Out of Sample Fusion

in the Estimation of Small Tail Probabilities

Benjamin Kedem

Department of Mathematics & Institute 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.)

NCSU, April 5, 2013

Abstract: "samples" not "sample"

• A great deal of the statistical literature deals with a single sample coming from a distribution, univariate or multivariate.

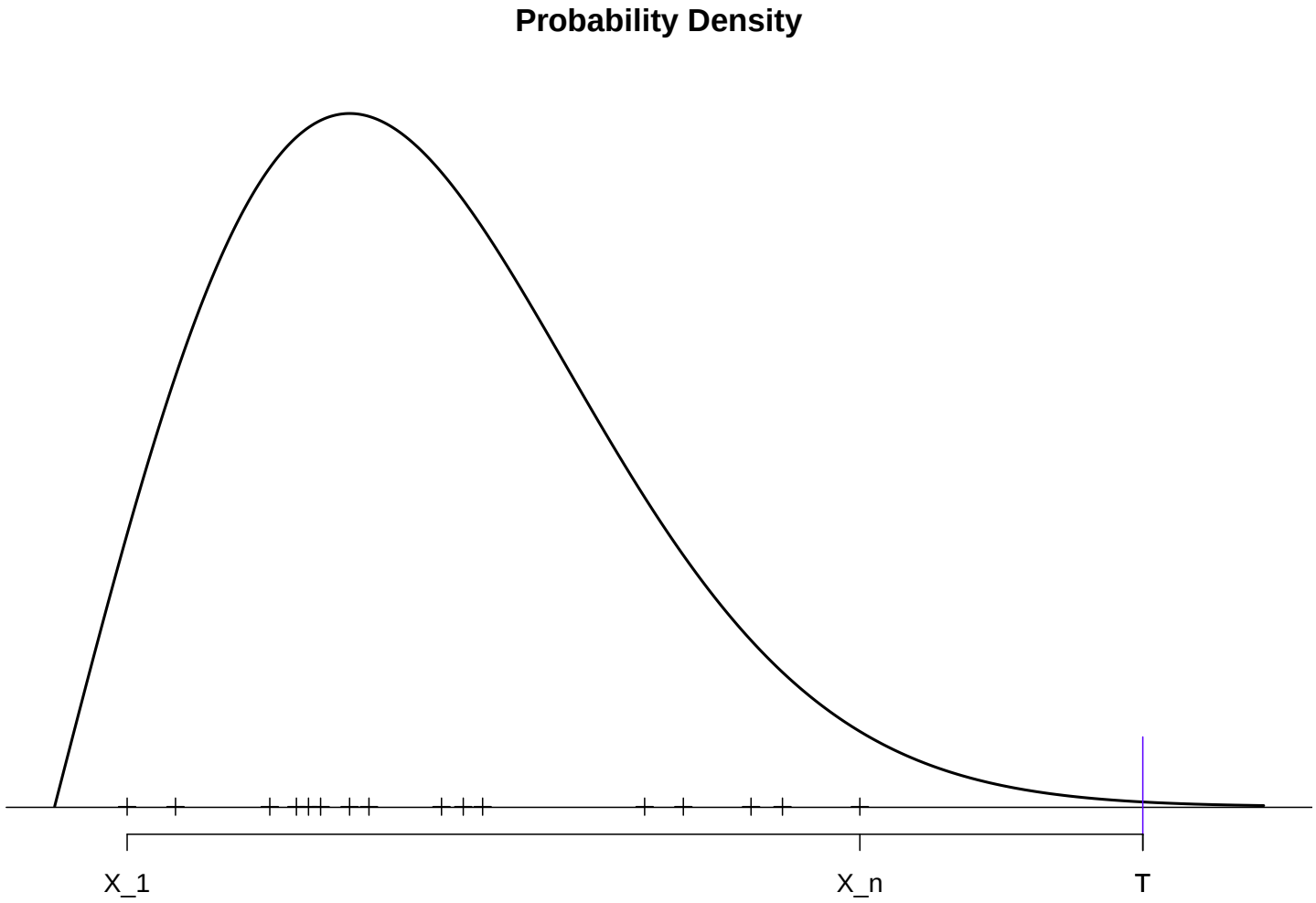
• As such, this practice neglects to bring in information from other sources, including information from **artificial data**, which could improve the desired inference.

• An **out of sample fusion method** is presented where an original real data sample is **fused** or combined with independent **computer generated samples** in the estimation of small tail probabilities assuming a density ratio model.

The basic idea: Relationships between probability distributions from many sources.

- The bootstrap is a "within sample" idea: Generation of many samples internally.
- Out of sample fusion is an "out of sample" idea: Generation of many samples externally.

We shall be concerned with the tail problem: Estimation of exceedance probabilitis $P(X > T)$.



Outline

Anything in common?

- Satellite sensors and their ground truth counterpart.
- An array of sensors sensing a common target
- Multiple filtering of a stationary signal.
- Multivariate autoregression with Gaussian noise.
- Classical one-way analysis of variance with normal data.
- k-parameter exponential families.
- Multinomial logistic regression.
- Weighted distributions and biased sampling.
- Comparison of distributions.
- **Answer**: Deviation from a reference distribution.

Anything in common?

- Satellite sensors and their ground truth counterpart.
- An array of sensors sensing a common target
- Multiple filtering of a stationary signal.
- Multivariate autoregression with Gaussian noise.
- Classical one-way analysis of variance with normal data.
- k-parameter exponential families.
- Multinomial logistic regression.
- Weighted distributions and biased sampling.
- Comparison of distributions.
- **Answer**: Deviation from a reference distribution.

Anything in common?

- Satellite sensors and their ground truth counterpart.
- An array of sensors sensing a common target
- Multiple filtering of a stationary signal.
- Multivariate autoregression with Gaussian noise.
- Classical one-way analysis of variance with normal data.
- k-parameter exponential families.
- Multinomial logistic regression.
- Weighted distributions and biased sampling.
- Comparison of distributions.
- **Answer**: Deviation from a reference distribution.

Anything in common?

- Satellite sensors and their ground truth counterpart.
- An array of sensors sensing a common target
- Multiple filtering of a stationary signal.
- Multivariate autoregression with Gaussian noise.
- Classical one-way analysis of variance with normal data.
- k-parameter exponential families.
- Multinomial logistic regression.
- Weighted distributions and biased sampling.
- Comparison of distributions.
- **Answer**: Deviation from a reference distribution.

Anything in common?

- Satellite sensors and their ground truth counterpart.
- An array of sensors sensing a common target
- Multiple filtering of a stationary signal.
- Multivariate autoregression with Gaussian noise.
- Classical one-way analysis of variance with normal data.
- k-parameter exponential families.
- Multinomial logistic regression.
- Weighted distributions and biased sampling.
- Comparison of distributions.
- **Answer**: Deviation from a reference distribution.

Anything in common?

- Satellite sensors and their ground truth counterpart.
- An array of sensors sensing a common target
- Multiple filtering of a stationary signal.
- Multivariate autoregression with Gaussian noise.
- Classical one-way analysis of variance with normal data.
- k-parameter exponential families.
- Multinomial logistic regression.
- Weighted distributions and biased sampling.
- Comparison of distributions.
- **Answer**: Deviation from a reference distribution.

Anything in common?

- Satellite sensors and their ground truth counterpart.
- An array of sensors sensing a common target
- Multiple filtering of a stationary signal.
- Multivariate autoregression with Gaussian noise.
- Classical one-way analysis of variance with normal data.
- k-parameter exponential families.
- Multinomial logistic regression.
- Weighted distributions and biased sampling.
- Comparison of distributions.
- **Answer**: Deviation from a reference distribution.

Anything in common?

- Satellite sensors and their ground truth counterpart.
- An array of sensors sensing a common target
- Multiple filtering of a stationary signal.
- Multivariate autoregression with Gaussian noise.
- Classical one-way analysis of variance with normal data.
- k-parameter exponential families.
- Multinomial logistic regression.
- Weighted distributions and biased sampling.
- Comparison of distributions.
- **Answer**: Deviation from a reference distribution.

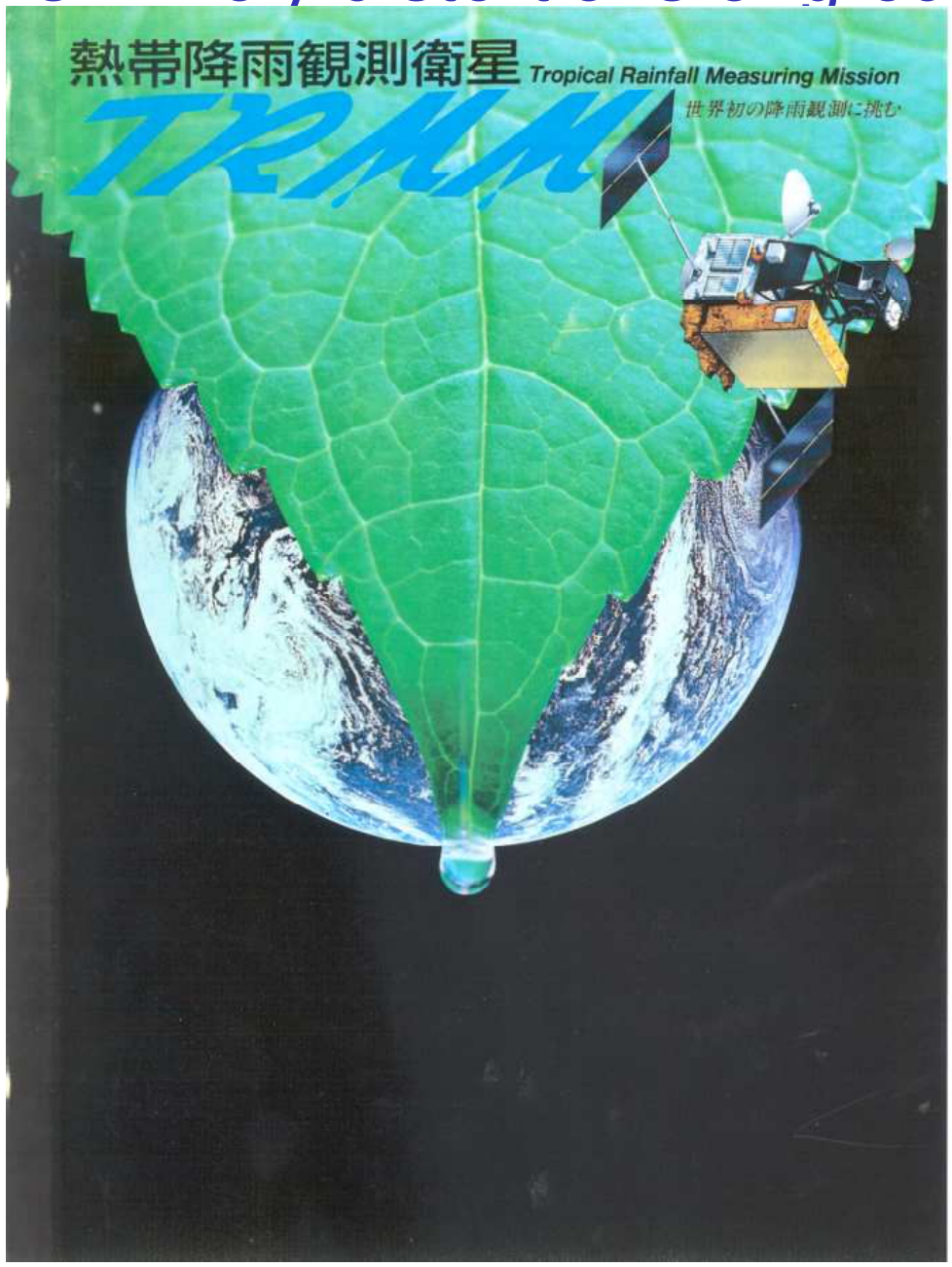
Anything in common?

- Satellite sensors and their ground truth counterpart.
- An array of sensors sensing a common target
- Multiple filtering of a stationary signal.
- Multivariate autoregression with Gaussian noise.
- Classical one-way analysis of variance with normal data.
- k-parameter exponential families.
- Multinomial logistic regression.
- Weighted distributions and biased sampling.
- Comparison of distributions.
- **Answer**: Deviation from a reference distribution.

Anything in common?

- Satellite sensors and their ground truth counterpart.
- An array of sensors sensing a common target
- Multiple filtering of a stationary signal.
- Multivariate autoregression with Gaussian noise.
- Classical one-way analysis of variance with normal data.
- k-parameter exponential families.
- Multinomial logistic regression.
- Weighted distributions and biased sampling.
- Comparison of distributions.
- **Answer**: Deviation from a reference distribution.

TRMM sensors: Likely distortions of ground truth



Reference: Ground truth

An array of sensors sensing a common target

We don't have to go very far...

- Consider a "system" of two ears, and a whisper coming from the right.
- The right ear is the more reliable sensor. Think of it as a reference.
- The left ear is a distortion of the reference.

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.

Multivariate autoregression with Gaussian noise

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

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

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

Analysis of variance

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

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

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

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

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

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

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 $x = G^{-1}(u)$:

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

- A general structure emerges of a reference behavior (distribution) and its many distortions:

$$\begin{aligned} g_1 &= w_1 g \\ &\vdots \\ g_q &= w_q g \end{aligned}$$

- How can we take advantage of this?
- Assume we have data from each of $g, g_1, g_2, \ldots, g_q$.
- Then, the relationship between a reference distribution and its distortions or tilts opens the door to a useful general statistical approach based on *fused* or *combined* data from many sources.

