Locally robust methods and near-parametric asymptotics

Spiridon Penev

UNSW Sydney, Australia

4th December 2017

Joint work with Kanta Naito (Shimane University, Japan)

Outline

Outline

- Past papers$\rightarrow$infusing a little localisation in the likelihood-based methods for regression and for density estimation can improve the resulting estimators w.r. to some **global** risk measures.
- We: similar effect can also be observed with respect to robust estimation procedures. Localised versions of robust density estimation procedures perform better with respect to a global risk measures based on minimization of Bregman divergence measures.

- Past papers$\rightarrow$infusing a little localisation in the likelihood-based methods for regression and for density estimation can improve the resulting estimators w.r. to some **global** risk measures.
- We: similar effect can also be observed with respect to robust estimation procedures. Localised versions of robust density estimation procedures perform better with respect to a global risk measures based on minimization of Bregman divergence measures.

Outline

Local likelihood estimation: introduced in Tibshirani and Hastie (1987) (regression) and by Loader (1996) and Hjort & Jones (1996) (density).
Represents a semiparametric approach: “The estimators run the gamut from a fully parametric fit to almost fully nonparametric with only a single smoothing parameter to be chosen".
Later: Eguchi & Copas (1998) and Park et. al. (2006): perform more detailed bandwidth analysis.
In a nutshell: the run is controlled by h, the bandwidth. “Small h": closer to the fully non-parametric; “Large h": closer to parametric. Depending on how close the “true" density to the parametric model$\rightarrow$ different terms in the expansion of estimation risk may dominate.

Local procedure: at each value x of the argument→parameter estimation.
Notation: cdf $F(x)$, density $f(x)$: nominal parametric model $g(x,\theta), \theta \in \Theta \in \mathbb{R}^p$. But: more realistic: assume that f belongs to a tubular neighbourhood $\cup_{\theta\in\Theta}\{f : D(f, g_\theta \leq \varepsilon)\}$ with, e.g.,

$$D(f, g_\theta) = E_f \log\left\{\frac{f(X)}{g(X,\theta)}\right\}.$$

Infuse a local adaptation to the global likelihood: maximize

$$\sum_{i=1}^{n} K\left(\frac{x_i - t}{h}\right) \log\left\{g(x_i, \theta)\right\} \tag{1}$$

with $K(\frac{x-t}{h})$: kernel. Note: normalization is needed:

$$\hat{g}_h(t) = g(t, \hat{\theta}_{t,h}) / \int g(x, \hat{\theta}_{x,h}) dx. \tag{2}$$

If ideal nominal parametric model is not expected to hold: replace KL: robust density estimation: Windham (1995), Basu et. al. (1998). We consider similar measures defined via the Bregman divergence (BD) and parameterised by $\lambda \geq 0$ (governing the compromise). However, we deal with local versions of the robust divergence measures.
Bregman divergence by using convex function $U(.)$ locally:

$$d_U(x,y) = U(x) - U(y) - < gradU(y), x - y > .$$

If $x(.)$ and $y(.)$ are positive functions then for $t \in \mathbb{R}$ (or $t \in \mathbb{R}^d$) the above defines point-wise divergence between $x(t)$ and $y(t)$ (in which case $gradU(y) = \frac{d}{dt}y(t)$). Using this: get globally

$$\int d_U(x(t), y(t))v(t)dt. \tag{3}$$

If $U(s) := -\log(s)$ and $v(t) = y(t) \rightarrow$ get the KL divergence between densities x and y.

It is not always appropriate to use $y(t)$ for weighting & uniform weight of $v(t) = 1$: more appropriate. The class of BD is large (not all of them useful in applications)$\rightarrow$we focus on a particular class. Starting with the Box-Cox transformation

$$G_\lambda(x) = \begin{cases} \frac{1}{\lambda}(x^\lambda - 1), \lambda > 0 \\ \log x, \lambda = 0 \end{cases}$$

define $U_\lambda(x) = x(G_\lambda(x) - 1)$. Note that $U_\lambda(x) \rightarrow_{\lambda \rightarrow 0} x \log x - x$ which is the $U(.)$ function of the von Neumann divergence.
Note: $U_\lambda(x)$ is convex when $x > 0, \lambda > 0$ and
$U'_\lambda(x) = \frac{1+\lambda}{\lambda}\left[x^\lambda - 1\right] = (1+\lambda)G_\lambda(x)$. For a particular t :

$$d_\lambda(f(t), g_\theta(t)) = U_\lambda(f(t)) - U_\lambda(g(t,\theta)) - (f(t) - g(t,\theta))U'_\lambda(g(t,\theta)).$$

We then define the global distance:
$d_\lambda(f, g_\theta) = \int d_\lambda(f(t), g_\theta(t))dt$. Can show: as $\lambda \rightarrow 0 \rightarrow$ get KL.

To locally adapt: θ is made locally dependent on the point at which the density is evaluated hoping that $g_h(t) \propto g(t, \hat{\theta}_{t,h})$ would be “better" than just $g(t, \hat{\theta})$.
Replacing log by the $G_\lambda \rightarrow$ end up analysing at t :

$$K_h(x-t)\{U_\lambda(f(x)) - U_\lambda(g(x,\theta)) - (f(x) - g(x,\theta))U'_\lambda(g(x,\theta))\}.$$

For a fixed t, the global divergence $d_{\lambda,t}(f,g)$ equals

$$\int K_h(x-t)\{U_\lambda(f(x)) - U_\lambda(g(x,\theta)) - (f(x) - g(x,\theta))U'_\lambda(g(x,\theta))\}dx.$$

For $\lambda > 0 \rightarrow$ maximizing

$$\int_{\mathbb{R}^d} \rho_\lambda(t,x,\theta)dF(x) \tag{4}$$

w.r. θ where the integrand is

$$\left(\frac{\lambda+1}{\lambda}\right) K\left(\frac{x-t}{h}\right)\{g(x,\theta)^\lambda - 1\} - \int_{\mathbb{R}^d} K\left(\frac{s-t}{h}\right) g(s,\theta)^{\lambda+1} ds \tag{5}$$

or: solve

$$\int_{\mathbb{R}^d} \psi_\lambda(t,x,\theta)dF(x) = 0, \tag{6}$$

where

$$\psi_\lambda = K\left(\frac{x-t}{h}\right) g(x,\theta)^\lambda u(x,\theta) - \int_{\mathbb{R}^d} K\left(\frac{s-t}{h}\right) g(s,\theta)^{\lambda+1} u(s,\theta)ds \tag{7}$$

and

$$u(x,\theta) = [u_1(x,\theta) \ \cdots \ u_p(x,\theta)]^T = (\partial/\partial\theta)\log g(x,\theta).$$

