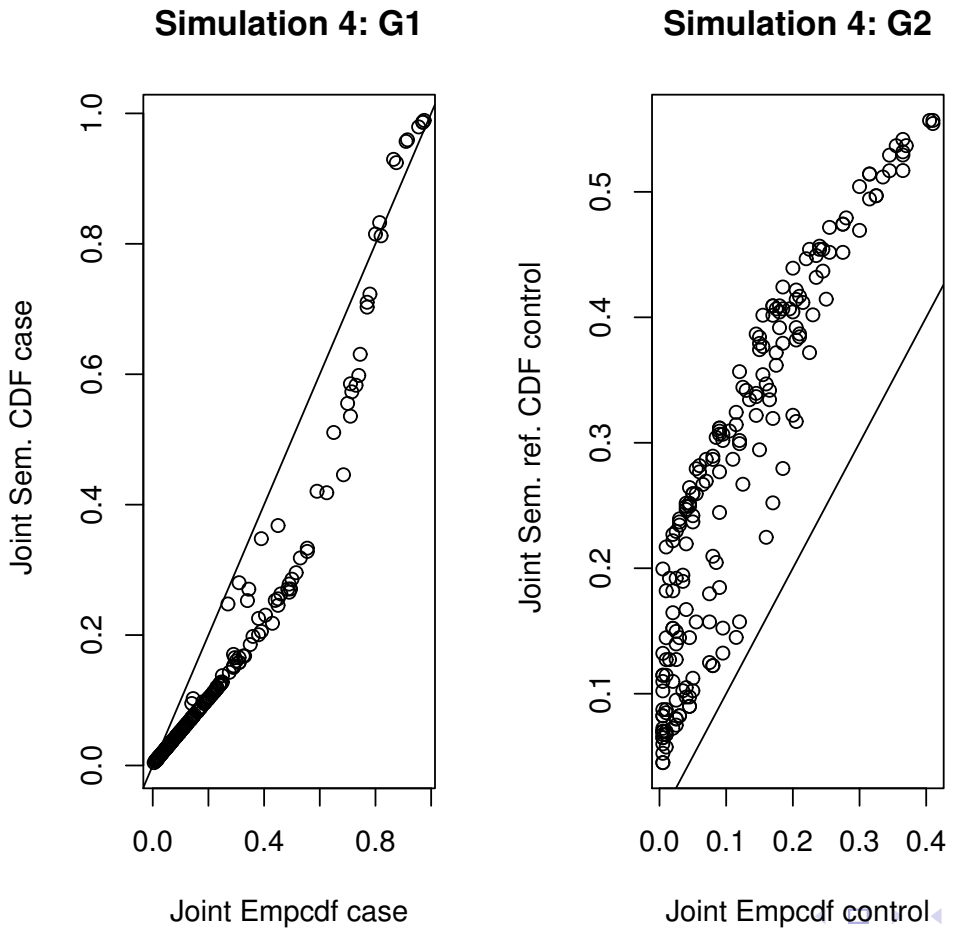


Table: Comparison of R_3^2 and $R_{.05,2}^2$ for 100 repetitions of case and control.

Simulation	Group	R_3^2	$R_{.05,2}^2$
(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

Bandwidth (BW) selection using cross validation.

	Case			Control		
	Same BW h	Diff. BW's h_1	h_2	Same BW h	Diff. BW's h_1	h_2
Leave One Out						
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
Alternative (12)						
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

Outline

1 Density Ratio Examples

2 Semiparametric Statistical Formulation

3 Combined semiparametric density estimators

4 Semiparametric regression

5 Application to Testicular Germ Cell Cancer

Suppose we have m multivariate samples of p-dim vectors:

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

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 $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ assume

$$\frac{g_i(\mathbf{x})}{g(\mathbf{x})} = w(\mathbf{x}, \alpha_i, \boldsymbol{\beta}_i) \equiv \exp(\alpha_i + \boldsymbol{\beta}_i' \mathbf{x}), \quad i = 1, ..., q \qquad (16)$$

where $\mathbf{x} = (x_1, \ldots, x_{(p-1)}, y)'$ and $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})'$.