- A general structure emerges of a reference behavior (distribution) and its many distortions:

$$\begin{aligned} g_1 &= w_1 g \\ &\vdots \\ g_q &= w_q g \end{aligned}$$

- How can we take advantage of this?
- Assume we have data from each of $g, g_1, g_2, \ldots, g_q$.
- Then, the relationship between a reference distribution and its distortions or tilts opens the door to a useful general statistical approach based on *fused* or *combined* data from many sources.

- A general structure emerges of a reference behavior (distribution) and its many distortions:

$$\begin{aligned} g_1 &= w_1 g \\ &\vdots \\ g_q &= w_q g \end{aligned}$$

- How can we take advantage of this?
- Assume we have data from each of $g, g_1, g_2, ..., g_q$.
- Then, the relationship between a reference distribution and its distortions or tilts opens the door to a useful general statistical approach based on *fused* or *combined* data from many sources.

- A general structure emerges of a reference behavior (distribution) and its many distortions:

$$\begin{aligned} g_1 &= w_1 g \\ &\vdots \\ g_q &= w_q g \end{aligned}$$

- How can we take advantage of this?
- Assume we have data from each of $g, g_1, g_2, ..., g_q$.
- Then, the relationship between a reference distribution and its distortions or tilts opens the door to a useful general statistical approach based on *fused* or *combined* data from many sources.

The previous structure suggests the following general semiparametric problem.

- Multiple data sources: $\boldsymbol{x}_1, ..., \boldsymbol{x}_q, \boldsymbol{x}_m$.
- Data fusion: $\mathbf{t} = (t_1, ..., t_n)' \equiv (\mathbf{x}_1', ..., \mathbf{x}_q', \mathbf{x}_m')'$.
- Fused data length: $n \equiv n_1 + \cdots + n_q + n_m$.
- Assume: $\boldsymbol{x}_j \sim g_j(x), \quad j = 1, ..., q, m$.
- Reference pdf: $g_m(x) = g(x)$.
- Tilting: $g_j(x) = \exp(\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(x)) g(x), \qquad j = 1, ..., q$.

Problem: Use the fused data $\mathbf{t} = (t_1, ..., t_n)'$ to

1. Estimate the reference pdf $g(x)$ and cdf $G(x)$.

2. Estimate $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_q)'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_q')'$.

3. Test hypotheses about the $\boldsymbol{\beta}_j$, and in particular test distribution equality, H_0: $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_q = 0$

Outline

Estimation

Follow Vardi (1982,1985), Qin and Zhang (1997), FKQS (2001).

MLE of $G(x)$ can be obtained by maximizing the likelihood over the class of step cdf's with jumps at the observed values $t_1, ..., t_n$. Accordingly, if $p_i = dG(t_i)$, $i = 1, .., n$, the empirical likelihood becomes,

$$\mathcal{L}(\alpha, \beta, G) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} \exp(\alpha_1 + \beta_1' \boldsymbol{h}(x_{1j})) \cdots \prod_{j=1}^{n_q} \exp(\alpha_q + \beta_q' \boldsymbol{h}(x_{qj}))$$

1. Get p_i

Fix α, β. Maximize $\prod_{i=1}^{n} p_i$ subject to the m constraints:

$$\sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i[w_j(t_i) - 1] = 0, \quad j = 1, ..., q,$$

where

$$w_j(t_i) = \exp(\ \alpha_j + \ \beta_j' \boldsymbol{h}(t_i)), \ j = 1, ..., q.$$

Use Lagrange multipliers $\lambda_0 = n, \quad \lambda_j = \nu_j n$. Then

$$(\star) \qquad p_i = \frac{1}{n_m} \cdot \frac{1}{1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)}$$

where

$$(\star) \qquad \rho_j = n_j / n_m, \quad j = 1, ..., q.$$

2. Estimate α, β

Profile log-likelihood up to a constant as a function of α, β only:

$$l = \sum_{j=1}^{n_1} [\alpha_1 + \beta_1' \boldsymbol{h}(x_{1j})] + \cdots + \sum_{j=1}^{n_q} [\alpha_q + \beta_q' \boldsymbol{h}(x_{qj})]$$
$$- \sum_{i=1}^{n} \log[1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)]$$

Score equations for $j = 1, ..., q$:

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

$$\frac{\partial l}{\partial \beta_j} = - \sum_{i=1}^{n} \frac{\rho_j h(t_i) w_j(t_i)}{1 + \rho_1 w_1(t_i) + \cdots + \rho_q w_q(t_i)}$$
$$+ \sum_{i=1}^{n_j} h(x_{ji}) = 0$$

3. Estimate $g(x), G(x)$

The solution of the score equations gives the maximum likelihood estimators $\hat{\alpha}, \hat{\beta}$, and consequently by substitution also $\hat{p}_i$. Thus,

$$\hat{p}_{\mathbf{i}} = \frac{1}{n_m} \cdot \frac{1}{1 + \sum_{j=1}^{q} \rho_j \exp(\hat{\alpha}_j + \hat{\beta}_j' \boldsymbol{h}(t_i))}.$$

Therefore,

$$\hat{g}(x) = Kernel(\hat{p}_{\mathbf{i}})$$

and

$$\hat{G}(t) = \sum_{i=1}^{n} I(t_i \leq t) \hat{p}_{\mathbf{i}}$$

Everything is estimated from everything

The reference $G(x)$ and all the parameters, and hence all the tilted distributions, are estimated from the entire fused data $\boldsymbol{t}$. Thus $G(x)$ is estimated from the fused data $\boldsymbol{t}$ and not just from the reference sample $\boldsymbol{x}_m$, and β_1 is estimated from $\boldsymbol{t}$ and not just from $\boldsymbol{x}_m$ and $\boldsymbol{x}_1$, etc. This **borrowing of strength** leads to more precise estimation. We shall quantify this in kernel density estimation.

Some asymptotic results

Assumptions

- The second moments of $h(t)$ with respect to each distribution are finite,

$$\int h^2(t) w_j(t) dG(t) < \infty,$$

$j = 1, \ldots, m$.

- The relative sample sizes $\rho_j = n_j / n_m$ are finite and remain fixed as the total sample size $\sum_{j=1}^{m} n_j = n \to \infty$, $j = 1, \ldots, m$.

Some asymptotic results

Assumptions

- The second moments of $h(t)$ with respect to each distribution are finite,

$$\int h^2(t) w_j(t) dG(t) < \infty,$$

  $j = 1, \ldots, m$.

- The relative sample sizes $\rho_j = n_j / n_m$ are finite and remain fixed as the total sample size $\sum_{j=1}^{m} n_j = n \to \infty$, $j = 1, \ldots, m$.

Fact (QZ (1997), FKQS (2001)): Assume

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

with true parameters $\boldsymbol{\alpha}_0 = (\alpha_{10}, ..., \alpha_{q0})'$, $\boldsymbol{\beta}_0 = (\beta_{10}, ..., \beta_{q0})'$. Then under regularity conditions the MLE's $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$ are strongly consistent and asymptotically normal,

$$\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \Rightarrow \mathrm{N}\left(\mathbf{0}, \boldsymbol{S}^{-1} \boldsymbol{V} \boldsymbol{S}^{-1}\right).$$

$\boldsymbol{V}, \boldsymbol{S}$ are defined below. We sometimes write $\boldsymbol{\Sigma} = \boldsymbol{S}^{-1} \boldsymbol{V} \boldsymbol{S}^{-1}$.

$$\boldsymbol{V} \equiv Var\left[\frac{1}{\sqrt{n}}\nabla l(\boldsymbol{\alpha},\boldsymbol{\beta})\right], \quad \boldsymbol{S} \equiv \lim_{n\to\infty} -\frac{1}{n}\nabla\nabla' l(\boldsymbol{\alpha},\boldsymbol{\beta})$$

where $\boldsymbol{V}$,$\boldsymbol{S}$ are $q(1+p)\times q(1+p)$ matrices.

Remark:

The entries in $\boldsymbol{S}$ are obtained by a repeated application of the facts

$$\int dG(t) = 1, \qquad \int w_j(t)dG(t) = 1, \qquad j = 1,...,q.$$

Remark:

It should be noted that due to profiling, the matrix $\boldsymbol{S}$ is not the usual information matrix but it plays a similar role.

Define $\rho_m \equiv 1,\ w_m(t) \equiv 1,$

$$E_j[\boldsymbol{h}(t)] \equiv \int \boldsymbol{h}(t) w_j(t) dG(t)$$

and,

$$\begin{aligned}
A_0(j,j') &\equiv \int \frac{w_j(t) w_{j'}(t) dG(t)}{1 + \sum_{k=1}^{q} \rho_k w_k(t)} \\
\boldsymbol{A}_1(j,j') &\equiv \int \frac{\boldsymbol{h}(t) w_j(t) w_{j'}(t) dG(t)}{1 + \sum_{k=1}^{q} \rho_k w_k(t)} \\
\boldsymbol{A}_2(j,j') &\equiv \int \frac{\boldsymbol{h}(t) \boldsymbol{h}'(t) w_j(t) w_{j'}(t) dG(t)}{1 + \sum_{k=1}^{q} \rho_k w_k(t)}
\end{aligned}$$

for $j, j' = 1, ..., q$.

The entries in $\boldsymbol{V}$:

$$
\begin{aligned}
\frac{1}{n} Var\left(\frac{\partial l}{\partial \alpha_j}\right) &= \frac{\rho_j^2}{1+\sum_{k=1}^{q}\rho_k}\{A_0(j,j) - \sum_{r=1}^{m}\rho_r A_0^2(j,r)\} \\
\frac{1}{n} Cov\left(\frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \alpha_{j'}}\right) &= \frac{\rho_j \rho_{j'}}{1+\sum_{k=1}^{q}\rho_k}\{A_0(j,j') \\
&- \sum_{r=1}^{m}\rho_r A_0(j,r)A_0(j',r)\} \\
\frac{1}{n} Cov\left(\frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \boldsymbol{\beta}_j}\right) &= \frac{\rho_j^2}{1+\sum_{k=1}^{q}\rho_k}\{A_0(j,j)E_j[\boldsymbol{h}'(t)] \\
&- \sum_{r=1}^{m}\rho_r A_0(j,r)\boldsymbol{A}_1'(j,r)\} \\
\frac{1}{n} Cov\left(\frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \boldsymbol{\beta}_{j'}}\right) &= \frac{\rho_j \rho_{j'}}{1+\sum_{k=1}^{q}\rho_k}\{A_0(j,j')E_{j'}[\boldsymbol{h}'(t)] \\
&- \sum_{r=1}^{m}\rho_r A_0(j,r)\boldsymbol{A}_1'(j',r)\}
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{n} Cov\left(\frac{\partial l}{\partial \boldsymbol{\beta}_j}, \frac{\partial l}{\partial \boldsymbol{\beta}_{j'}}\right) &= \frac{\rho_j \rho_{j'}}{1 + \sum_{k=1}^{q} \rho_k} \{ -\boldsymbol{A}_2(j, j') + E_j[\boldsymbol{h}(t)] \boldsymbol{A}_1'(j, j') \\
&+ \boldsymbol{A}_1(j, j') E_{j'}[\boldsymbol{h}'(t)] \\
&- \sum_{r=1}^{m} \rho_r \boldsymbol{A}_1(j, r) \boldsymbol{A}_1'(j', r) \} \\
&+ \frac{1}{n} \sum_{i=1}^{n_j} \sum_{k=1}^{n_{j'}} Cov[\boldsymbol{h}(\epsilon_{ji}), \boldsymbol{h}(\epsilon_{j'k})]
\end{aligned}
$$

The last term is 0 for $j \neq j'$ and $(n_j/n) Var[\boldsymbol{h}(\epsilon_{j1})]$ for $j = j'$.