Empirically: get an M-estimator equation $0 =$

$$\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{X_i-t}{h}\right) g(X_i,\theta)^\lambda u(X_i,\theta) - \int K\left(\frac{s-t}{h}\right) g(s,\theta)^{\lambda+1} u(s,\theta)ds. \tag{8}$$

Outline

This requires investigation to limit as $\lambda \searrow 0$. It can easily be seen that it leads to a corresponding local M-estimator:

$$\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - t)\left[\frac{\partial}{\partial\theta}\log g(X_i, \theta_0)\right] - \int K_h(s-t)\frac{\partial}{\partial\theta} g(s, \theta_0) ds = 0.$$

We recover Hjort & Jones's local likelihood.

We: interested in $\lambda > 0$ to achieve a compromise. In the global setting (h is very large): Basu et. al.: $\lambda = 0$: efficiency, $\lambda = 1$: very robust. In our localised setting: h is still large but grows with a rate that depends on the sample size & $\lambda > 0$ allows us to attain a compromise between robustness and efficiency along the curve.

New insight: not focusing on modifying non-robust estimators of *parameters* where the inference part: finalized once the parameters have been estimated. We estimate local features of the density that has generated the data; $g(t, \hat{\theta}(t))$ does not belong to the class $g(t, \theta), \theta \in \Theta$ and perhaps gives us a better idea about the density that has generated the data. But: similarities, too. The belief that the true density is "not too far away" from a particular model density $g(x, \theta_0)$ is common:

Structural Assumption A*: $f(x) = (1 - \delta_n)g(x, \theta^*) + \delta_n h(x)$, $\delta_{n n \to \infty} \to 0$ and finite first moments for $f(x), g(x, \theta^*)$.

Without **A*** a purely non-parametric estimator would certainly be better in the limit as $n \to \infty$. Now: local and global robustness-based estimators are expected to perform better than the non-parametric: it is legitimate to compare them.
For comparison: global $\hat{\theta}$:

$$\sum_{i=1}^{n} \{g(x_i, \theta)^{\lambda} u(x_i, \theta) - \int g(x, \theta)^{\lambda+1} u(x, \theta) dx\} = 0.$$

Local estimator $\hat{\theta}(t)$: for each fixed t

$$\sum_{i=1}^{n} \{K(\frac{x_i - t}{h}) g(x_i, \theta)^{\lambda} u(x_i, \theta) - \int K(\frac{x - t}{h}) g(x, \theta)^{\lambda+1} u(x, \theta) dx\} = 0.$$

For kernel:$K(t) = 1 - \kappa_2 t^2 + \kappa_4 t^4 + o(t^4)$ with $\kappa_2, \kappa_4 > 0$. Typical: $K(t) = \exp(-\frac{t^2}{2})$.

Introduce vector functions

$$\psi_t(x,\theta) = K(\frac{x-t}{h})g(x,\theta)^{\lambda}u(x,\theta) - \int K(\frac{x-t}{h})g(x,\theta)^{\lambda+1}u(x,\theta)dx,$$

$$\psi(x,\theta) = g(x,\theta)^{\lambda}u(x,\theta) - \int g(x,\theta)^{\lambda+1}u(x,\theta)dx$$

$$\psi_t^{(k)}(x,\theta) = (x-t)^k g(x,\theta)^{\lambda}u(x,\theta) - \int (x-t)^k g(x,\theta)^{\lambda+1}u(x,\theta)dx.$$

- The Fisher consistency: $\int \psi(x,\theta_0)dF(x) = 0$.
- The local: $\hat{\theta}(t) : \int \psi_t(x,\theta)dF_n(x) = 0$.
- The global: $\hat{\theta} : \int \psi(x,\theta)dF_n(x) = 0$.

Introduce vector functions

$$\psi_t(x,\theta) = K(\frac{x-t}{h})g(x,\theta)^{\lambda}u(x,\theta) - \int K(\frac{x-t}{h})g(x,\theta)^{\lambda+1}u(x,\theta)dx,$$

$$\psi(x,\theta) = g(x,\theta)^{\lambda}u(x,\theta) - \int g(x,\theta)^{\lambda+1}u(x,\theta)dx$$

$$\psi_t^{(k)}(x,\theta) = (x-t)^k g(x,\theta)^{\lambda}u(x,\theta) - \int (x-t)^k g(x,\theta)^{\lambda+1}u(x,\theta)dx.$$

- The Fisher consistency: $\int \psi(x,\theta_0)dF(x) = 0$.
- The local: $\hat{\theta}(t) : \int \psi_t(x,\theta)dF_n(x) = 0$.
- The global: $\hat{\theta} : \int \psi(x,\theta)dF_n(x) = 0$.

Introduce vector functions

$$\psi_t(x,\theta) = K(\frac{x-t}{h})g(x,\theta)^{\lambda}u(x,\theta) - \int K(\frac{x-t}{h})g(x,\theta)^{\lambda+1}u(x,\theta)dx,$$

$$\psi(x,\theta) = g(x,\theta)^{\lambda}u(x,\theta) - \int g(x,\theta)^{\lambda+1}u(x,\theta)dx$$

$$\psi_t^{(k)}(x,\theta) = (x-t)^k g(x,\theta)^{\lambda}u(x,\theta) - \int (x-t)^k g(x,\theta)^{\lambda+1}u(x,\theta)dx.$$

- The Fisher consistency: $\int \psi(x,\theta_0)dF(x) = 0$.
- The local: $\hat{\theta}(t) : \int \psi_t(x,\theta)dF_n(x) = 0$.
- The global: $\hat{\theta} : \int \psi(x,\theta)dF_n(x) = 0$.