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 $\hat{\boldsymbol{\beta}}_j$:

$$\frac{\partial l}{\partial \alpha_j} = -\sum_{i=1}^{n} \frac{\rho_j w_j(\mathbf{t}_i)}{1 + \rho_1 w_1(\mathbf{t}_i) + \cdots + \rho_q w_q(\mathbf{t}_i)} + n_j = 0 \tag{17}$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}_j} = -\sum_{i=1}^{n} \frac{\rho_j w_j(\mathbf{t}_i)\mathbf{t}_i}{1 + \rho_1 w_1(\mathbf{t}_i) + \cdots + \rho_q w_q(\mathbf{t}_i)} + \sum_{i=1}^{n_j} (x_{ji1}, \ldots, y_{ji})' = 0 \tag{18}$$

$j = 1, \ldots, q$ and

$$\rho_j = n_j / n_m$$

$$w_j(\mathbf{t}_i) = \exp(\alpha_j + \boldsymbol{\beta}_j' \mathbf{t}_i)$$

$$\hat{p}_i = \frac{1}{n_m} \cdot \frac{1}{1 + \rho_1 \hat{w}_1(\mathbf{t}_i) + \cdots + \rho_q \hat{w}_q(\mathbf{t}_i)} \tag{19}$$

$$\hat{G}(\mathbf{t}) = \frac{1}{n_m} \cdot \sum_{i=1}^{n} \frac{I(\mathbf{t}_i \le \mathbf{t})}{1 + \rho_1 \hat{w}_1(\mathbf{t}_i) + \cdots + \rho_q \hat{w}_q(\mathbf{t}_i)} \tag{20}$$

where

$$\hat{w}_j(\mathbf{t}_i) = \exp(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}_j' \mathbf{t}_i)$$

As $n \to \infty$, the estimators $\hat{\boldsymbol{\theta}} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_q, \hat{\boldsymbol{\beta}}_1, \cdots, \hat{\boldsymbol{\beta}}_q)'$ 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:

$$\hat{E}_j(y \mid x_1, \ldots, x_{(p-1)}) = \sum_i^{n_j} y_i \frac{\hat{g}_j(x_1, \ldots, x_{(p-1)}, y_i)}{\sum_{y_i} \hat{g}_j(x_1, \ldots, x_{(p-1)}, y_i)}, \; j = 1, \ldots, q, m. \tag{21}$$

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 $nh^p \to \infty$,

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

in the mean square sense.

(b) If, in addition $0 < A < g(\mathbf{x})$, then

$$\int |\hat{\mathrm{E}}(y|\mathbf{x}) - \mathrm{E}(y|\mathbf{x})| d\mathbf{x} \to 0$$

in the mean square sense.