The entries in $\boldsymbol{S}$:

$$-\frac{1}{n}\frac{\partial^2 l}{\partial \alpha_j^2} \quad \to \quad \frac{\rho_j}{1+\sum_{k=1}^q \rho_k} \int \frac{[1+\sum_{k\neq j}^q \rho_k w_k(t)] w_j(t)}{1+\sum_{k=1}^q \rho_k w_k(t)} dG(t)$$

$$-\frac{1}{n}\frac{\partial^2 l}{\partial \alpha_j \alpha_{j'}} \quad \to \quad \frac{-\rho_j \rho_{j'}}{1+\sum_{k=1}^q \rho_k} \int \frac{w_j(t) w_{j'}(t)}{1+\sum_{k=1}^q \rho_k w_k(t)} dG(t)$$

$$-\frac{1}{n}\frac{\partial^2 l}{\partial \alpha_j \partial \boldsymbol{\beta}_j'} \quad \to \quad \frac{\rho_j}{1+\sum_{k=1}^q \rho_k} \int \frac{[1+\sum_{k\neq j}^q \rho_k w_k(t)] w_j(t) \boldsymbol{h}'(t)}{1+\sum_{k=1}^q \rho_k w_k(t)} dG(t)$$

$$-\frac{1}{n}\frac{\partial^2 l}{\partial \alpha_j \partial \boldsymbol{\beta}_{j'}'} \quad \to \quad \frac{-\rho_j \rho_{j'}}{1+\sum_{k=1}^q \rho_k} \int \frac{w_j(t) w_{j'}(t) \boldsymbol{h}'(t)}{1+\sum_{k=1}^q \rho_k w_k(t)} dG(t)$$

$$-\frac{1}{n}\frac{\partial^2 l}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_j'} \quad \to \quad \frac{\rho_j}{1+\sum_{k=1}^q \rho_k} \int \frac{[1+\sum_{k\neq j}^q \rho_k w_k(t)] w_j(t) \boldsymbol{h}(t) \boldsymbol{h}'(t)}{1+\sum_{k=1}^q \rho_k w_k(t)} dG(t)$$

$$-\frac{1}{n}\frac{\partial^2 l}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_{j'}'} \quad \to \quad \frac{-\rho_j \rho_{j'}}{1+\sum_{k=1}^q \rho_k} \int \frac{w_j(t) w_{j'}(t) \boldsymbol{h}(t) \boldsymbol{h}'(t)}{1+\sum_{k=1}^q \rho_k w_k(t)} dG(t)$$

Theorem 1 (B. Zhang 2000, G. Lu 2007)

The process $\sqrt{n}(\hat{G}(t) - G(t))$ converges weakly to a zero-mean Gaussian process in $D[-\infty, \infty]$, with covariance matrix given by

$$\begin{aligned}
&\mathrm{Cov}\{\sqrt{n}(\hat{G}(t) - G(t)), \sqrt{n}(\hat{G}(s) - G(s))\} = \\
&\sum_{k=0}^{m} \rho_k \left(G(t \wedge s) - G(t)G(s) - \sum_{j=1}^{m} \rho_j A_j(t \wedge s) \right) \\
&+ \left(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes \mathbf{1}_p) \right) S^{-1} \begin{pmatrix} \boldsymbol{\rho}\bar{A}(t) \\ (\boldsymbol{\rho} \otimes \mathbf{1}_p)\bar{B}(t) \end{pmatrix}. \quad (2)
\end{aligned}$$

where $\boldsymbol{\rho} = \mathrm{diag}\{\rho_1, \ldots, \rho_m\}$, ρ_j's are sample fractions, $\bar{A}$ and $\bar{B}$ are vectors of first and second weighted moments of the distortion function $\boldsymbol{h}$ with respect to different samples, and $\mathbf{1}_p = (1, ..., 1)'$. (Note: here $\boldsymbol{h}$ is vector-valued).

Hypothesis testing

Under $\mathrm{H}_0 : \boldsymbol{\beta} = \mathbf{0}$, all the moments are taken with respect to the reference g.

Define a $q \times q$ matrix $\mathbf{A}_{11}$ whose jth diagonal element is

$$\frac{\rho_j[1 + \sum_{k \neq j}^{q} \rho_k]}{[1 + \sum_{k=1}^{q} \rho_k]^2}.$$

For $j \neq j'$, the jj' element is

$$\frac{-\rho_j \rho_{j'}}{[1 + \sum_{k=1}^{q} \rho_k]^2}.$$

The elements are bounded by 1 and the matrix is nonsingular,

$$|\mathbf{A}_{11}| = \frac{\prod_{k=1}^{q} \rho_k}{[1 + \sum_{k=1}^{q} \rho_k]^m} > 0.$$

Under $\mathrm{H}_0 : \boldsymbol{\beta} = (\boldsymbol{\beta}_1^{'}, ..., \boldsymbol{\beta}_{\mathrm{q}}^{'})' = \mathbf{0}$,

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{11} \otimes E[\boldsymbol{h}'(t)] \\ \boldsymbol{A}_{11} \otimes E[\boldsymbol{h}(t)] & \boldsymbol{A}_{11} \otimes E[\boldsymbol{h}(t)\boldsymbol{h}'(t)] \end{pmatrix}$$

and

$$\boldsymbol{V} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{11} \otimes Var[\boldsymbol{h}(t)] \end{pmatrix}$$

$$(\star) \quad \mathcal{X}_1 = n\hat{\boldsymbol{\beta}}'(\boldsymbol{A}_{11} \otimes Var[\boldsymbol{h}(t)])\hat{\boldsymbol{\beta}} \tag{3}$$

$Var[\boldsymbol{h}(t)]$ is the covariance matrix of $\boldsymbol{h}(t)$, and all moments are evaluated with respect to the reference distribution.

$$\mathcal{X}_1 \longrightarrow \chi^2_{(qp)}$$

Testing Linear Hypotheses (vector β)

$$\mathcal{X}_{\mathbf{2}} = n(\boldsymbol{H}\hat{\theta} - \mathbf{c})'(\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}')^{-1}(\boldsymbol{H}\hat{\theta} - \mathbf{c}) \tag{4}$$

where $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_q, \beta_1', \ldots, \beta_q')'$, $\boldsymbol{H}$ is $p' \times [(1+p)q)]$ predetermined matrix of rank p', $p' < (1+p)q$, $\mathbf{c}$ is a vector in $\Re^{p'}$, and the variance-covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{S}^{-1}\boldsymbol{V}\boldsymbol{S}^{-1}$. It follows under $\mathrm{H}_0 : \boldsymbol{H}\boldsymbol{\theta} = \mathbf{c}$ that χ_2 is asymptotically distributed as χ^2 with (p') degrees of freedom provided the inverse exists , and H_0 is rejected for large values.

Likelihood Ratio Test ($\ell = l$)

$$
\begin{aligned}
LR &= -2[\ell(0,0) - \ell(\hat{\alpha}, \hat{\boldsymbol{\beta}})] = \\
&- 2\sum_{i=1}^{n} \log[1 + \rho_1 \hat{w}_1(t_i) + \ldots + \rho_q \hat{w}_q(t_i)] \\
&+ 2\sum_{i=1}^{q}\sum_{j=1}^{n_i}[\hat{\alpha}_i + \hat{\boldsymbol{\beta}}_i' \boldsymbol{h}(x_{ij})] + 2n \log\left[1 + \sum_{i=1}^{q} \rho_i\right]
\end{aligned}
$$

Under $\mathrm{H}_0 : \boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_q')' = \mathbf{0}$, LR is asymptotically approximately distributed as $\chi^2_{(qp)}$.

Traditional single sample kernel density estimator (Rosenblatt 1956, Parzen 1962)

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_n^p} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right) \tag{5}$$

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. $\int_{\mathbf{R}^p} K(\mathbf{x}) d\mathbf{x} = 1$.

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

$$\hat{g}_l(\mathbf{x}) = \frac{1}{h_n^p} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{p}_{ij} \hat{w}_l(\mathbf{x}_{ij}) K\left(\frac{\mathbf{x} - \mathbf{x}_{ij}}{h_n}\right) \tag{6}$$

Fact: The semiparametric kernel density estimator has the same asymptotic bias but smaller variance as compared with the traditional kernel density estimator.

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:

$$MISE(\hat{g}_l(\mathbf{x})) = E\left(\int \left|\hat{g}_l(\mathbf{x}) - g_l(\mathbf{x})\right|^2 d\mathbf{x}\right) \tag{7}$$

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

$$h_n^* = \left(\frac{(p/n)\int w_l(\mathbf{x}) g_l(\mathbf{x}) / \left[\sum_{k=1}^{m} \zeta_k w_k(\mathbf{x})\right] d\mathbf{x} \int K^2(\mathbf{u}) d\mathbf{u}}{\int \left(\int \mathbf{u}' (\partial^2 g_l(\mathbf{x}) / \partial \mathbf{x} \partial \mathbf{x}') \mathbf{u} K(\mathbf{u}) d\mathbf{u} \right)^2 d\mathbf{x}} \right)^{\frac{1}{4+p}} \tag{8}$$

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

$$h_n^{-p} \sum_{i=1}^{n} \sum_{i'=1}^{n} \hat{p}(\mathbf{t}_i) \hat{w}_l(\mathbf{t}_i) \hat{p}(\mathbf{t}_{i'}) \hat{w}_l(\mathbf{t}_{i'}) \int K(\mathbf{z}) K\left(\mathbf{z} + \frac{\mathbf{t}_i - \mathbf{t}_{i'}}{h_n}\right) d\mathbf{z}$$
$$- \frac{2}{n_l(n_l - 1)h_n^p} \sum_{i \neq j} K\left(\frac{\mathbf{x}_{li} - \mathbf{x}_{lj}}{h_n}\right). \quad (9)$$

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

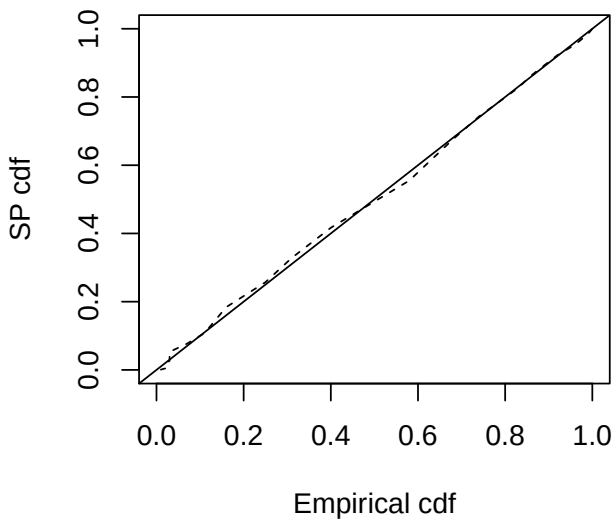
$$eff(\hat{f}, \hat{g}_l) \equiv \frac{AMISE^*(\hat{g}_l)}{AMISE^*(\hat{f})} \leq 1$$

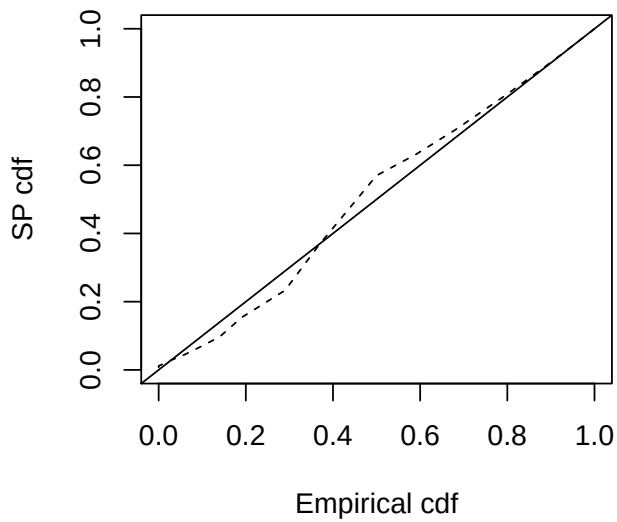
where $AMISE^*$ is the optimal $AMISE$.

Graphical Goodness of Fit: $\hat{G}$ vs. $\tilde{G}$

Plots of pairs $(\tilde{G}(x), \hat{G}(x))$ (Empirical cdf vs SP cdf).

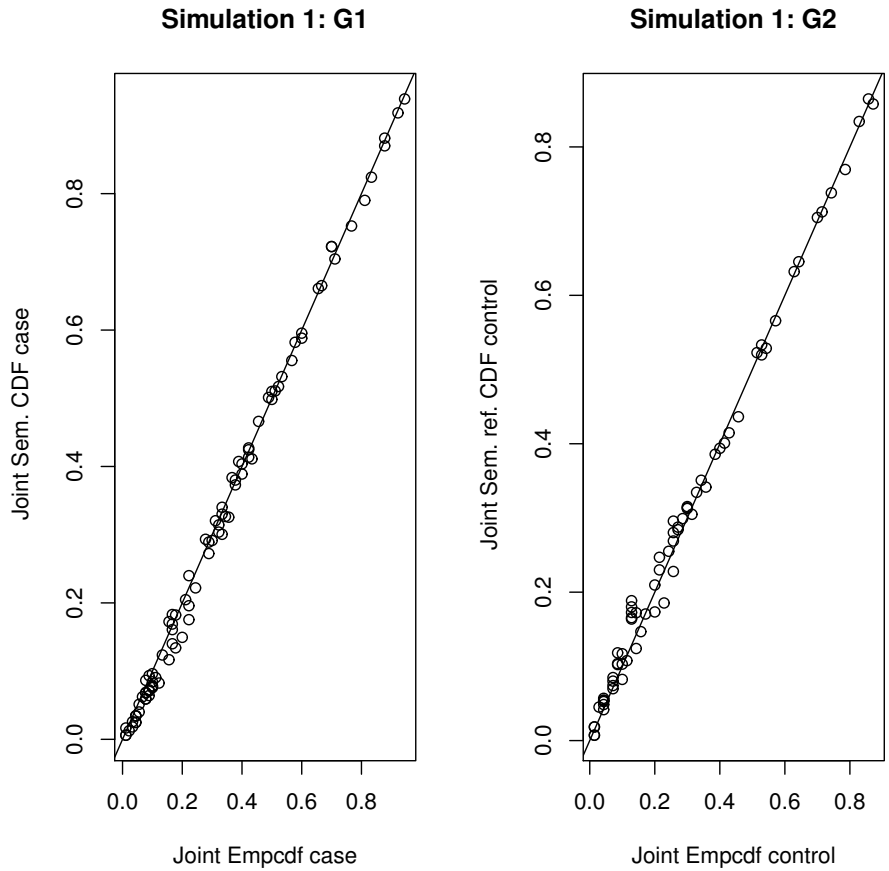
Correct model: $X_0 \sim \mathrm{N}(0,1), \mathrm{X}_1 \sim \mathrm{N}(1,1)$
Incorrect Model:$X_0 \sim \mathrm{N}(0,1), \mathrm{X}_1 \sim \mathrm{Exp}(1)$
In all cases $n_0 = n_1 = 100$.