We define $p \times p$ matrices:

$$\Psi_t(\theta) = \int \frac{\partial}{\partial \theta} \psi_t(x,\theta)^T dF(x), \widehat{\Psi}_t(\theta) = \int \frac{\partial}{\partial \theta} \psi_t(x,\theta)^T dF_n(x),$$

$$\Psi(\theta) = \int \frac{\partial}{\partial \theta} \psi(x,\theta)^T dF(x), \widehat{\Psi}(\theta) = \int \frac{\partial}{\partial \theta} \psi(x,\theta)^T dF_n(x),$$

$$\Psi_t^{(\ell)}(\theta) = \int \frac{\partial}{\partial \theta} \psi_t^{(\ell)}(x,\theta)^T dF(x), \widehat{\Psi}_t^{(\ell)}(\theta) = \int \frac{\partial}{\partial \theta} \psi_t^{(\ell)}(x,\theta)^T dF_n(x).$$

- **Condition A1.** $n \to \infty, h \to \infty$ and $h^2 = O(\sqrt{n})$.
- **Condition A2.** For any $t \in \mathbb{R}^d$ and any vector $\tilde{\theta}$ satisfying $||\tilde{\theta} - \hat{\theta}|| < ||\hat{\theta}(t) - \hat{\theta}||$,
  $\max_{1 \le i,j \le p} \left| \Psi_t(\tilde{\theta})_{ij} - \Psi_t(\theta_0)_{ij} \right| = o\left(\frac{1}{h^2}\right)$ and
  $\max_{1 \le i,j \le p} \left| \widehat{\Psi}_t(\tilde{\theta})_{ij} - \widehat{\Psi}_t(\theta_0)_{ij} \right| = o_p\left(\frac{1}{h^2}\right)$ hold as $n, h \to \infty$.
- **Condition A3.** $\int \frac{\partial}{\partial \theta^T} \psi_t(x,\tilde{\theta}) dF_n(x)$ is a.s. non-singular for $\tilde{\theta}$ satisfying $\| \tilde{\theta} - \hat{\theta} \| < \| \hat{\theta}(t) - \hat{\theta} \|$.

We define $p \times p$ matrices:

$$\Psi_t(\theta) = \int \frac{\partial}{\partial \theta} \psi_t(x,\theta)^T dF(x), \widehat{\Psi}_t(\theta) = \int \frac{\partial}{\partial \theta} \psi_t(x,\theta)^T dF_n(x),$$

$$\Psi(\theta) = \int \frac{\partial}{\partial \theta} \psi(x,\theta)^T dF(x), \widehat{\Psi}(\theta) = \int \frac{\partial}{\partial \theta} \psi(x,\theta)^T dF_n(x),$$

$$\Psi_t^{(\ell)}(\theta) = \int \frac{\partial}{\partial \theta} \psi_t^{(\ell)}(x,\theta)^T dF(x), \widehat{\Psi}_t^{(\ell)}(\theta) = \int \frac{\partial}{\partial \theta} \psi_t^{(\ell)}(x,\theta)^T dF_n(x).$$

- **Condition A1.** $n \to \infty, h \to \infty$ and $h^2 = O(\sqrt{n})$.
- **Condition A2.** For any $t \in \mathbb{R}^d$ and any vector $\tilde{\theta}$ satisfying $||\tilde{\theta} - \hat{\theta}|| < ||\hat{\theta}(t) - \hat{\theta}||$,
  $\max_{1 \le i,j \le p} \left| \Psi_t(\tilde{\theta})_{ij} - \Psi_t(\theta_0)_{ij} \right| = o\left(\frac{1}{h^2}\right)$ and
  $\max_{1 \le i,j \le p} \left| \widehat{\Psi}_t(\tilde{\theta})_{ij} - \widehat{\Psi}_t(\theta_0)_{ij} \right| = o_p\left(\frac{1}{h^2}\right)$ hold as $n, h \to \infty$.
- **Condition A3.** $\int \frac{\partial}{\partial \theta^T} \psi_t(x,\tilde{\theta}) dF_n(x)$ is a.s. non-singular for $\tilde{\theta}$ satisfying $\| \tilde{\theta} - \hat{\theta} \| < \| \hat{\theta}(t) - \hat{\theta} \|$.

We define $p \times p$ matrices:

$$\Psi_t(\theta) = \int \frac{\partial}{\partial \theta} \psi_t(x,\theta)^T dF(x), \widehat{\Psi}_t(\theta) = \int \frac{\partial}{\partial \theta} \psi_t(x,\theta)^T dF_n(x),$$

$$\Psi(\theta) = \int \frac{\partial}{\partial \theta} \psi(x,\theta)^T dF(x), \widehat{\Psi}(\theta) = \int \frac{\partial}{\partial \theta} \psi(x,\theta)^T dF_n(x),$$

$$\Psi_t^{(\ell)}(\theta) = \int \frac{\partial}{\partial \theta} \psi_t^{(\ell)}(x,\theta)^T dF(x), \widehat{\Psi}_t^{(\ell)}(\theta) = \int \frac{\partial}{\partial \theta} \psi_t^{(\ell)}(x,\theta)^T dF_n(x).$$

- **Condition A1.** $n \to \infty, h \to \infty$ and $h^2 = O(\sqrt{n})$.
- **Condition A2.** For any $t \in \mathbb{R}^d$ and any vector $\tilde{\theta}$ satisfying $||\tilde{\theta} - \hat{\theta}|| < ||\hat{\theta}(t) - \hat{\theta}||$,
  $\max_{1 \le i,j \le p} \left| \Psi_t(\tilde{\theta})_{ij} - \Psi_t(\theta_0)_{ij} \right| = o\left(\frac{1}{h^2}\right)$ and
  $\max_{1 \le i,j \le p} \left| \widehat{\Psi}_t(\tilde{\theta})_{ij} - \widehat{\Psi}_t(\theta_0)_{ij} \right| = o_p\left(\frac{1}{h^2}\right)$ hold as $n, h \to \infty$.
- **Condition A3.** $\int \frac{\partial}{\partial \theta^T} \psi_t(x,\tilde{\theta}) dF_n(x)$ is a.s. non-singular for $\tilde{\theta}$ satisfying $\| \tilde{\theta} - \hat{\theta} \| < \| \hat{\theta}(t) - \hat{\theta} \|$.