Outline

1 Density Ratio Examples

2 Semiparametric Statistical Formulation

3 Combined semiparametric density estimators

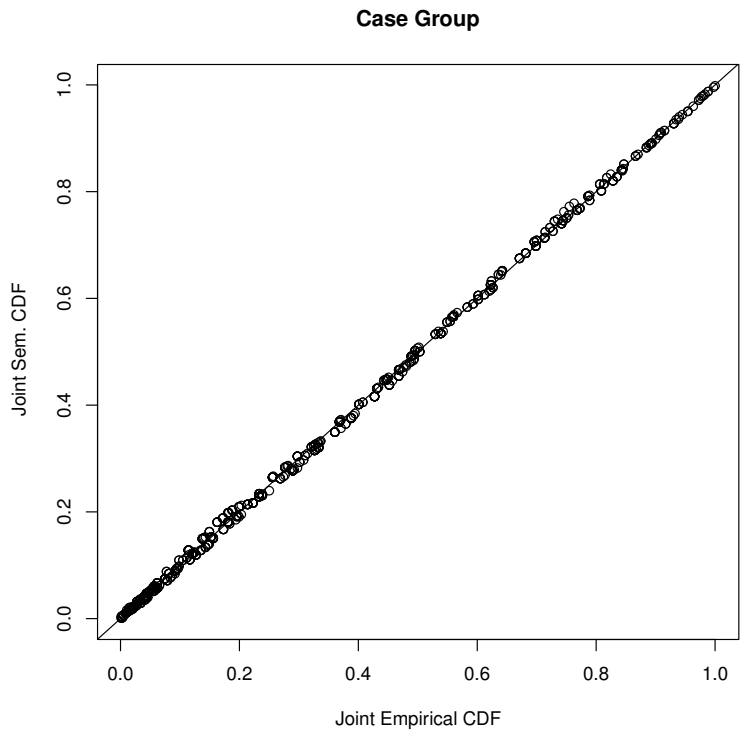
4 Semiparametric regression

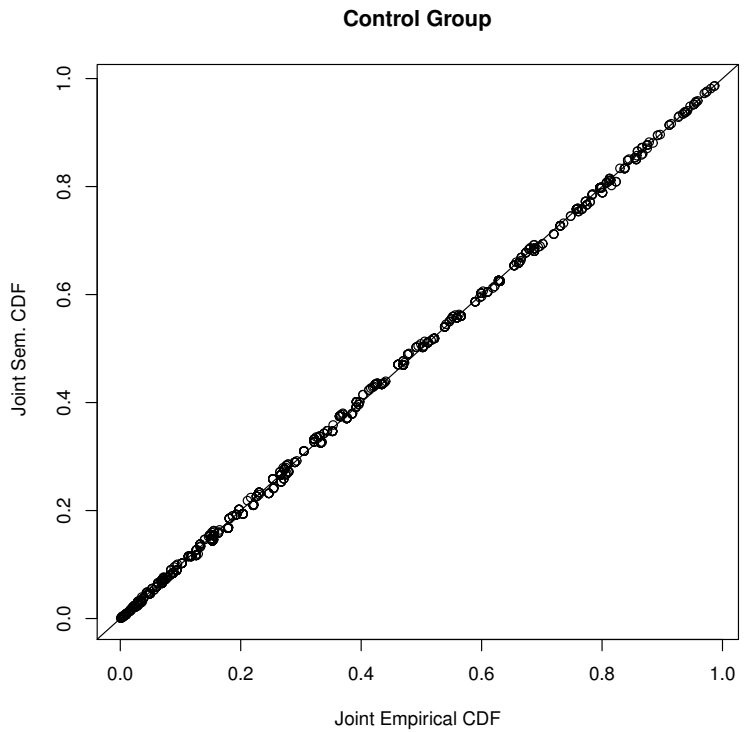
5 Application to Testicular Germ Cell Cancer

- 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).
- 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.

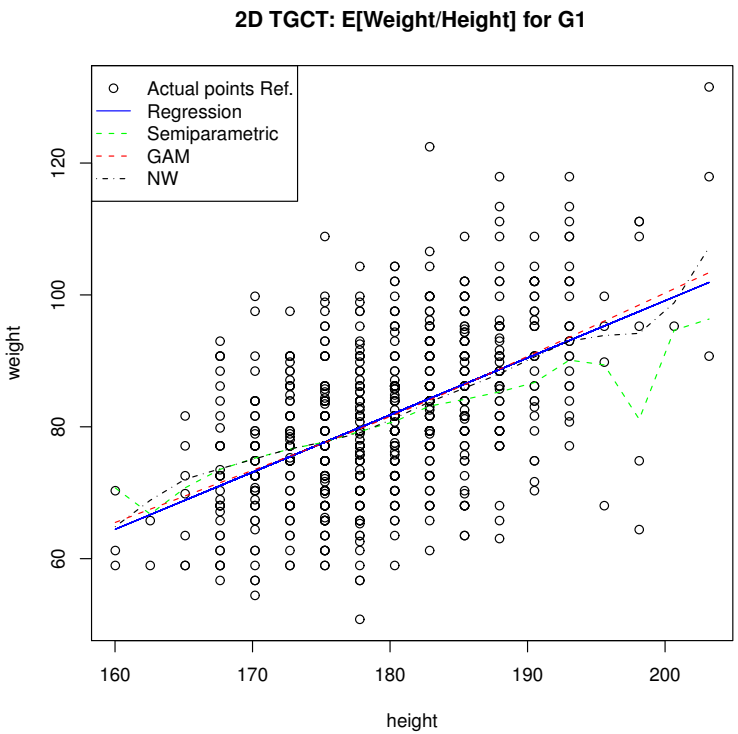
2D Problem: E(Weight | Height)