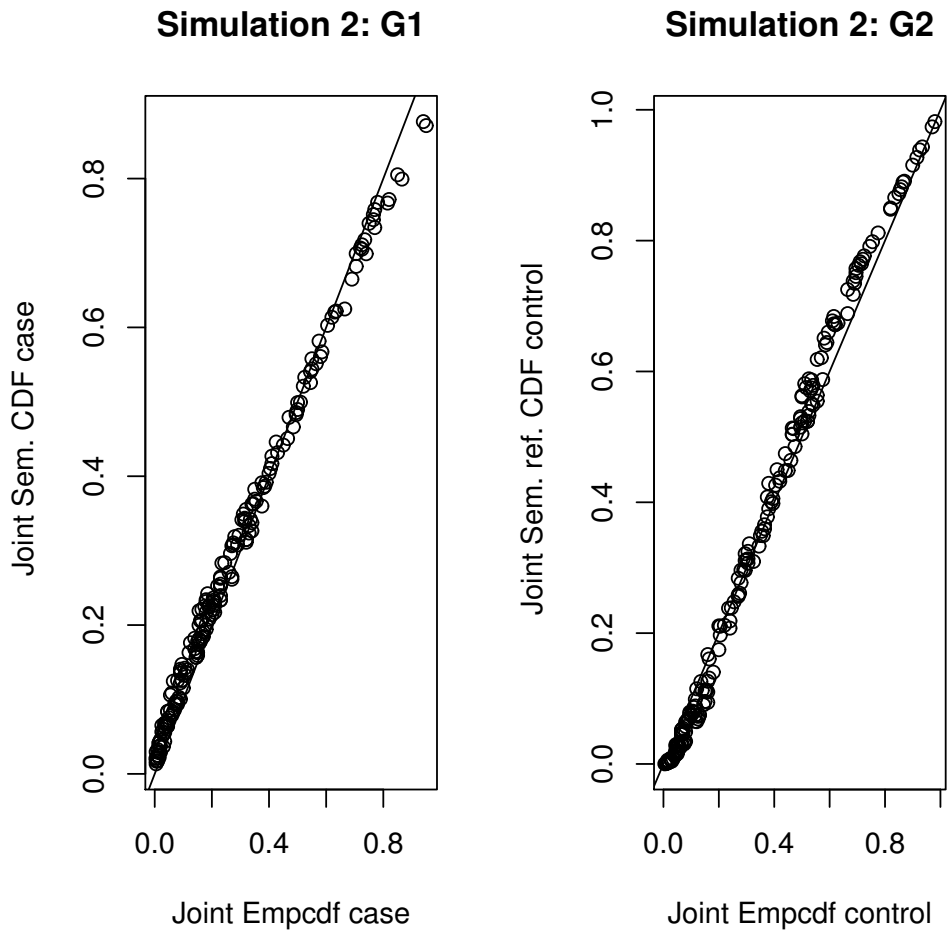


Simulation Results (VKG 2012): $\hat{G}$ vs. $\tilde{G}$

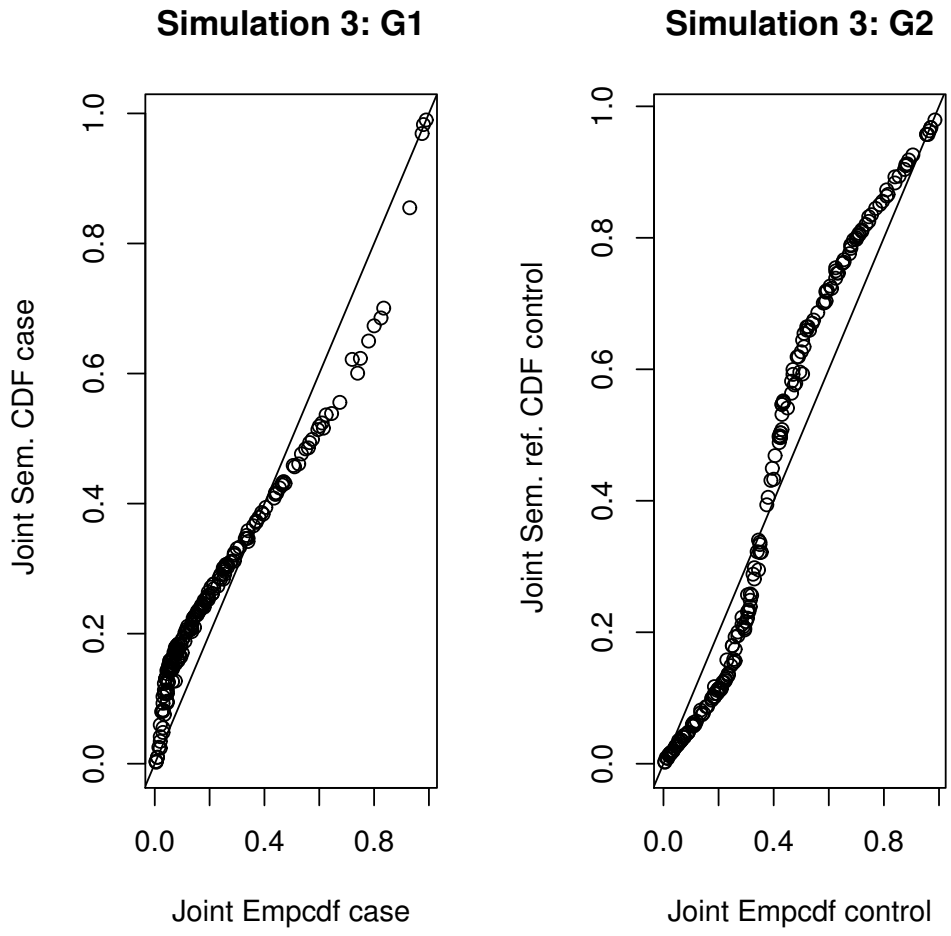
Simulation 1: $g_1 \sim N((0,0)',))$ (case), $g_2 \sim N((0,0)',))$ (control), $= \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$, $n_1 = 40$, $n_2 = 30$.



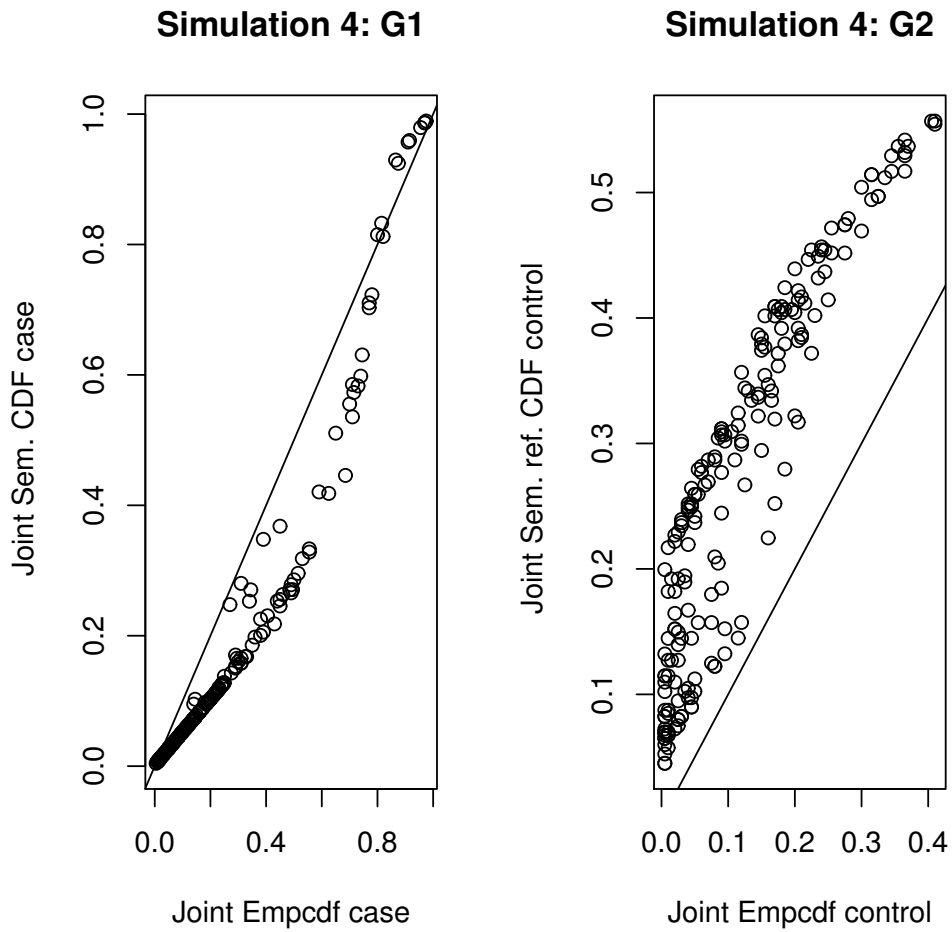
Simulation 2: $g_1 \sim N((0,0)',))$ (case), $g_2 \sim N((1,1)',))$ (control) with $= \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$, $n_1 = 200$, $n_2 = 200$.



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



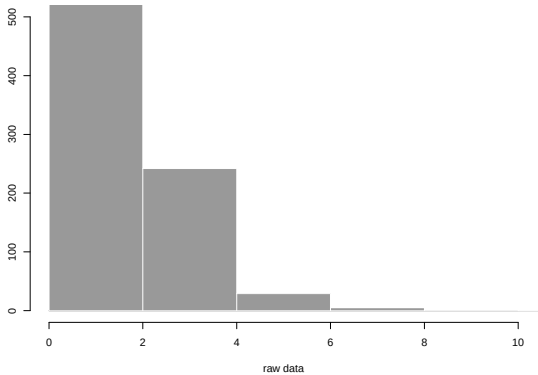
Simulation 4: g_1 from standard two dimensional Multivariate Cauchy (case) and g_2 from uniform distribution on the triangle $(0,0),(6,0),(-3,4)$ (control), and $n_1 = 200$, $n_2 = 200$.

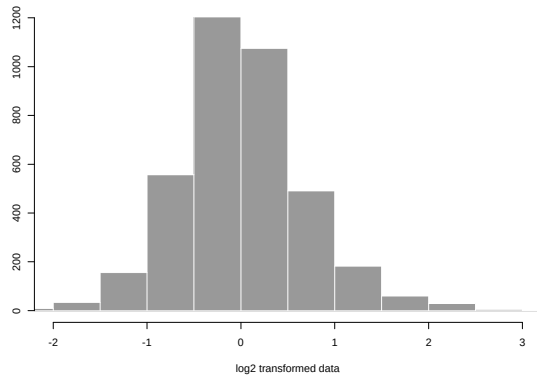


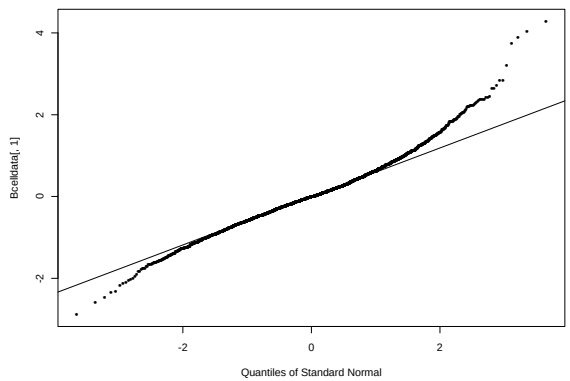
Outline

Application to microarrays (Qi 2002)

Histograms of gene expression GCAC4026 raw and $\log_2$-data.
QQ-plot of the $\log_2$-data shows the data are far from normal.



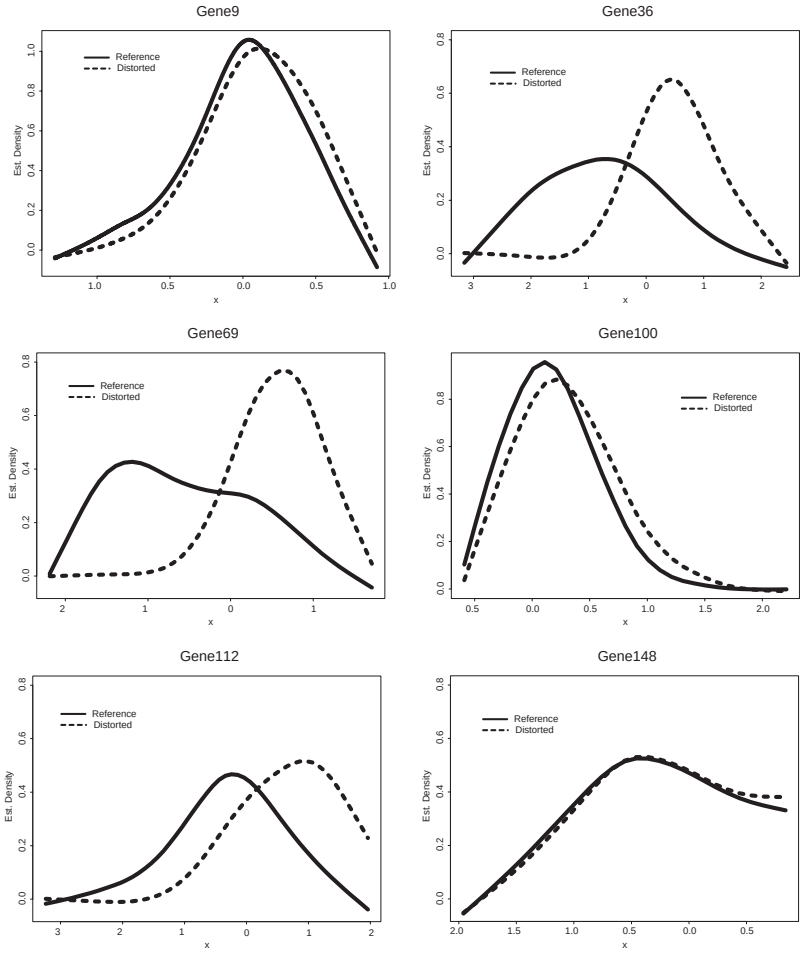




Two groups, GC, AC (reference), of $\log_2$ gene expression data. $m = 2, h(x) = x$. We get practically the same results with $\boldsymbol{h}(x) = (x, x^2)'$. Plots of the estimated pdf's support the numerical results. Use $\mathcal{X}_1$, df=1, in testing similarity.