Statements:

Lemma

- ***Statement 1).*** *Under A1) and A2), as $n \to \infty, h \to \infty$, for any vector $\tilde{\theta}$ satisfying $||\tilde{\theta} - \hat{\theta}|| < ||\hat{\theta}(t) - \hat{\theta}||$:*
  $\widehat{\Psi}_t(\tilde{\theta}) = \Psi(\theta_0) - \frac{1}{h^2}\widehat{V}_t(\theta_0) + o_p\left(\frac{1}{h^2}\right)$, *where*
  $\widehat{V}_t(\theta) = \kappa_2 \Psi_t^{(2)}(\theta) - \left(\frac{h^2}{\sqrt{n}}\right)\sqrt{n}\left\{\widehat{\Psi}(\theta) - \Psi(\theta)\right\}$.
- ***Statement 2).*** *Under A1), A2), A3), as $n \to \infty, h \to \infty$: for any $t \in \mathbb{R}^d$ and any vector $\tilde{\theta}$ with $||\tilde{\theta} - \hat{\theta}|| < ||\hat{\theta}(t) - \hat{\theta}||$:*
  $\widehat{\Psi}_t(\tilde{\theta})^{-1} = \Psi(\theta_0)^{-1} + \frac{1}{h^2}\Psi(\theta_0)^{-1}\widehat{V}_t(\theta_0)\Psi(\theta_0)^{-1} + o_p\left(\frac{1}{h^2}\right)$.
- ***Statement 3).*** $\int \psi_t^{(\ell)}(x, \hat{\theta}) dF_n(x) =$
  $\int \psi_t^{(\ell)}(x, \theta_0) dF(x) + \frac{1}{\sqrt{n}}\hat{v}_t^{(\ell)}(\theta_0) + o_p\left(\frac{1}{\sqrt{n}}\right)$, *where*
  $\hat{v}_t^{(\ell)}(\theta) = \sqrt{n}\left\{\int \psi_t^{(\ell)}(x, \theta) dF_n(x) - \int \psi_t^{(\ell)}(x, \theta) dF(x)\right\} +$
  $\Psi_t^{(\ell)}(\theta)\sqrt{n}(\hat{\theta} - \theta), \ell = 2, 4.$

Statements:

Lemma

- ***Statement 1).*** *Under A1) and A2), as $n \to \infty, h \to \infty$, for any vector $\tilde{\theta}$ satisfying $||\tilde{\theta} - \hat{\theta}|| < ||\hat{\theta}(t) - \hat{\theta}||$:*
  $\widehat{\Psi}_t(\tilde{\theta}) = \Psi(\theta_0) - \frac{1}{h^2}\widehat{V}_t(\theta_0) + o_p\left(\frac{1}{h^2}\right)$, *where*
  $\widehat{V}_t(\theta) = \kappa_2 \Psi_t^{(2)}(\theta) - \left(\frac{h^2}{\sqrt{n}}\right)\sqrt{n}\left\{\widehat{\Psi}(\theta) - \Psi(\theta)\right\}$.
- ***Statement 2).*** *Under A1), A2), A3), as $n \to \infty, h \to \infty$: for any $t \in \mathbb{R}^d$ and any vector $\tilde{\theta}$ with $||\tilde{\theta} - \hat{\theta}|| < ||\hat{\theta}(t) - \hat{\theta}||$:*
  $\widehat{\Psi}_t(\tilde{\theta})^{-1} = \Psi(\theta_0)^{-1} + \frac{1}{h^2}\Psi(\theta_0)^{-1}\widehat{V}_t(\theta_0)\Psi(\theta_0)^{-1} + o_p\left(\frac{1}{h^2}\right)$.
- ***Statement 3).*** $\int \psi_t^{(\ell)}(x, \hat{\theta}) dF_n(x) =$
  $\int \psi_t^{(\ell)}(x, \theta_0) dF(x) + \frac{1}{\sqrt{n}}\hat{v}_t^{(\ell)}(\theta_0) + o_p\left(\frac{1}{\sqrt{n}}\right)$, *where*
  $\hat{v}_t^{(\ell)}(\theta) = \sqrt{n}\left\{\int \psi_t^{(\ell)}(x, \theta) dF_n(x) - \int \psi_t^{(\ell)}(x, \theta) dF(x)\right\} +$
  $\Psi_t^{(\ell)}(\theta)\sqrt{n}(\hat{\theta} - \theta), \ell = 2, 4.$

Statements:

Lemma

- ***Statement 1).*** *Under A1) and A2), as $n \to \infty, h \to \infty$, for any vector $\tilde{\theta}$ satisfying $||\tilde{\theta} - \hat{\theta}|| < ||\hat{\theta}(t) - \hat{\theta}||$:*
  $\widehat{\Psi}_t(\tilde{\theta}) = \Psi(\theta_0) - \frac{1}{h^2}\widehat{V}_t(\theta_0) + o_p\left(\frac{1}{h^2}\right)$, *where*
  $\widehat{V}_t(\theta) = \kappa_2 \Psi_t^{(2)}(\theta) - \left(\frac{h^2}{\sqrt{n}}\right)\sqrt{n}\left\{\widehat{\Psi}(\theta) - \Psi(\theta)\right\}$.
- ***Statement 2).*** *Under A1), A2), A3), as $n \to \infty, h \to \infty$: for any $t \in \mathbb{R}^d$ and any vector $\tilde{\theta}$ with $||\tilde{\theta} - \hat{\theta}|| < ||\hat{\theta}(t) - \hat{\theta}||$:*
  $\widehat{\Psi}_t(\tilde{\theta})^{-1} = \Psi(\theta_0)^{-1} + \frac{1}{h^2}\Psi(\theta_0)^{-1}\widehat{V}_t(\theta_0)\Psi(\theta_0)^{-1} + o_p\left(\frac{1}{h^2}\right)$.
- ***Statement 3).*** $\int \psi_t^{(\ell)}(x, \hat{\theta}) dF_n(x) =$
  $\int \psi_t^{(\ell)}(x, \theta_0) dF(x) + \frac{1}{\sqrt{n}}\hat{v}_t^{(\ell)}(\theta_0) + o_p\left(\frac{1}{\sqrt{n}}\right)$, *where*
  $\hat{v}_t^{(\ell)}(\theta) = \sqrt{n}\left\{\int \psi_t^{(\ell)}(x, \theta) dF_n(x) - \int \psi_t^{(\ell)}(x, \theta) dF(x)\right\} +$
  $\Psi_t^{(\ell)}(\theta)\sqrt{n}(\hat{\theta} - \theta), \ell = 2, 4.$

Statements:

Lemma

- ***Statement 1).*** *Under A1) and A2), as $n \to \infty, h \to \infty$, for any vector $\tilde{\theta}$ satisfying $||\tilde{\theta} - \hat{\theta}|| < ||\hat{\theta}(t) - \hat{\theta}||$:*
  $\widehat{\Psi}_t(\tilde{\theta}) = \Psi(\theta_0) - \frac{1}{h^2}\widehat{V}_t(\theta_0) + o_p\left(\frac{1}{h^2}\right)$, *where*
  $\widehat{V}_t(\theta) = \kappa_2 \Psi_t^{(2)}(\theta) - \left(\frac{h^2}{\sqrt{n}}\right)\sqrt{n}\left\{\widehat{\Psi}(\theta) - \Psi(\theta)\right\}$.
- ***Statement 2).*** *Under A1), A2), A3), as $n \to \infty, h \to \infty$: for any $t \in \mathbb{R}^d$ and any vector $\tilde{\theta}$ with $||\tilde{\theta} - \hat{\theta}|| < ||\hat{\theta}(t) - \hat{\theta}||$:*
  $\widehat{\Psi}_t(\tilde{\theta})^{-1} = \Psi(\theta_0)^{-1} + \frac{1}{h^2}\Psi(\theta_0)^{-1}\widehat{V}_t(\theta_0)\Psi(\theta_0)^{-1} + o_p\left(\frac{1}{h^2}\right)$.
- ***Statement 3).*** $\int \psi_t^{(\ell)}(x, \hat{\theta}) dF_n(x) =$
  $\int \psi_t^{(\ell)}(x, \theta_0) dF(x) + \frac{1}{\sqrt{n}}\hat{v}_t^{(\ell)}(\theta_0) + o_p\left(\frac{1}{\sqrt{n}}\right)$, *where*
  $\hat{v}_t^{(\ell)}(\theta) = \sqrt{n}\left\{\int \psi_t^{(\ell)}(x, \theta) dF_n(x) - \int \psi_t^{(\ell)}(x, \theta) dF(x)\right\} +$
  $\Psi_t^{(\ell)}(\theta)\sqrt{n}(\hat{\theta} - \theta), \ell = 2, 4.$

Lemma

Under conditions A1), A2) and A3), for any $t \in \mathbb{R}^d$, it holds

$$\begin{aligned}
\hat{\theta}(t) - \hat{\theta} \quad = \quad & \left(\frac{\kappa_2}{h^2}\right) \Psi(\theta_0)^{-1} \int_{\mathbb{R}^d} \psi_t^{(2)}(x, \theta_0) dF(x) \\
& + \frac{\kappa_2}{h^2\sqrt{n}} \Psi(\theta_0)^{-1} \hat{v}_t^{(2)}(\theta_0) \\
& + \frac{\kappa_2}{h^4} \Psi(\theta_0)^{-1} \widehat{V}_t(\theta_0) \Psi(\theta_0)^{-1} \int_{\mathbb{R}^d} \psi_t^{(2)}(x, \theta_0) dF(x) \\
& - \frac{\kappa_4}{h^4} \Psi(\theta_0)^{-1} \int_{\mathbb{R}^d} \psi_t^{(4)}(x, \theta_0) dF(x) + o_p\left(\frac{1}{h^4}\right).
\end{aligned}$$

Theorem

Under Conditions A1), A2) and A3), for any $t \in \mathbb{R}^d$ we have

$$E_f[\hat{\theta}(t) - \hat{\theta}] = -\frac{\kappa_2}{h^2}\left[-\int \frac{\partial}{\partial \theta^T}\psi(x,\theta_0)dF(x)\right]^{-1}\int \psi_t^{(2)}(x,\theta_0)dF(x) +$$
$$\frac{\kappa_4}{h^4}\left[-\int \frac{\partial}{\partial \theta^T}\psi(x,\theta_0)dF(x)\right]^{-1}\int \frac{\partial}{\partial \theta^T}\psi_t^{(4)}(x,\theta_0)dF(x) +$$
$$\frac{\kappa_2^2}{h^4}\left[-\int \frac{\partial}{\partial \theta^T}\psi(x,\theta_0)dF(x)\right]^{-1}\int \frac{\partial}{\partial \theta^T}\psi_t^{(2)}(x,\theta_0)dF(x) *$$
$$\left[-\int \frac{\partial}{\partial \theta^T}\psi(x,\theta_0)dF(x)\right]^{-1}\int \psi_t^{(2)}(x,\theta_0)dF(x) + o(\frac{1}{h^4}).$$

Note. This generalizes in many directions of Lemma 1 of Eguchi & Copas (1998).

Now: we introduce a risk measure when estimating the density f by an estimator $\bar{f}$: $R(\bar{f},f) = \mathbf{E}_f \int d_\lambda[f(t),\bar{f}(t)]dt$.

$$\mathbf{E}_f \int \bar{f}^\lambda(t)\{\bar{f}(t) - \frac{\lambda+1}{\lambda} f(t)\}dt.$$

Theorem

(simplified) There exists $\lambda_0 > 0$ such that for all $\lambda \in [0,\lambda_0)$

$$R(\bar{f},f) - R(\hat{f},f) = \frac{2(1+\lambda)\kappa_2}{h^2} B^T I_\lambda(\theta_0)^{-1} B + o(\frac{1}{h^2}),$$

$$B = E_{g_{\theta_0}-f}\left[x g(x,\theta_0)^\lambda u(x,\theta_0)\right],$$

$$I_\lambda(\theta_0) = E_f[g(x,\theta_0)^\lambda u(x,\theta_0) u(x,\theta_0)^T].$$

Note: If δ_n is sufficiently small in (*) there is no expectation for the proposed localization to bring significant improvement. We can show that if $\delta_n\sqrt{n} \to c \geq 0$ then $\sqrt{n}(\hat{f}(x) - f(x)$ is asymptotically normal with mean $c\{g(x,\theta^*) - h(x)\}$ and finite variance and if $\delta_n = o(1/\sqrt{n})$ then there is no effect of the localization.