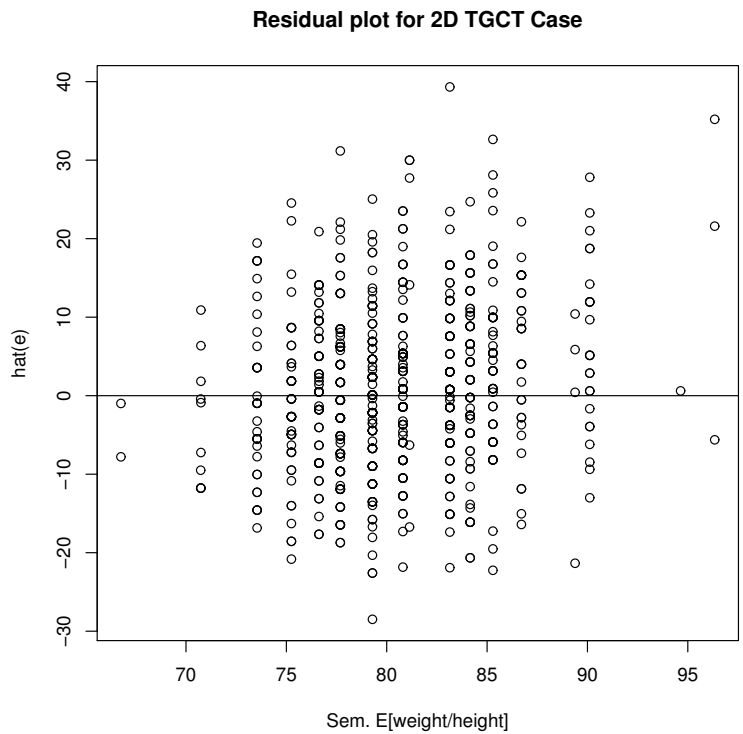
Diagnostic plots of $\hat{G}_i$ versus $\tilde{G}_i$, $i = 1, 2$ evaluated at (height,weight) pairs.