Gene	n_1	n_2	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\mathcal{X}_1$	p-value
9	24	23	-0.049(0.06)	0.810(0.74)	1.46	0.226
36	24	21	0.452(0.35)	3.199(1.02)	97.17	0.000
69	24	22	-0.073(0.30)	2.415(0.71)	47.35	0.000
100	23	22	-0.159(0.13)	0.751(0.63)	1.24	0.265
112	24	23	-0.241(0.21)	1.698(0.53)	31.86	0.000
148	23	23	0.030(0.13)	0.100(0.44)	0.05	0.821

Estimated pdf's of Gene Expression



Application to radar meteorology (KWF 2004)

Reflectivity data obtained from two different radars (or "algorithms" or "sensors") at two different time periods. Data are random samples of reflectivity.

Kwajalein radar: S-band (10 cm) KPOL radar, located on Kwajalein Island at the southern end of the Kwajalein Atoll, Marshall Islands.

Brown Radar: C-band radar aboard NOAA ship Ronald H. Brown (RHB) at sea near Kwajalein Island.

The data obtained during the first period are referred to suggestively as **Kwajalein1, Brown1**, and those from the second period are called **Kwajalein2, Brown2**.

Kwajalein1, Brown1 (Reference)

$m = 2$, $n_1 = n_2 = 500$. The hypothesis that the data come from the same radar (algorithm) is **rejected** quite conclusively.

$h(x)$	Data	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\mathcal{X}_1$	p-value
x	1	0.784	-0.027	14.503	1.399e-03
	2	1.244	-0.042	33.476	7.216e-09
	3	0.707	-0.024	12.204	4.768e-04
	4	0.909	-0.030	17.292	3.206e-05
$\log(x)$	1	1.319	-0.396	6.520	0.011
	2	1.908	-0.575	12.744	3.572e-04
	3	1.871	-0.562	11.202	8.169e-04
	4	2.050	-0.621	16.510	4.838e-05

Brown1, Brown1 (reference)

$m = 2$, $n_1 = n_2 = 500$. The hypothesis that the data come from the same radar (algorithm) is **accepted** quite conclusively.

$h(x)$	Data	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\mathcal{X}_1$	p-value
x	1	-0.078	0.003	0.140	0.709
	2	0.005	-0.000	0.001	0.980
	3	-0.139	0.005	0.457	0.499
	4	0.112	-0.004	0.274	0.601
$\log(x)$	1	-0.584	0.175	1.723	0.189
	2	0.095	-0.028	0.042	0.838
	3	-0.225	0.067	0.250	0.617
	4	-0.027	0.008	0.003	0.959

Kwajalein2, Brown2 (reference)

$m = 2$, $n_1 = n_2 = 500$. The hypothesis that the data come from the same radar (algorithm) is **rejected** quite conclusively.

$h(x)$	Data	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\mathcal{X}_1$	p-value
x	1	5.323	-0.164	88.332	0
	2	3.975	-0.123	52.279	4.815e-13
	3	4.695	-0.146	74.950	0
	4	5.016	-0.156	85.325	0
$\log(x)$	1	14.359	-4.142	54.526	1.534e-13
	2	18.625	-5.367	79.723	0
	3	14.880	-4.302	60.788	6.328e-15
	4	13.580	-3.921	49.771	1.727e-12

Kwajalein2, Kwajalein2, Kwajalein2 (reference)

$m = 3$, $n_1 = n_2 = n_3 = 500$. The hypothesis that the data come from the same radar (algorithm) is **accepted** quite conclusively.

Data	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\mathcal{X}_1$	p-value
$h(x) = x$						
1	0.108	0.049	-0.003	-0.002	0.283	0.868
2	0.065	-0.003	-0.002	0.000	0.135	0.935
3	0.227	-0.041	-0.007	0.001	1.896	0.388
4	0.239	-0.220	-0.008	0.007	4.707	0.095
$h(x) = \log x$						
1	0.453	2.278	-0.132	-0.665	1.929	0.381
2	-0.792	-0.223	0.231	0.065	0.250	0.882
3	-0.359	0.735	0.105	-0.215	0.553	0.758
4	1.665	1.246	-0.485	-0.363	1.014	0.602

Estimation of small threshold probabilities

- From Theorem 1 we know that $\sqrt{n}(\hat{G}(t) - G(t))$ converges to a zero-mean Gaussian process.
- Let $\hat{V}(t)$ denote the estimated variance of $\hat{G}(t)$ obtained from the theorem by replacing parameters by their estimates.
- A $1 - \alpha$ level pointwise confidence interval for $G(t)$ is approximated by

$$\left(\hat{G}(t) - z_{\alpha/2}\sqrt{\hat{V}(t)}, \;\; \hat{G}(t) + z_{\alpha/2}\sqrt{\hat{V}(t)} \right), \tag{10}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution.

Idea:

If only the reference distribution is of interest then there is no reason why the fusion samples cannot be computer generated samples.

In particular, we can obtain confidence intervals for small threhold or survival probabilities $1 - G(T)$ for relatively large thresholds T by fusing a given reference sample X_0 with external artificially generated independent data. We refer to this as *out of sample fusion.*

Advantages:

• Control the generated samples to ensure the density ratio model holds to a reasonable degree.

• ***And if the density ratio assumption holds, we can populate the tails of the reference distribution with data related to*** X_0***.***

• Control the number of artificial samples and their sizes.

• Can repeat the integration of the real and artificial data many times to check the reliability of the results.

• Since more data (albeit artificial) are used in addition to X_0, the method produces on average short confidence intervals for small threshold probabilities relative to existing methods.

$100(1-\alpha)\%$ CI's for Binomial Probabilities

n Bernoulli trials. Success probability p. Number of successes X. $\hat{p} = X/n$.

- "Text book" (Empirical or "EP")

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Agresti-Coull (1998) ("AC")

Recommended in Brown et al (2001) for relatively large samples.

Replace n by $\tilde{n} = n + z^2_{\alpha/2}$

Replace $\hat{p}$ by

$$\frac{X + z^2_{\alpha/2}/2}{\tilde{n}}.$$

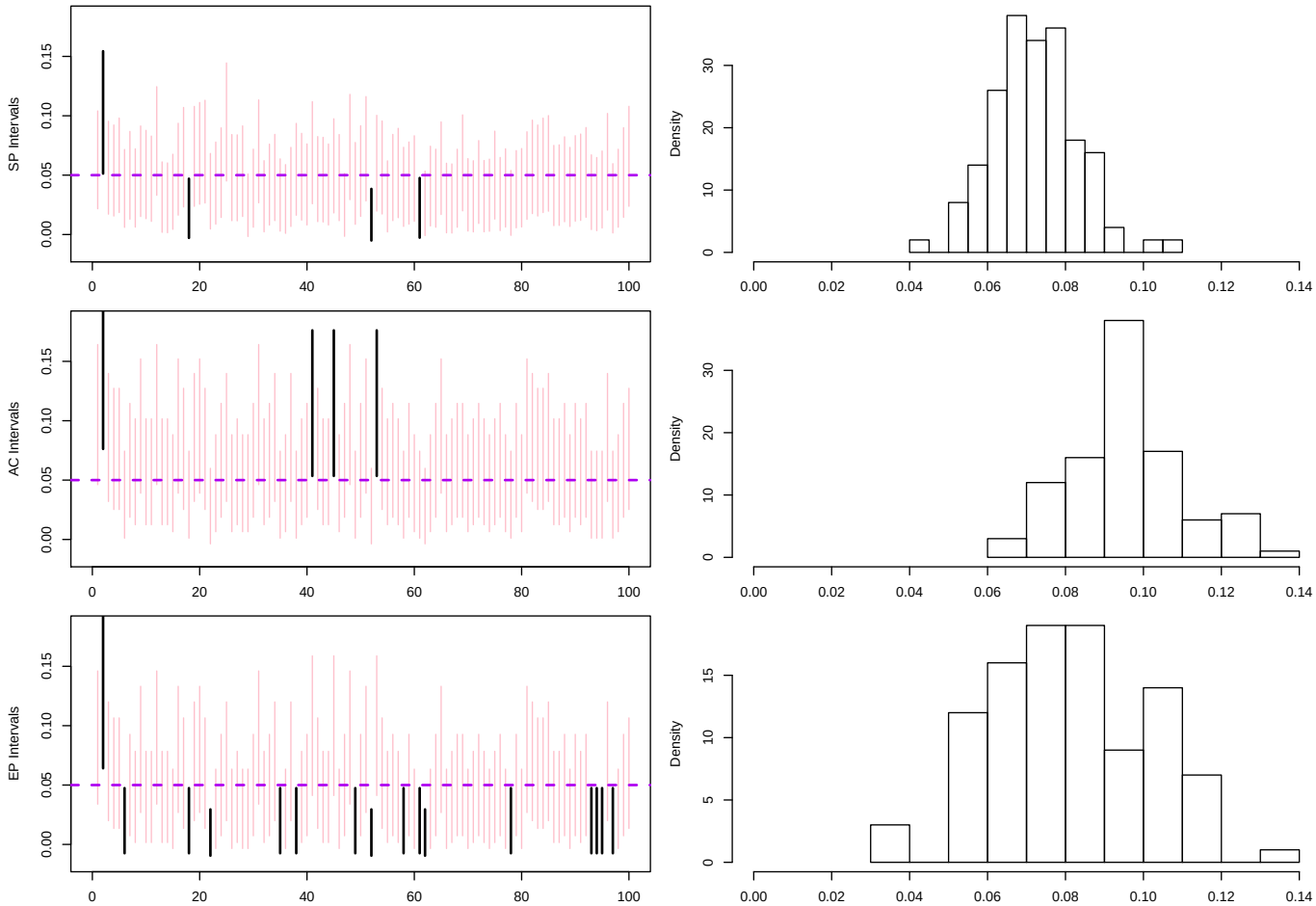
- Wilson (1927) ("WL")

$$\frac{\hat{p} + \frac{1}{2n}z^2_{\alpha/2} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + \frac{1}{n}z^2_{\alpha/2}}$$

Coverage from 100 CI's

$h = (x, x^2)$, $p = 1 - G(1.645) = 0.05$. All $n_i = 100$.

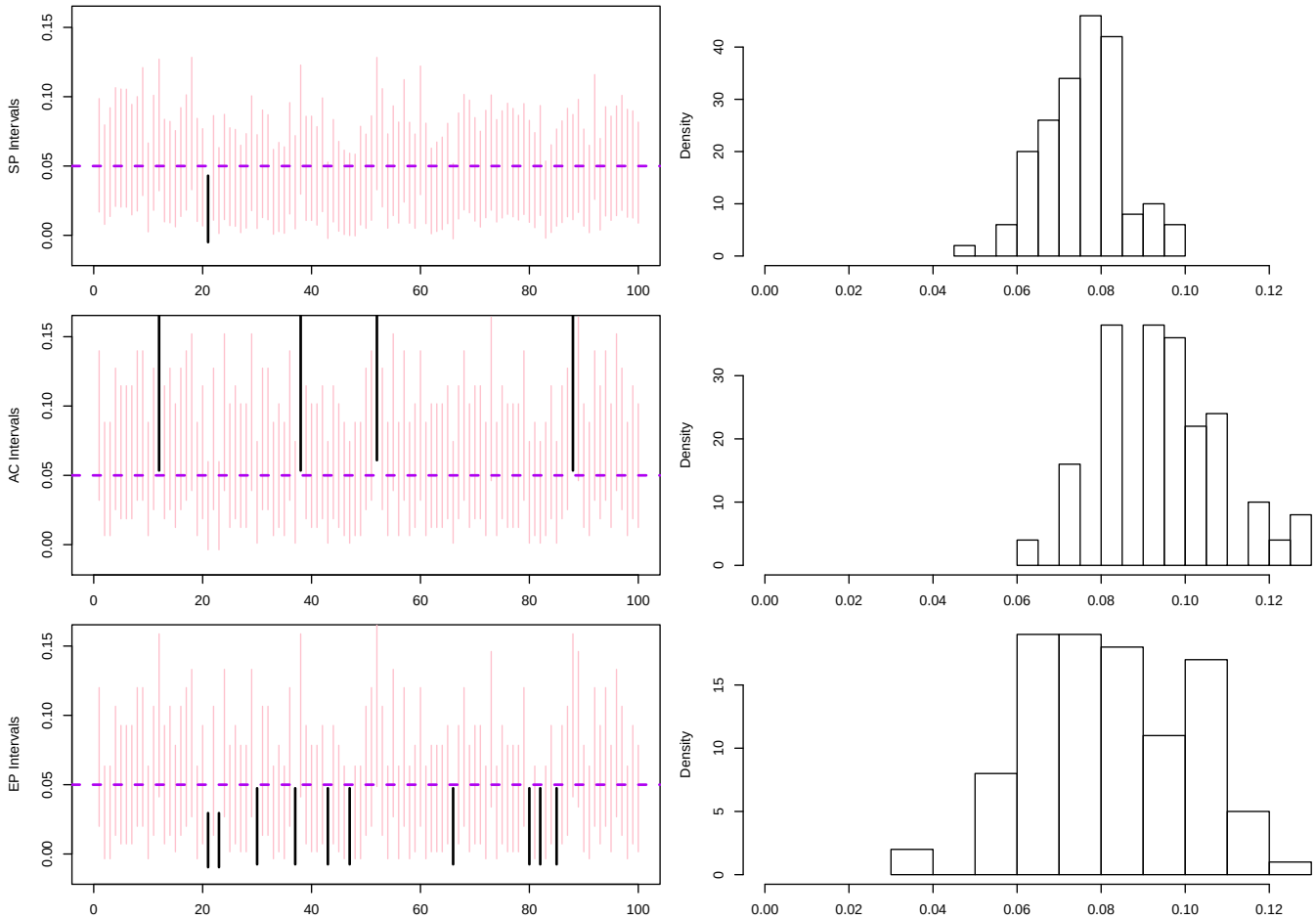
$X_0 \sim \mathrm{N}(0,1), \mathrm{X}_1 \sim \mathrm{Exp}(1), \mathrm{X}_2 \sim \mathrm{Poisson}(1), \mathrm{X}_3 \sim \mathrm{t}_{(5)}$