Take true distribution: skew normal, small $\alpha > 0 : g(x) = 2\phi(x)\Phi(\alpha x)$ where $\phi(.)(\alpha = 0 :$ st. normal).
Practitioner assumes that perhaps there is contamination & would like to use robust method to estimate $\theta = (\mu, \sigma)^T$.
If a globally robust method based on minimization of the λ-Bregman distance is applied: need to solve

$$\mu = \frac{\int t\phi_{\sigma/\sqrt{\lambda}}(t-\mu)f(t)dt}{\int \phi_{\sigma/\sqrt{\lambda}}(t-\mu)f(t)dt}$$

$$-\frac{\lambda}{\sqrt{\lambda+1}} = \int e^{-\frac{(t-\mu)^2\lambda}{2\sigma^2}} f(t)dt - \frac{1}{\sigma^2}\int (t-\mu)^2 e^{-\frac{(t-\mu)^2\lambda}{2\sigma^2}} f(t)dt. \quad (9)$$

Get non-linear equation system:

$$\mu = \frac{\sum_{i=1}^{n} X_i \phi_{\sigma/\sqrt{\lambda}}(X_i - \mu)}{\sum_{i=1}^{n} \phi_{\sigma/\sqrt{\lambda}}(X_i - \mu)}$$

$$-\frac{\lambda}{\sqrt{\lambda+1}} = \sum_{i=1}^{n} e^{-\frac{(X_i-\mu)^2\lambda}{2\sigma^2}} - \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 e^{-\frac{(X_i-\mu)^2\lambda}{2\sigma^2}}. \qquad (10)$$

(Note: if $\lambda = 0$: reduce: $\hat{\mu} = \bar{X}, \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$).

For **locally robust** (and with $K(x) = e^{-\frac{1}{2}x^2}$: utilising:

$$\phi_\sigma(x-\mu)\phi_{\sigma'}(x-\mu') = \phi_{(\sigma^2+\sigma'^2)^{1/2}}(\mu-\mu')\phi_{\sigma\sigma'/(\sigma^2+\sigma'^2)^{1/2}}(x-\mu^*) \qquad (11)$$

where $\mu^* = \frac{\sigma'^2\mu+\sigma^2\mu'}{\sigma^2+\sigma'^2}$: Let $\mu_\lambda^* = \frac{\frac{\sigma^2}{\lambda}x+h^2\mu}{h^2+\frac{\sigma^2}{\lambda}}$ and $\sigma_\lambda^* = \sqrt{\frac{\frac{h^2\sigma^2}{\lambda}}{h^2+\frac{\sigma^2}{\lambda}}}$.

Then for each fixed x, the local estimators $\mu = \mu(x)$ and $\sigma = \sigma(x)$ satisfy:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{\lambda}(2\pi\sigma^2)^{(\lambda-1)/2}}\phi_{\sqrt{h^2+\frac{\sigma^2}{\lambda}}}(x-\mu)\phi_{\sigma^*_\lambda}(X_i-\mu^*_\lambda)\frac{X_i-\mu}{\sigma^2}=$$

$$\frac{1}{\sqrt{\lambda+1}(2\pi\sigma^2)^{\lambda/2}}\phi_{\sqrt{h^2+\frac{\sigma^2}{\lambda+1}}}(x-\mu)\frac{\mu^*_{\lambda+1}-\mu}{\sigma^2}$$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{\lambda}(2\pi\sigma^2)^{(\lambda-1)/2}}\phi_{\sqrt{h^2+\frac{\sigma^2}{\lambda}}}(x-\mu)\phi_{\sigma^*_\lambda}(X_i-\mu^*_\lambda)[-\frac{1}{2\sigma^2}+\frac{(X_i-\mu)^2}{2\sigma^4}]=$$

$$\frac{1}{\sqrt{\lambda+1}(2\pi\sigma^2)^{\lambda/2}}\phi_{\sqrt{h^2+\frac{\sigma^2}{\lambda+1}}}(x-\mu)[-\frac{1}{2\sigma^2}+\frac{\sigma^{*2}_{\lambda+1}+(\mu^*_{\lambda+1}-\mu)^2}{2\sigma^4}]$$

We have implemented the global and the locally robust estimation procedures. Adopted MATLAB's `fmincon`.