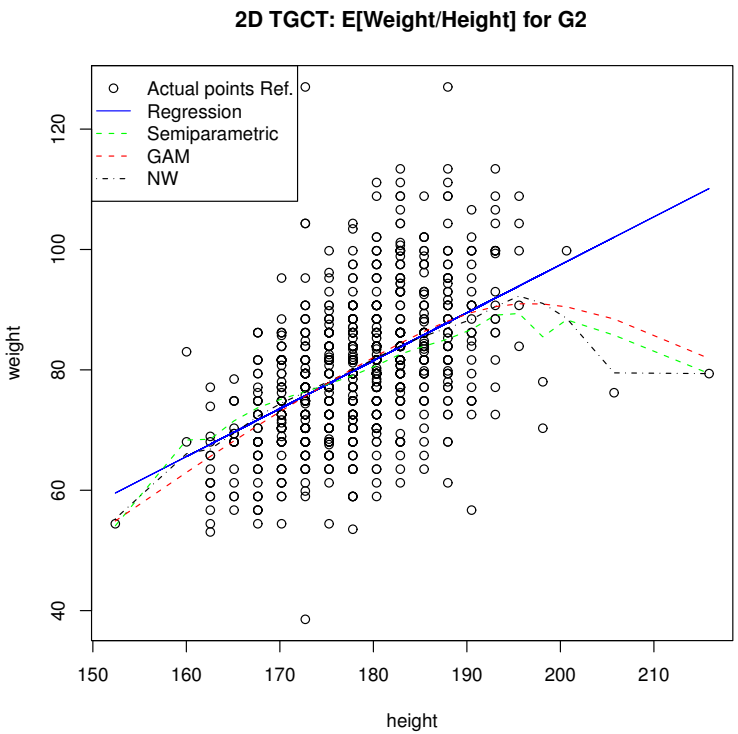


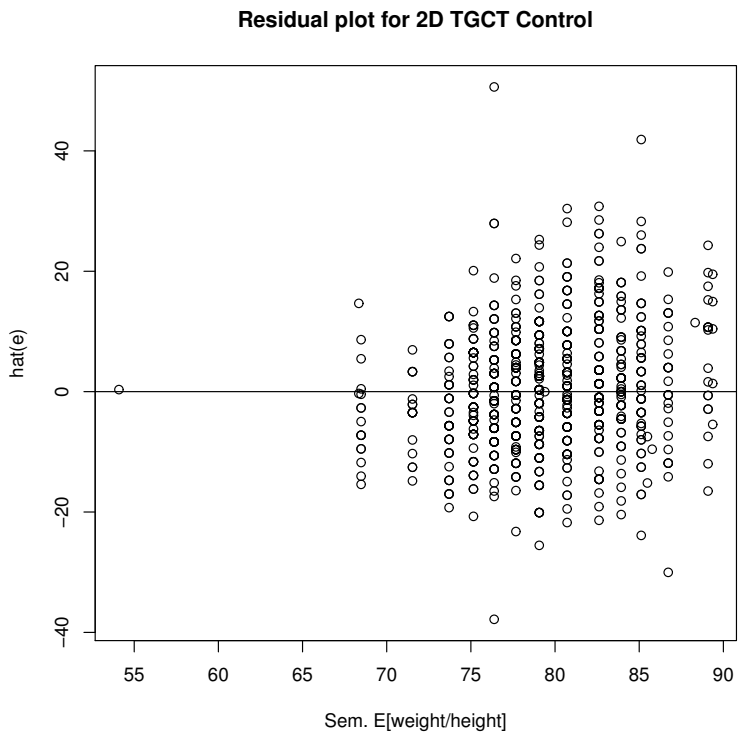
2D Case Regression





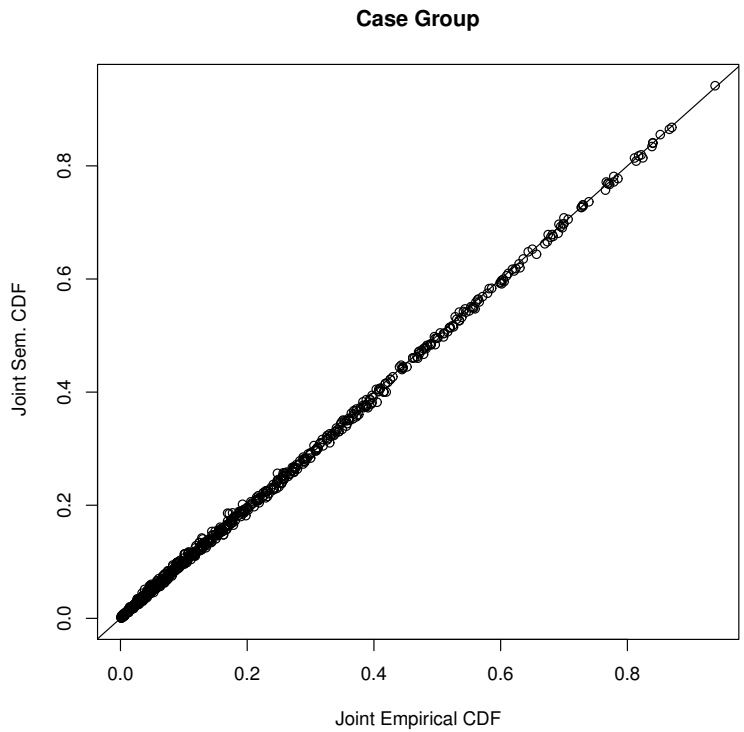
2D Control Regression

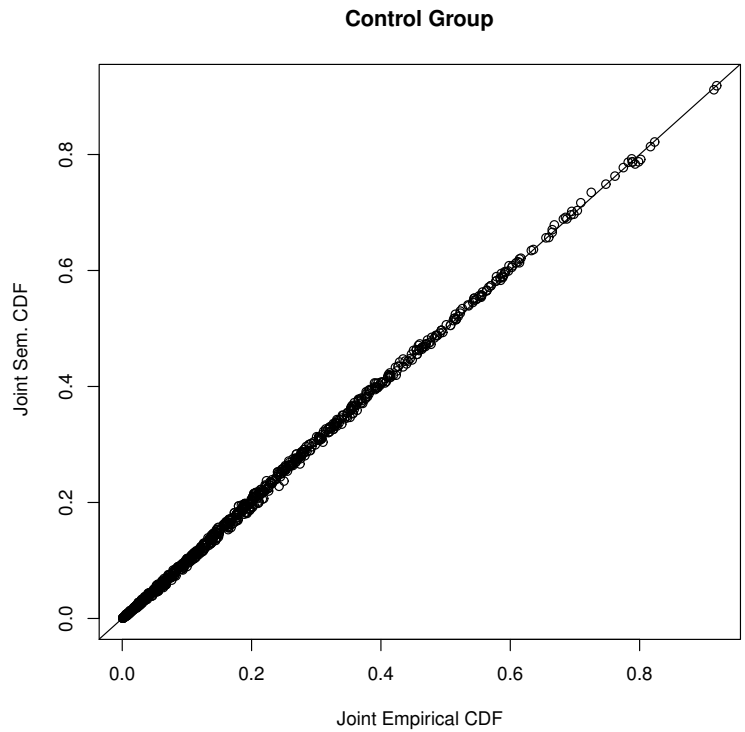




3D Problem: E(Weight | Height, Age)