Coverage from 100 CI's

$h = (x, x^2)$, $p = 1 - G(1.645) = 0.05$. All $n_i = 100$.

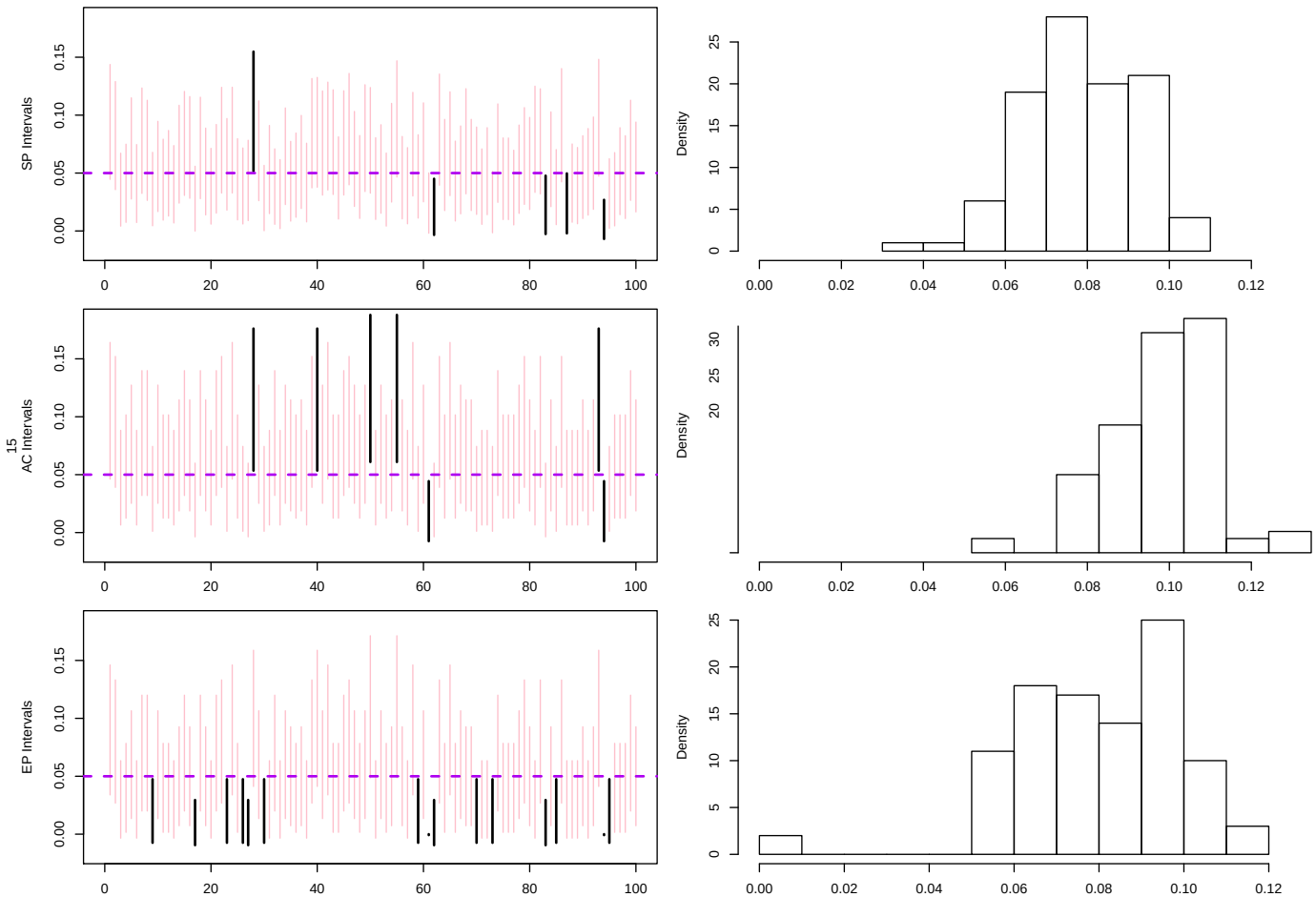
$X_0 \sim \mathrm{N}(0,1), \mathrm{X}_1 \sim \mathrm{Exp}(1), \mathrm{X}_2 \sim \mathrm{b}(5, 0.6), \mathrm{X}_3 \sim \mathrm{Poisson}(1), \mathrm{X}_4 \sim \mathrm{t}_{(5)}$



Coverage from 100 CI's

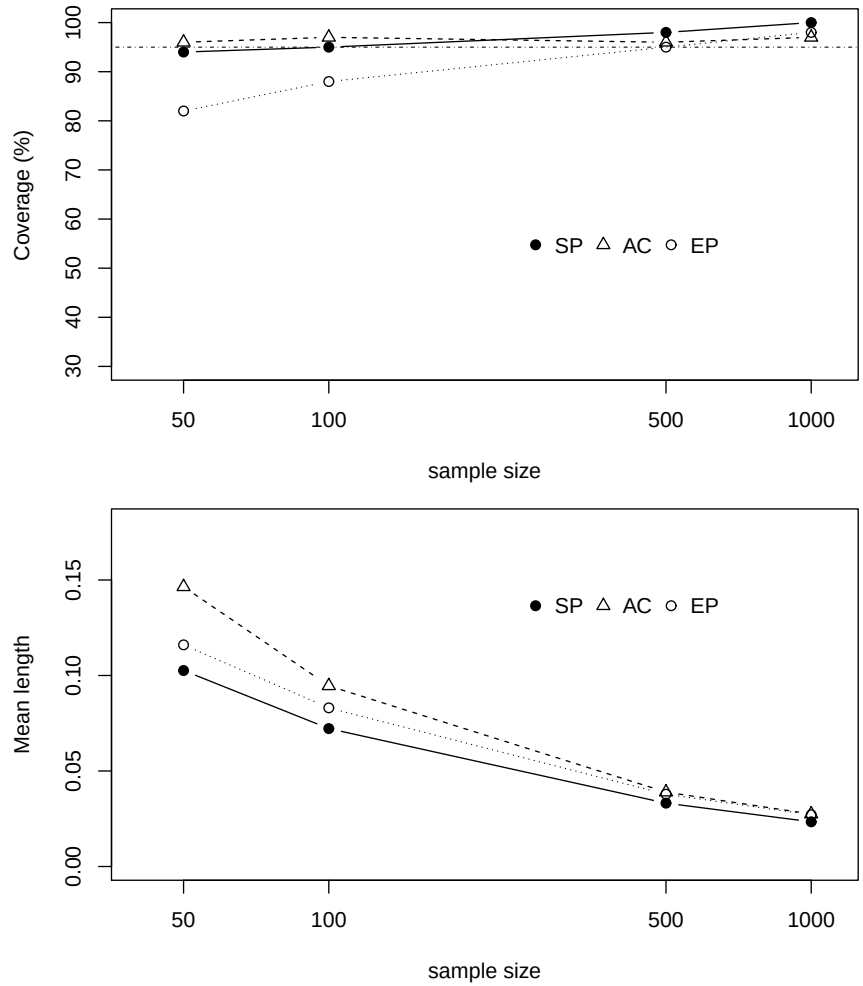
$h = (x, x^2)$, $p = 1 - G(1.645) = 0.05$. All $n_i = 100$.

$X_0 \sim \mathrm{N}(0, 1), \mathrm{X}_1 \sim \mathrm{Unif}(1, 7), \mathrm{X}_2 \sim \mathrm{Unif}(-1, 3), \mathrm{X}_3 \sim \mathrm{Unif}(-10, 20),$



Coverage (top) and length (bottom) corresponding to 95% confidence as functions of sample size n_0 on a log scale. $p = 1 - G(1.645) = 0.05$.

SP: Single fusion of $X_0 \sim \mathrm{N}(0,1), \mathrm{X}_1 \sim \mathrm{N}(0,1), \mathrm{X}_2 \sim \mathrm{N}(1,1), \mathrm{X}_3 \sim \mathrm{N}(0.5,1),$



$p = 0.0001$, $h = (x, x^2)$

```
X <- cbind(rnorm(100,-1,1),rnorm(100,-1,2),rnorm(100,2,3),rnorm(100))
SE(X,3.719016,"normal")
```

Method	$\hat{p}$	SE	L	U
SP	0.0001564856	0.001083197	-0.001966580	0.002279551
AC	0.0184974037	0.013222556	-0.007418807	0.044413614
WL	0.0184974037	0.009437451	0.000000000	0.036994807
EP	0.0000000000	0.000000000	0.000000000	0.000000000

Saving:

$$0.013222556/0.001083197 = 12.20697$$

Method	$\hat{p}$	SE	L	U
SP	0.0001231609	0.000955043	-0.001748723	0.001995045
AC	0.0184974037	0.013222556	-0.007418807	0.044413614
WL	0.0184974037	0.009437451	0.000000000	0.036994807
EP	0.0000000000	0.000000000	0.000000000	0.000000000

Saving:

$$0.013222556/0.000955043 = 13.84498$$

$$p = 0.0001, h = (x, x^2)$$

```
X <- cbind(runif(100,-3,3),runif(100,-10,10),runif(100,-20,20),rnorm(100))
SE(X,3.719016,"normal")
```

Method	$\hat{p}$	SE	L	U
SP	4.208654e-05	0.0005613172	-0.001058095	0.001142268
AC	1.849740e-02	0.0132225564	-0.007418807	0.044413614
WL	1.849740e-02	0.0094374509	0.000000000	0.036994807
EP	0.00000000	0.0000000000	0.000000000	0.000000000

Saving:

$$0.0132225564/0.0005613172 = 23.5563$$

Method	$\hat{p}$	SE	L	U
SP	0.0002167695	0.001275283	-0.002282784	0.002716323
AC	0.0184974037	0.013222556	-0.007418807	0.044413614
WL	0.0184974037	0.009437451	0.000000000	0.036994807
EP	0.0000000000	0.000000000	0.000000000	0.000000000

Saving:

$$0.013222556/0.001275283 = 10.36833$$

$$p = 0.0001, h = (x, x^2)$$

```
X <- cbind(rgamma(100,5,1),rgamma(100,5,2),rgamma(100,5,3),rnorm(100))
SE(X,3.719016,"normal")
```

Method	$\hat{p}$	SE	L	U
SP	0.0003753147	0.001676662	-0.002910944	0.003661573
AC	0.0184974037	0.013222556	-0.007418807	0.044413614
WL	0.0184974037	0.009437451	0.000000000	0.036994807
EP	0.0000000000	0.000000000	0.000000000	0.000000000

Saving:

$$0.013222556/0.001676662 = 7.886238$$

Method	$\hat{p}$	SE	L	U
SP	0.000512219	0.001959691	-0.003328775	0.004353213
AC	0.018497404	0.013222556	-0.007418807	0.044413614
WL	0.018497404	0.009437451	0.000000000	0.036994807
EP	0.000000000	0.000000000	0.000000000	0.000000000

Saving:

$$0.013222556/0.001959691 = 6.747266$$

$$p = 0.0001, h = (x, x^2)$$

```
X <- cbind(rpois(100,1),rt(100,5),rbinom(100,5,0.6),rnorm(100))
SE(X,3.719016,"normal")
```

Method	$\hat{p}$	SE	L	U
SP	0.002133187	0.003995519	-0.005698030	0.009964405
AC	0.018497404	0.013222556	-0.007418807	0.044413614
WL	0.018497404	0.009437451	0.000000000	0.036994807
EP	0.000000000	0.000000000	0.000000000	0.000000000