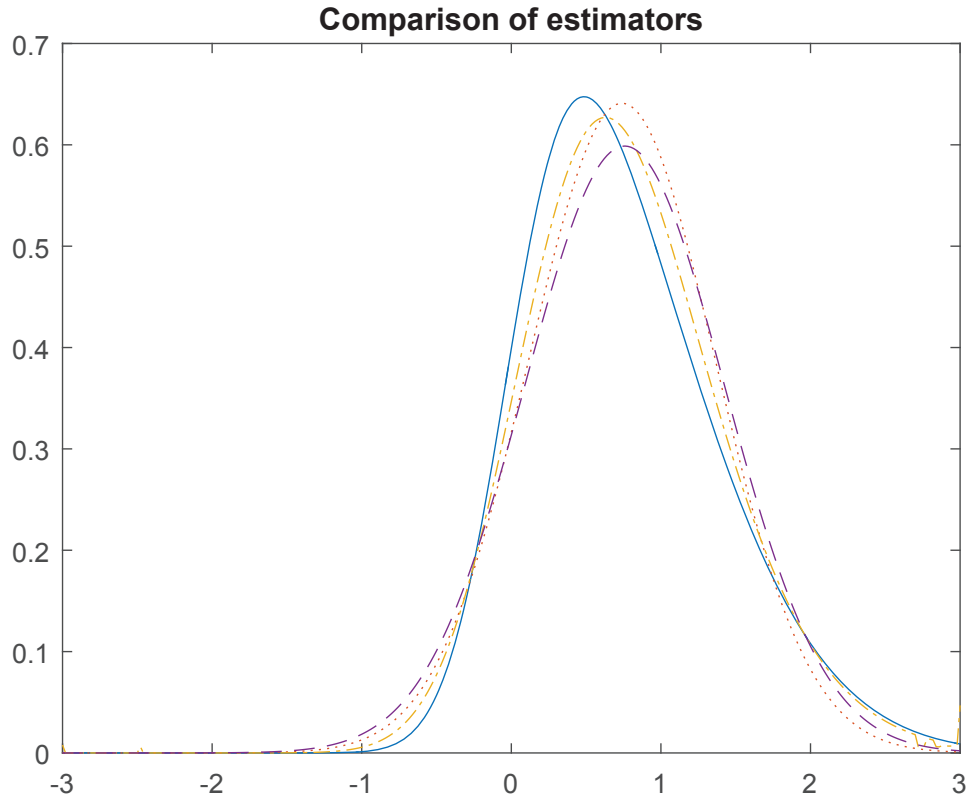
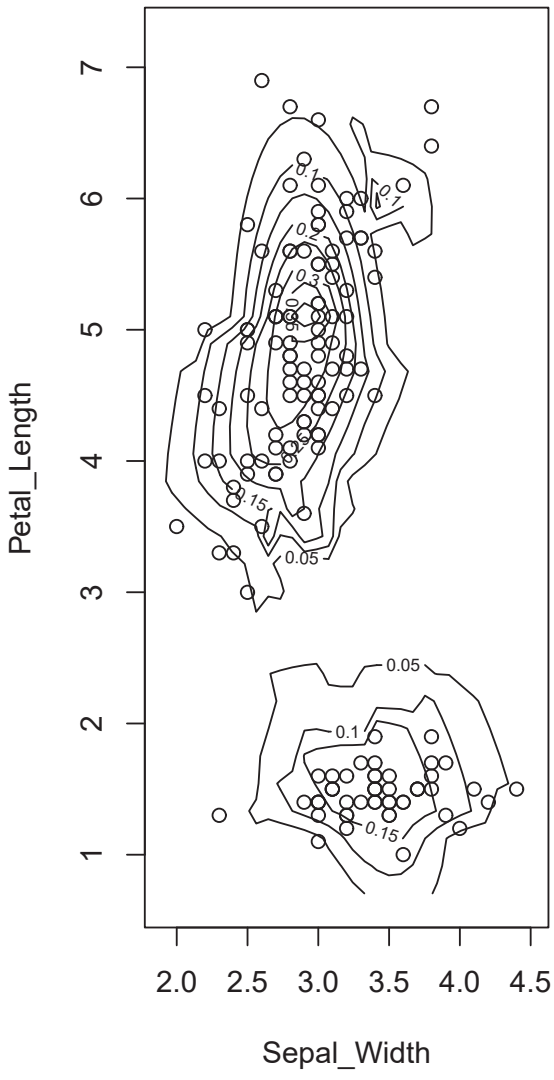
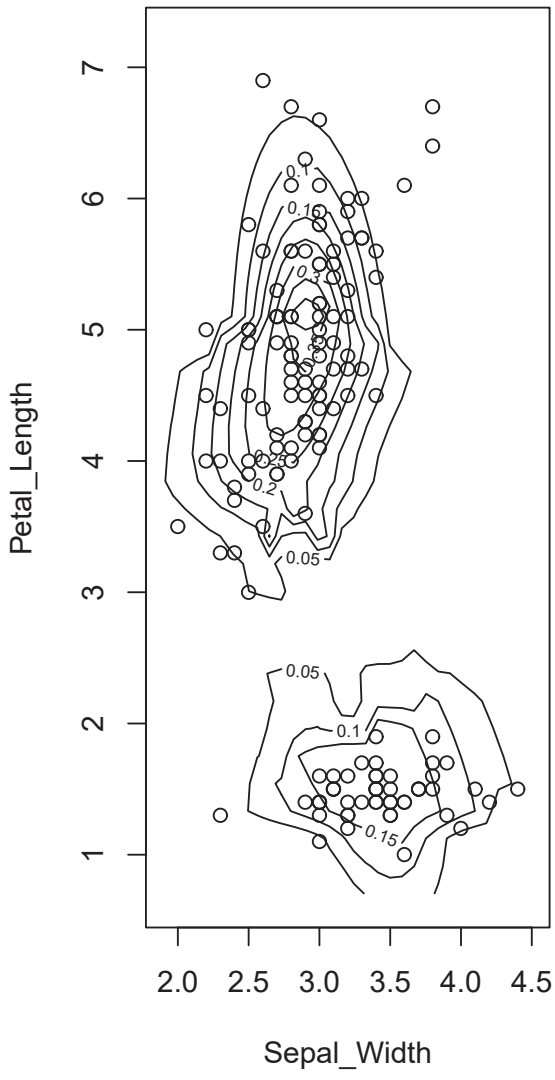


Figure: *True density: solid line; Locally robust: dot-dashed line; globally robust: dotted line; normal approximation: dashed line*

Real Data Example in dim. 2 Use: Part of Fisher's Iris data with (`Sepal Width`, `Petal Length`), where $h = \sqrt{3/8}$ was utilized. The obtained contour plots of the density estimates with $\lambda = 0$ and $\lambda = 0.2$ are drawn in the next Figure. An apparent difference can be observed in the right upper area in both panels: a data point have been included in contour of density estimate in the left panel with $\lambda = 0$, meanwhile it has been shaved off from the contour in the right panel with $\lambda = 0.2$. This again indicates that a nonzero $\lambda > 0$ contributes to robustification.


Petal_Length
Sepal_Width
1
2
3
4
5
6
7
2.0
2.5
3.0
3.5
4.0
4.5
0.05
0.1
0.15
0.2
0.3
0.4


Petal_Length
Sepal_Width
1
2
3
4
5
6
7
2.0
2.5
3.0
3.5
4.0
4.5
0.05
0.1
0.15
0.2
0.3

- Basu, A., Harris, I., Hjort, N. and Jones, M.C. (1998) *Biometrika*, 85, 549–559.
- Hjort, N. and Glad, I. (1995) *AS*, 23, 882–904.
- Hjort, N. and Jones, M.C. (1996) *AS*, 24, 1619–1647.
- Huber, P. and Ronchetti, E. (2009) *Robust Statistics*, 2nd Edition. Wiley, New York.
- Jones, M.C., Hjort, N., Harris, I. and Basu, A. (2001) *Biometrika*, 88, 865–873.
- Loader, C. (1996) *AS*, 24, 1602–1618.
- Park, B., Lee, Y., Kim, T., Park, C. and Eguchi, S. (2006) *JSPI*, (136), 3, 839–859.
- Tibshirani, R. and Hastie, T. (1987) *JASA* 82, 559–567.
- Windham, M. (1995) *JRSS B*, 57, 599–609.

- Basu, A., Harris, I., Hjort, N. and Jones, M.C. (1998) *Biometrika*, 85, 549–559.
- Hjort, N. and Glad, I. (1995) *AS*, 23, 882–904.
- Hjort, N. and Jones, M.C. (1996) *AS*, 24, 1619–1647.
- Huber, P. and Ronchetti, E. (2009) *Robust Statistics*, 2nd Edition. Wiley, New York.
- Jones, M.C., Hjort, N., Harris, I. and Basu, A. (2001) *Biometrika*, 88, 865–873.
- Loader, C. (1996) *AS*, 24, 1602–1618.
- Park, B., Lee, Y., Kim, T., Park, C. and Eguchi, S. (2006) *JSPI*, (136), 3, 839–859.
- Tibshirani, R. and Hastie, T. (1987) *JASA* 82, 559–567.
- Windham, M. (1995) *JRSS B*, 57, 599–609.