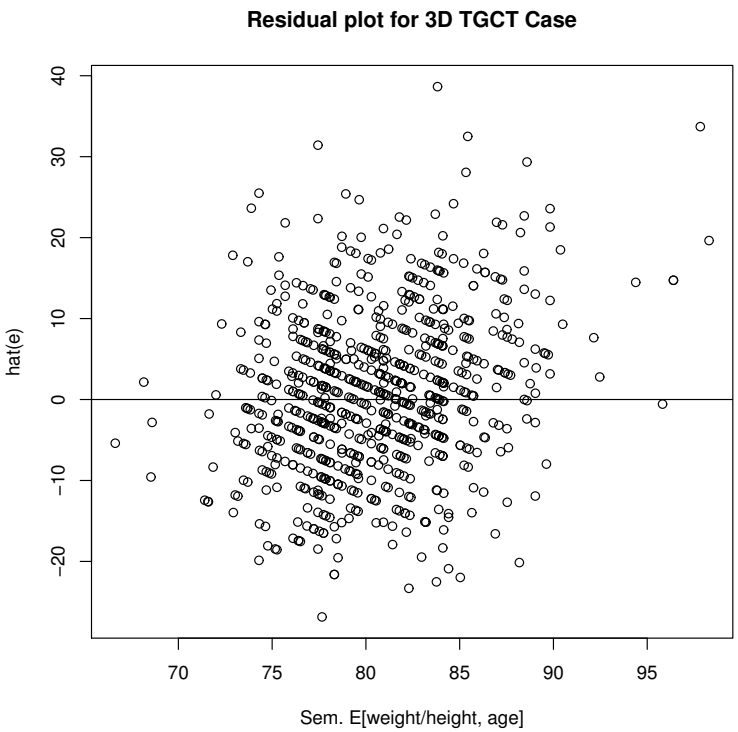
Diagnostic plots of $\hat{G}_i$ versus $\tilde{G}_i$, $i = 1, 2$ evaluated at (age,height,weight) triplets.

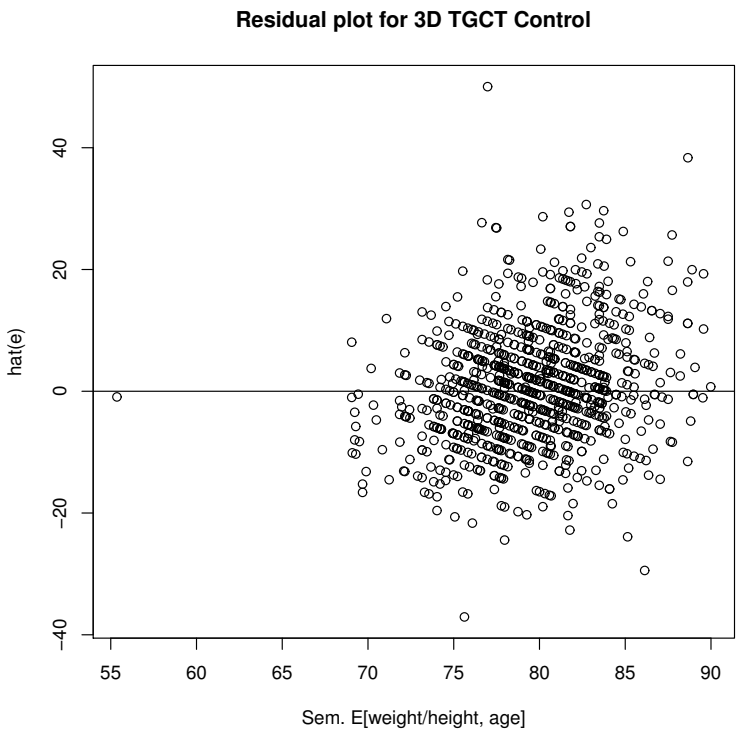




3D Residuals

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.

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

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	$\hat{E}[W \mid H, A]$	Weight	$\hat{E}[W \mid H, A]$
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	

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	$\hat{E}[W \mid H, A]$	Weight	$\hat{E}[W \mid H, A]$
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
		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	

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	$\hat{E}[W \mid H, A]$	Weight	$\hat{E}[W \mid H, A]$
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