Saving:

$$0.013222556/0.003995519 = 3.309346$$

Method	$\hat{p}$	SE	L	U
SP	0.003771944	0.005308733	-0.006633172	0.01417706
AC	0.018497404	0.013222556	-0.007418807	0.04441361
WL	0.018497404	0.009437451	0.000000000	0.03699481
EP	0.000000000	0.000000000	0.000000000	0.00000000

Saving:

$$0.013222556/0.005308733 = 2.490718$$

$$p = 0.0001, h = (x, logx)$$

```
X <- cbind(runif(100,0,5),runif(100,0,10),runif(100,0,40),rgamma(100,5,3))
SE(X,5.927336,"gamma")
```

Method	$\hat{p}$	SE	L	U
SP	1.363207e-05	0.0003072173	-0.0005885139	0.000615778
AC	1.849740e-02	0.0132225564	-0.0074188068	0.044413614
WL	1.849740e-02	0.0094374509	0.0000000000	0.036994807
EP	0.000000e+00	0.0000000000	0.0000000000	0.000000000

Saving:

$$0.0132225564/0.0003072173 = 43.03975$$

Method	$\hat{p}$	SE	L	U
SP	9.977693e-06	0.0002700859	-0.0005193906	0.000539346
AC	1.849740e-02	0.0132225564	-0.0074188068	0.044413614
WL	1.849740e-02	0.0094374509	0.0000000000	0.036994807
EP	0.000000e+00	0.0000000000	0.0000000000	0.000000000

Saving:

$$0.0132225564/0.0002700859 = 48.95686$$

$$p = 0.0001, h = (x, logx)$$

```
X <- cbind(runif(100,0,50),runif(100,0,30),rlnorm(100,1,1))
SE(X,41.22383,"gamma")
```

Method	$\hat{p}$	SE	L	U
SP	0.0001429428	0.0009762795	-0.001770565	0.002056451
AC	0.0184974037	0.0132225564	-0.007418807	0.044413614
WL	0.0184974037	0.0094374509	0.000000000	0.036994807
EP	0.0000000000	0.0000000000	0.000000000	0.000000000

Saving:

$$0.0132225564/0.0009762795 = 13.54382$$

Method	$\hat{p}$	SE	L	U
SP	0.0001009011	0.0008193674	-0.001505059	0.001706861
AC	0.0184974037	0.0132225564	-0.007418807	0.044413614
WL	0.0184974037	0.0094374509	0.000000000	0.036994807
EP	0.0000000000	0.0000000000	0.000000000	0.00000000

Saving:

$$0.0132225564/0.0008193674 = 16.13752$$

$$p = 0.0001, h = (x, x^2)$$

```
X <- cbind(rgamma(500,5,3),runif(500,-3,5),rnorm(500))
SE(X,3.719016,"normal")
```

Method	$\hat{p}$	SE	L	U
SP	0.0001432455	0.000437294	-7.138508e-04	0.001000342
AC	0.0038123093	0.002745476	-1.568824e-03	0.009193442
WL	0.0038123093	0.001945056	4.336809e-19	0.007624619
EP	0.0000000000	0.000000000	0.000000e+00	0.000000000

Saving:

$$0.002745476/0.000437294 = 6.27833$$

Method	$\hat{p}$	SE	L	U
SP	0.0001046779	0.0003731569	-6.267097e-04	0.0008360654
AC	0.0038123093	0.0027454759	-1.568824e-03	0.0091934420
WL	0.0038123093	0.0019450557	4.336809e-19	0.0076246185
EP	0.0000000000	0.0000000000	0.000000e+00	0.0000000000

Saving:

$$0.0027454759/0.0003731569 = 7.35743$$

$$p = 0.0001, h = (x, x^2)$$

```
X <- cbind(rgamma(1000,5,3),runif(1000,-3,5),rnorm(1000))
SE(X,3.719016,"normal")
```

Method	$\hat{p}$	SE	L	U
SP	0.0001204543	0.0002833001	-4.348139e-04	0.0006757226
AC	0.0019134493	0.0013793040	-7.899865e-04	0.0046168851
WL	0.0019134493	0.0009762496	2.168404e-19	0.0038268986
EP	0.0000000000	0.0000000000	0.000000e+00	0.0000000000

Saving:

$$0.0013793040/0.0002833001 = 4.868703$$

Method	$\hat{p}$	SE	L	U
SP	0.0001017656	0.0002604297	-4.086765e-04	0.0006122078
AC	0.0019134493	0.0013793040	-7.899865e-04	0.0046168851
WL	0.0019134493	0.0009762496	2.168404e-19	0.0038268986
EP	0.0000000000	0.0000000000	0.000000e+00	0.0000000000

Saving:

$$0.0013793040/0.0002604297 = 5.296262$$

$$p = 0.0001, h = (x, x^2)$$

```
X <- cbind(rgamma(5000,5,3),runif(5000,-3,5),rnorm(5000))
SE(X,3.719016,"normal")
```

Method	$\hat{p}$	SE	L	U
SP	0.0001128325	0.0001226638	-1.275885e-04	0.0003532536
AC	0.0005837115	0.0003414449	-8.552039e-05	0.0012529434
WL	0.0005837115	0.0002797993	3.530488e-05	0.0011321182
EP	0.0002000000	0.0001999800	-1.919608e-04	0.0005919608

Saving:

$$0.0003414449/0.0001226638 = 2.783583$$

Method	$\hat{p}$	SE	L	U
SP	0.0001076335	0.0001197527	-0.0001270818	0.0003423488
AC	0.0003838651	0.0002769199	-0.0001588979	0.0009266281
WL	0.0003838651	0.0001958495	0.0000000000	0.0007677301
EP	0.0000000000	0.0000000000	0.0000000000	0.0000000000

Saving:

$$0.0002769199/0.0001197527 = 2.312431$$

Testicular Germ Cell Tumor (TGCT) Data

- TGCT data from the Servicemen's Testicular Tumor Environmental and Endocrine Determinants Study (2002-2005) contain 763 case and 928 control observations.
- Out of sample fusion: Each of the TGCT case-control samples is combined separately with the **same** computer generated artificial sample to detect differences in the case-control distributions.
- "Choosing different artificial samples lets us view the reference sample from different angles".
- We validated the results of KKVG (2009) by finding a difference between the case and control distributions.

The TGCT Case-Control data are very similar in terms of summary statistics.

Table: TGCT case-control summary statistics.

Variables	CCTL	Range	Mean	SD
Age	Control	18.00 ,46.00	27.91	5.93
	Case	18.00 ,45.00	27.82	5.99
Height (cm)	Control	152.4, 215.9	178.3	7.06
	Case	160.0, 203.2	179.6	7.03
Weight (kg)	Control	38.55, 127.01	80.13	11.14
	Case	50.80, 131.54	81.43	11.69

$$\rho_{\text{Control}} = \begin{array}{c} \\ \text{A} \\ \text{H} \\ \text{W} \end{array} \begin{array}{c} \begin{array}{ccc} \text{A} & \text{H} & \text{W} \end{array} \\ \begin{pmatrix} 1.000 & -0.021 & 0.115 \\ -0.021 & 1.000 & 0.505 \\ 0.115 & 0.505 & 1.000 \end{pmatrix} \end{array}$$

$$\rho_{\text{Case}} = \begin{array}{c} \\ \text{A} \\ \text{H} \\ \text{W} \end{array} \begin{array}{c} \begin{array}{ccc} \text{A} & \text{H} & \text{W} \end{array} \\ \begin{pmatrix} 1.000 & 0.021 & 0.162 \\ 0.021 & 1.000 & 0.521 \\ 0.162 & 0.521 & 1.000 \end{pmatrix} \end{array}$$

Consider multivariate normal distributions with the same covariance matrices:

$$\mathbf{x}_0 \sim g_0(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}), \quad \mathbf{x}_1 \sim g_1(\mathbf{x}) = \mathcal{N}(\boldsymbol{mu}_1, \boldsymbol{\Sigma}).$$

$$\frac{g_1(\mathbf{x})}{g_0(\mathbf{x})} = \exp\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1)\}. \tag{11}$$

Denote

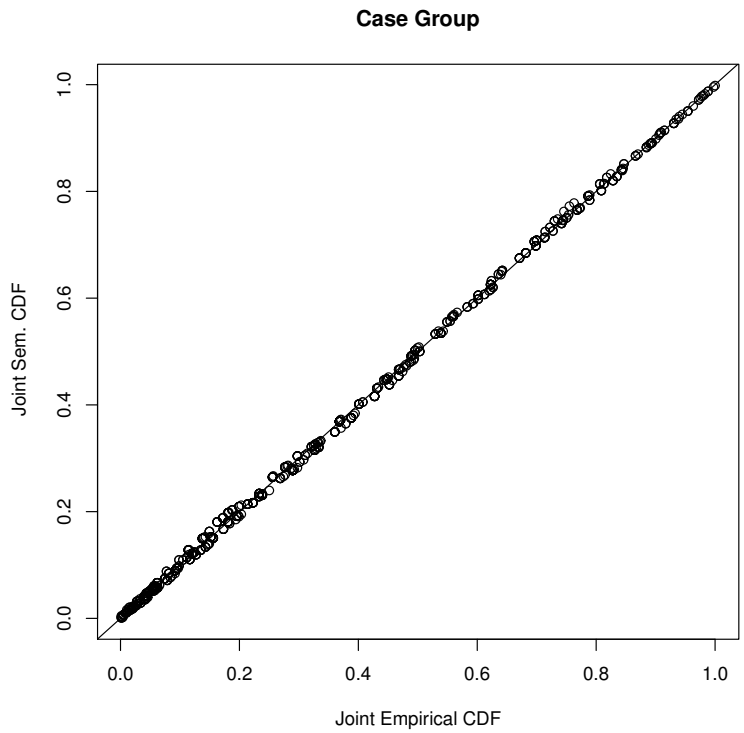
$$\alpha = -\frac{1}{2}(\boldsymbol{\mu}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1), \quad \boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0),$$

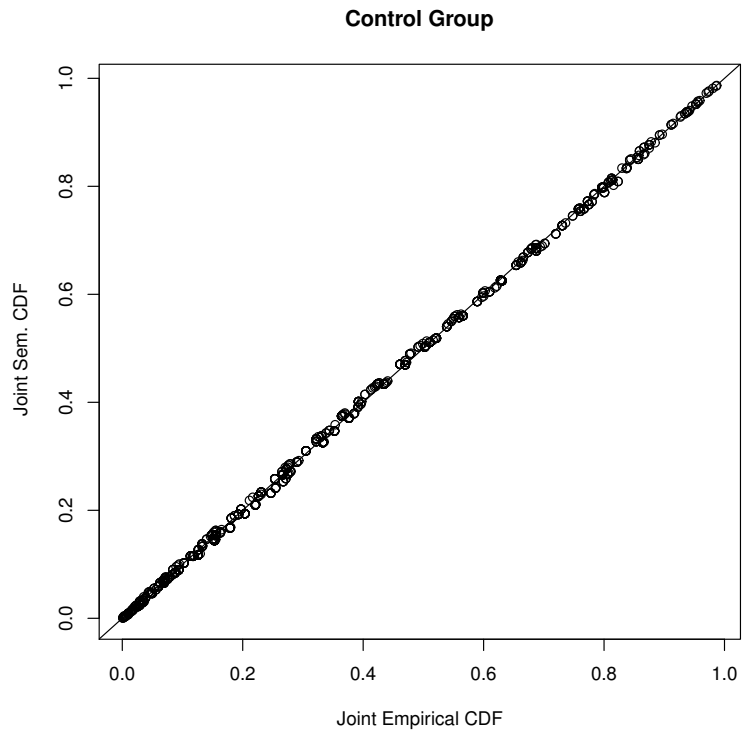
to obtain a density ratio model with tilt function: $\mathbf{h}(\mathbf{x}) = \mathbf{x}$.

$$\frac{g_1(\mathbf{x})}{g_0(\mathbf{x})} = \exp(\alpha + \boldsymbol{\beta}'\mathbf{x}), \tag{12}$$

By combining the real case-control data VKG (2012) show that (12) is justified by plotting $\hat{G}$ vs $\tilde{G}$ for both case and control.

Diagnostic plots of $\hat{G}_i$ versus $\tilde{G}_i$, $i = 1, 2$ evaluated at (height,weight) pairs (VKG 2012).