- Basu, A., Harris, I., Hjort, N. and Jones, M.C. (1998) *Biometrika*, 85, 549–559.
- Hjort, N. and Glad, I. (1995) *AS*, 23, 882–904.
- Hjort, N. and Jones, M.C. (1996) *AS*, 24, 1619–1647.
- Huber, P. and Ronchetti, E. (2009) *Robust Statistics*, 2nd Edition. Wiley, New York.
- Jones, M.C., Hjort, N., Harris, I. and Basu, A. (2001) *Biometrika*, 88, 865–873.
- Loader, C. (1996) *AS*, 24, 1602–1618.
- Park, B., Lee, Y., Kim, T., Park, C. and Eguchi, S. (2006) *JSPI*, (136), 3, 839–859.
- Tibshirani, R. and Hastie, T. (1987) *JASA* 82, 559–567.
- Windham, M. (1995) *JRSS B*, 57, 599–609.

- Basu, A., Harris, I., Hjort, N. and Jones, M.C. (1998) *Biometrika*, 85, 549–559.
- Hjort, N. and Glad, I. (1995) *AS*, 23, 882–904.
- Hjort, N. and Jones, M.C. (1996) *AS*, 24, 1619–1647.
- Huber, P. and Ronchetti, E. (2009) *Robust Statistics*, 2nd Edition. Wiley, New York.
- Jones, M.C., Hjort, N., Harris, I. and Basu, A. (2001) *Biometrika*, 88, 865–873.
- Loader, C. (1996) *AS*, 24, 1602–1618.
- Park, B., Lee, Y., Kim, T., Park, C. and Eguchi, S. (2006) *JSPI*, (136), 3, 839–859.
- Tibshirani, R. and Hastie, T. (1987) *JASA* 82, 559–567.
- Windham, M. (1995) *JRSS B*, 57, 599–609.

- Basu, A., Harris, I., Hjort, N. and Jones, M.C. (1998) *Biometrika*, 85, 549–559.
- Hjort, N. and Glad, I. (1995) *AS*, 23, 882–904.
- Hjort, N. and Jones, M.C. (1996) *AS*, 24, 1619–1647.
- Huber, P. and Ronchetti, E. (2009) *Robust Statistics*, 2nd Edition. Wiley, New York.
- Jones, M.C., Hjort, N., Harris, I. and Basu, A. (2001) *Biometrika*, 88, 865–873.
- Loader, C. (1996) *AS*, 24, 1602–1618.
- Park, B., Lee, Y., Kim, T., Park, C. and Eguchi, S. (2006) *JSPI*, (136), 3, 839–859.
- Tibshirani, R. and Hastie, T. (1987) *JASA* 82, 559–567.
- Windham, M. (1995) *JRSS B*, 57, 599–609.

- Basu, A., Harris, I., Hjort, N. and Jones, M.C. (1998) *Biometrika*, 85, 549–559.
- Hjort, N. and Glad, I. (1995) *AS*, 23, 882–904.
- Hjort, N. and Jones, M.C. (1996) *AS*, 24, 1619–1647.
- Huber, P. and Ronchetti, E. (2009) *Robust Statistics*, 2nd Edition. Wiley, New York.
- Jones, M.C., Hjort, N., Harris, I. and Basu, A. (2001) *Biometrika*, 88, 865–873.
- Loader, C. (1996) *AS*, 24, 1602–1618.
- Park, B., Lee, Y., Kim, T., Park, C. and Eguchi, S. (2006) *JSPI*, (136), 3, 839–859.
- Tibshirani, R. and Hastie, T. (1987) *JASA* 82, 559–567.
- Windham, M. (1995) *JRSS B*, 57, 599–609.

- Basu, A., Harris, I., Hjort, N. and Jones, M.C. (1998) *Biometrika*, 85, 549–559.
- Hjort, N. and Glad, I. (1995) *AS*, 23, 882–904.
- Hjort, N. and Jones, M.C. (1996) *AS*, 24, 1619–1647.
- Huber, P. and Ronchetti, E. (2009) *Robust Statistics*, 2nd Edition. Wiley, New York.
- Jones, M.C., Hjort, N., Harris, I. and Basu, A. (2001) *Biometrika*, 88, 865–873.
- Loader, C. (1996) *AS*, 24, 1602–1618.
- Park, B., Lee, Y., Kim, T., Park, C. and Eguchi, S. (2006) *JSPI*, (136), 3, 839–859.
- Tibshirani, R. and Hastie, T. (1987) *JASA* 82, 559–567.
- Windham, M. (1995) *JRSS B*, 57, 599–609.

- Basu, A., Harris, I., Hjort, N. and Jones, M.C. (1998) *Biometrika*, 85, 549–559.
- Hjort, N. and Glad, I. (1995) *AS*, 23, 882–904.
- Hjort, N. and Jones, M.C. (1996) *AS*, 24, 1619–1647.
- Huber, P. and Ronchetti, E. (2009) *Robust Statistics*, 2nd Edition. Wiley, New York.
- Jones, M.C., Hjort, N., Harris, I. and Basu, A. (2001) *Biometrika*, 88, 865–873.
- Loader, C. (1996) *AS*, 24, 1602–1618.
- Park, B., Lee, Y., Kim, T., Park, C. and Eguchi, S. (2006) *JSPI*, (136), 3, 839–859.
- Tibshirani, R. and Hastie, T. (1987) *JASA* 82, 559–567.
- Windham, M. (1995) *JRSS B*, 57, 599–609.

- Basu, A., Harris, I., Hjort, N. and Jones, M.C. (1998) *Biometrika*, 85, 549–559.
- Hjort, N. and Glad, I. (1995) *AS*, 23, 882–904.
- Hjort, N. and Jones, M.C. (1996) *AS*, 24, 1619–1647.
- Huber, P. and Ronchetti, E. (2009) *Robust Statistics*, 2nd Edition. Wiley, New York.
- Jones, M.C., Hjort, N., Harris, I. and Basu, A. (2001) *Biometrika*, 88, 865–873.
- Loader, C. (1996) *AS*, 24, 1602–1618.
- Park, B., Lee, Y., Kim, T., Park, C. and Eguchi, S. (2006) *JSPI*, (136), 3, 839–859.
- Tibshirani, R. and Hastie, T. (1987) *JASA* 82, 559–567.
- Windham, M. (1995) *JRSS B*, 57, 599–609.