We fuse the case and control data (separately) with the same artificial normal sample with average mean and covariance matrix from the case-control data to obtain two bivariate distributions (Zhou 2013):

$$P(\textbf{Height} \leq \textit{Height}, \textbf{Weight} \leq \textit{Weight})$$

Height	Weight	Control	Case
170.	60.	0.3165	0.3031
175.	60.	0.4033	0.3747
180.	60.	0.5244	0.4987
185.	60.	0.6335	0.6149
170.	70.	0.3165	0.3031
175.	70.	0.4033	0.3747
180.	70.	0.5244	0.4987
185.	70.	0.6335	0.6149
170.	80.	0.3165	0.3031
175.	80.	0.4033	0.3747
180.	80.	0.5244	0.4987
185.	80.	0.6335	0.6149
170.	90.	0.3165	0.3031
175.	90.	0.4033	0.3747
180.	90.	0.5244	0.4987
185.	90.	0.6335	0.6149

Indeed, in KKVG (2009) we rejected $\boldsymbol{\beta} = \mathbf{0}$ in:

$$\frac{g_1(\mathbf{x})}{g_0(\mathbf{x})} = \exp(\alpha + \boldsymbol{\beta}'\mathbf{x})$$

Case-control kernel density estimates require optimal bandwidth matrices (VKG 2012):

$$\boldsymbol{H}_{\text{Control}} = \begin{array}{c} \\ \text{Height} \\ \text{Weight} \end{array} \begin{array}{c} \begin{array}{cc} \text{Height} & \text{Weight} \end{array} \\ \begin{pmatrix} 1.1010 & 0 \\ 0 & 1.7566 \end{pmatrix} \end{array} \qquad \boldsymbol{H}_{\text{Case}} = \begin{array}{c} \\ \text{Height} \\ \text{Weight} \end{array} \begin{array}{c} \begin{array}{cc} \text{Height} & \text{Weight} \end{array} \\ \begin{pmatrix} 1.1236 & 0 \\ 0 & 2.1079 \end{pmatrix} \end{array}$$

Case-control kernel density estimates require optimal bandwidth matrices (VKG 2012):

$$\boldsymbol{H}_{\text{Control}} = \begin{array}{c} \\ \text{Height} \\ \text{Weight} \end{array} \begin{array}{c} \begin{array}{cc} \text{Height} & \text{Weight} \end{array} \\ \begin{pmatrix} 1.1010 & 0 \\ 0 & 1.7566 \end{pmatrix} \end{array}$$

$$\boldsymbol{H}_{\text{Case}} = \begin{array}{c} \\ \text{Height} \\ \text{Weight} \end{array} \begin{array}{c} \begin{array}{cc} \text{Height} & \text{Weight} \end{array} \\ \begin{pmatrix} 1.1236 & 0 \\ 0 & 2.1079 \end{pmatrix} \end{array}$$

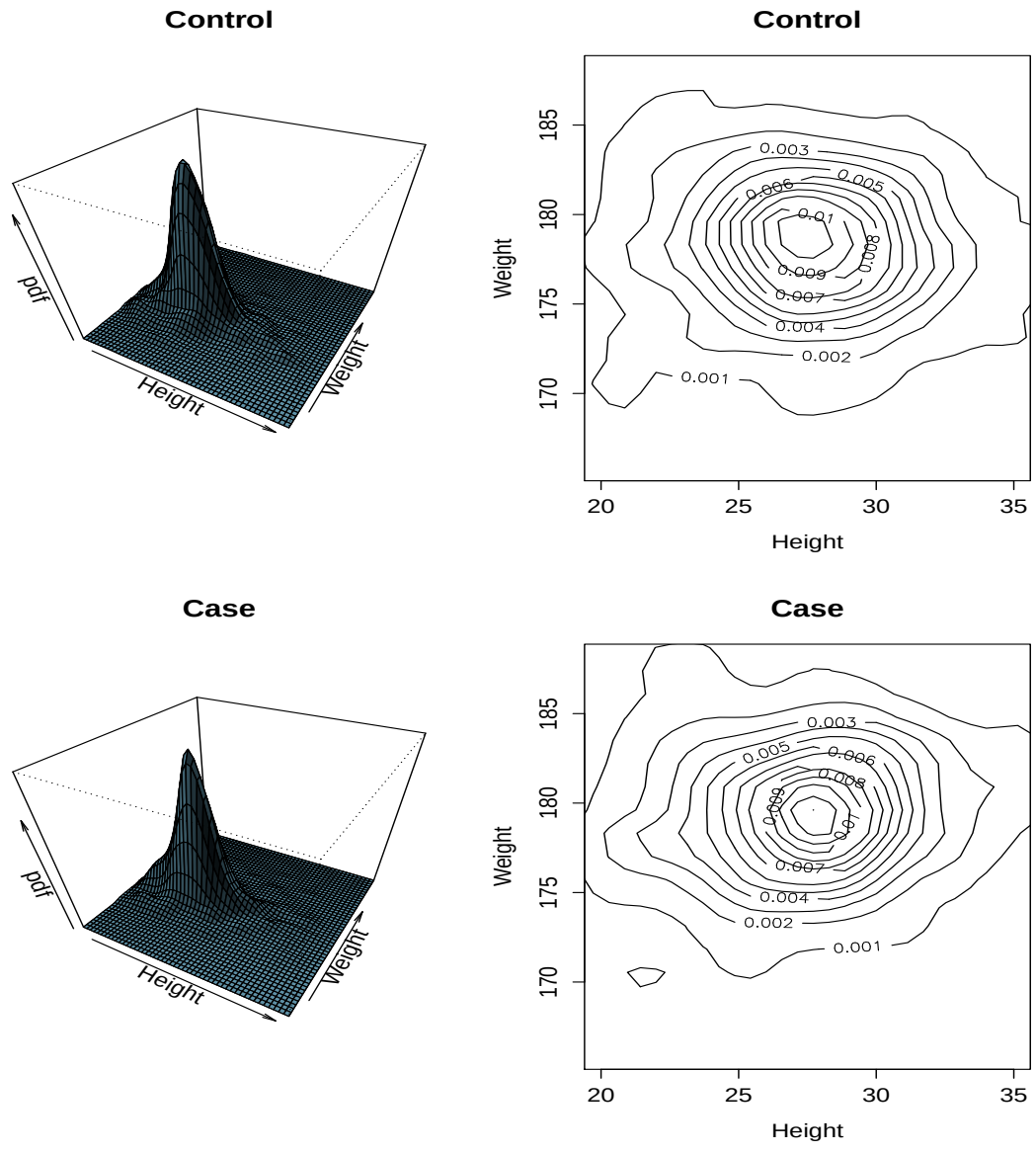
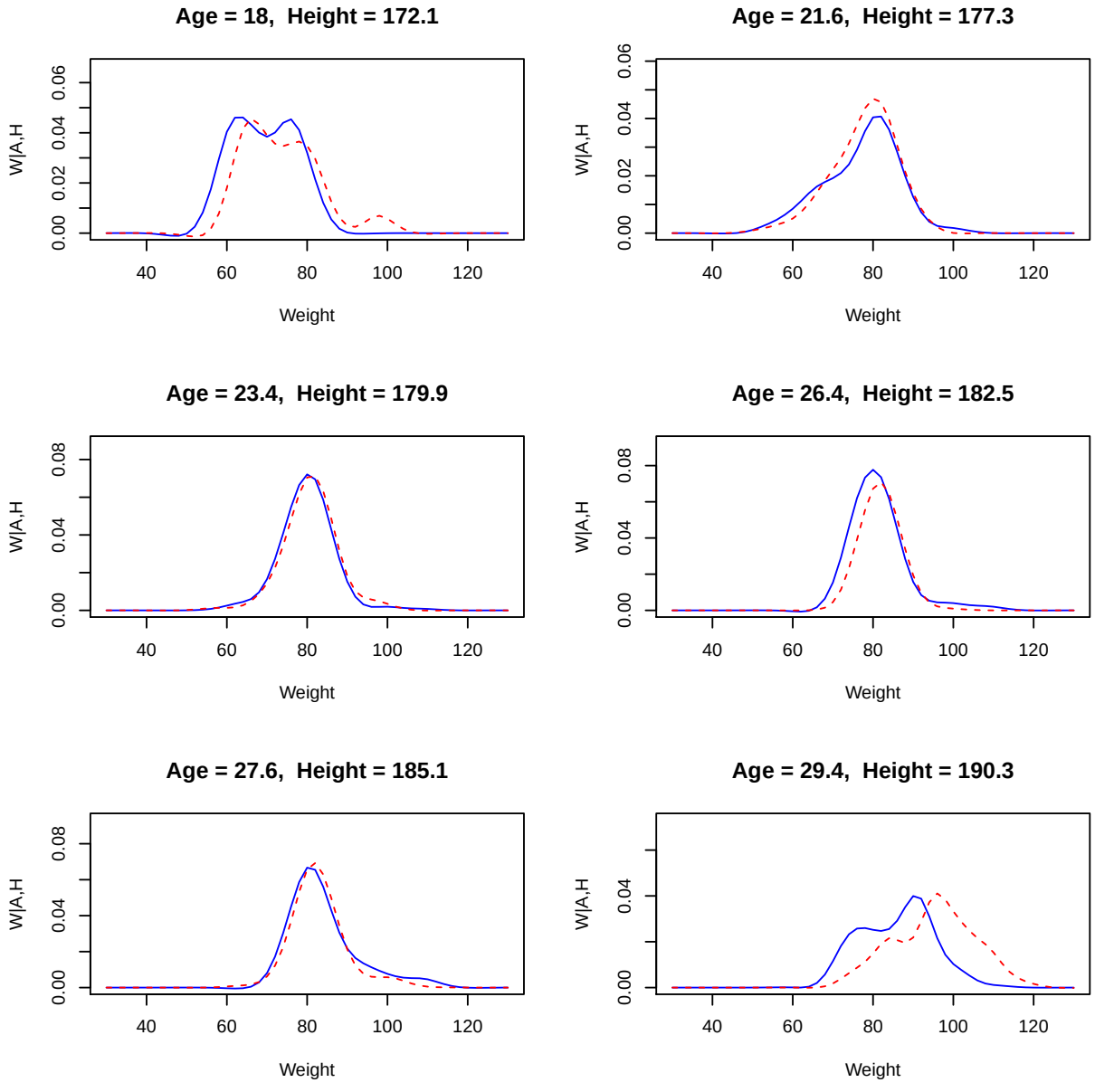
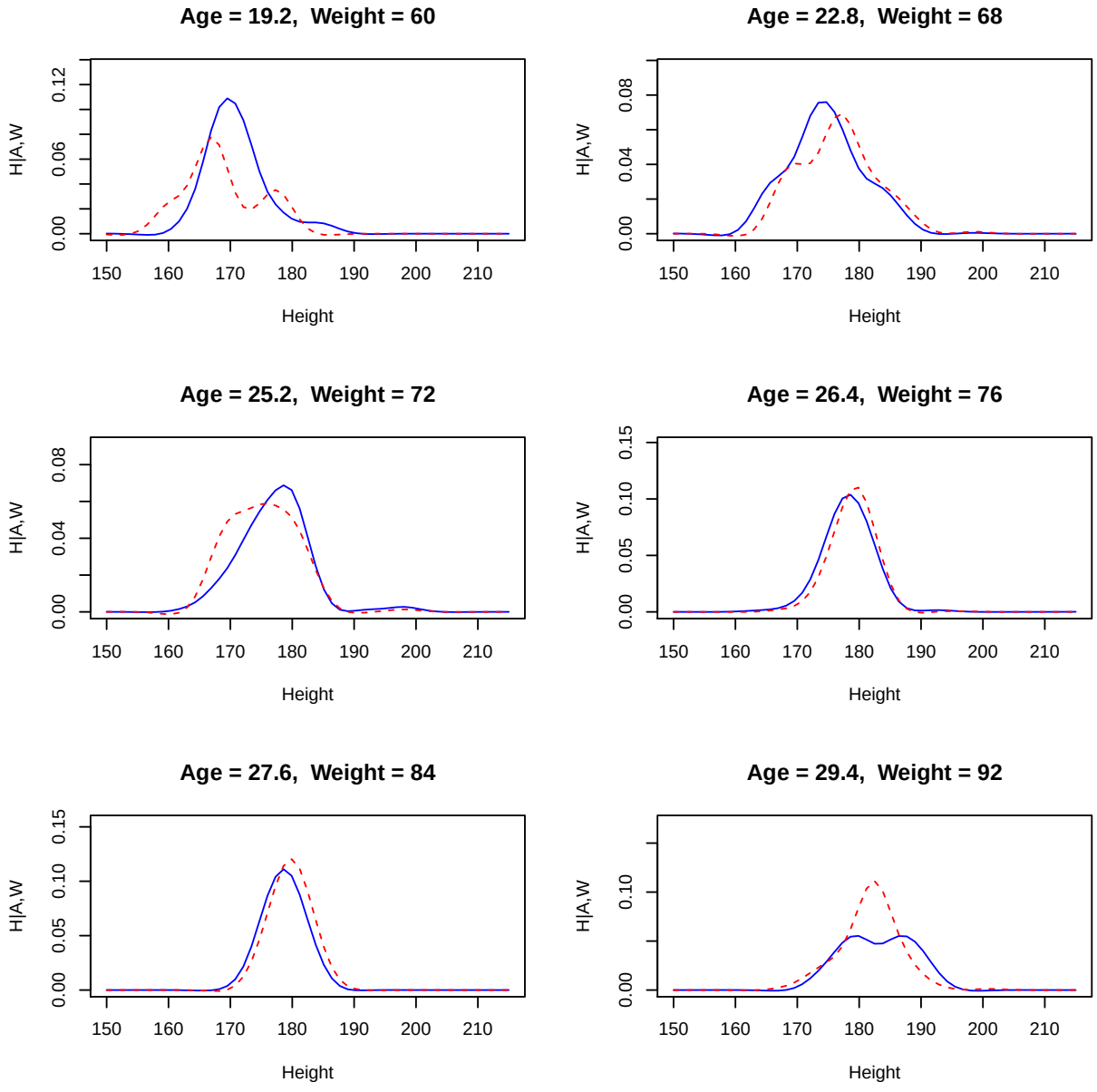


Figure: 3D plot for kernel density estimates of Height and Weight

$f(W|A,H)$. Control: solid line; Case: dashed line.



$f(H|A, W)$. Control: solid line; Case: dashed line.



Results obtained from different artificial data samples are strikingly similar.