NANZAN-TR-2006-03

The smallest and largest distributions of the absolute difference and their applications to robust scale estimation

Miyoshi Kimura

March 2007

Technical Report of the Nanzan Academic Society
Mathematical Sciences and Information Engineering

The smallest and largest distributions of the absolute difference and their applications to robust scale estimation

Miyoshi Kimura

Nanzan University

Abstract A certain class of symmetric and unimodal continuous distributions is considered. The smallest and largest distributions of the absolute difference of two independent random variables with a common distribution in the class are derived. The results are extended to the case that any distribution in the class may be contaminated. As an application of the results to robust estimation, the implosion and explosion biases of an Q-estimate for scale over a class of contaminated distributions are derived. This contaminated distribution class is related to (c, γ)-contamination.

AMS 2000 Subject classifications: Primary 62F35, Secondary 62G35, 62J05

Key words: symmetric unimodal distribution, smallest and largest distributions, absolute difference, robust scale estimation, Q-estimate, implosion bias, explosion bias, (c, γ)-contamination.

1 Introduction and basic results

For a given symmetric and unimodal continuous function we consider a certain class of symmetric unimodal continuous distributions and derive some basic results (Proposition, Corollary). From these results we obtain the stochastically smallest and largest distributions of the absolute difference of two independent random variables with a common distribution in the class (Theorem 1). Next we consider the case that any distribution in the class may be contaminated and extend the results to a broader class which consists of all the ε-contaminated distributions (Theorem 2). There are various statistics based on the absolute difference and the results have large applicability. As an application of Theorem 2 to robust estimation, we use the contaminated distribution class to descrive the departure from the model distribution and derive the implosion and explosion biases of an Q-estimate for scale (Theorems 3 and 4). The Q-estimate was proposed as an alternative to MAD (median absolute deviation) by Rousseeuw and Croux (1993). We notice that the class of contaminated distributions is related to (c, γ)-contamination introduced by Ando and Kimura (2003). See Ando and Kimura (2004, 2005, 2006) for applications of (c, γ)-contamination.

Let g be nonnegative continuous unimodal function defined on the real line R which is even and satisfies

$$1 \le \int_{-\infty}^{\infty} g(x)dx < \infty. \tag{1.1}$$

For g we define $\mathcal{F}_g$ as the class of all continuous distributions F which has an even unimodal continuous (on the support) density f such that $0 \le f \le g$. In what follows, for any F in $\mathcal{F}_g$ we use its even unimodal continuous density f. Let X and Y be independent and identically distributed with a common F in $\mathcal{F}_g$. There are a lot of important statistics based on the abosolute difference $|X-Y|$ and we are interested in its smallest and largest distributions.

Let $\hat{F}$ and $\bar{F}$ have the densities $\hat{f}$ and $\bar{f}$ defined as follows:

$$\hat{f}(x) = \begin{cases} g(x) & \text{if} \quad |x| \le a, \\ 0 & \text{if} \quad |x| > a, \end{cases} \tag{1.2}$$

$$\bar{f}(x) = \begin{cases} g(b) & \text{if} \quad |x| \le b, \\ g(x) & \text{if} \quad |x| > b, \end{cases} \tag{1.3}$$

where a and b are the positive constants such that $\hat{f}$ and $\bar{f}$ are densities. Note that $\hat{F}$ and $\bar{F}$ belong to $\mathcal{F}_g$.

The following basic results are used to derive main theorems.

Proposition. The following hold:

(i) $$\sup_{F\in\mathcal{F}_g} \int_{-\infty}^{\infty} \{F(x+t)-F(x-t)\}dF(x) = \int_{-\infty}^{\infty} \{\hat{F}(x+t)-\hat{F}(x-t)\}d\hat{F}(x), \quad t \ge 0,$$

(ii) $$\inf_{F\in\mathcal{F}_g} \int_{-\infty}^{\infty} \{F(x+t)-F(x-t)\}dF(x) = \int_{-\infty}^{\infty} \{\bar{F}(x+t)-\bar{F}(x-t)\}d\bar{F}(x), \quad t \ge 0.$$

Proof. (i) Let F be any element of $\mathcal{F}_g$. Since $F(x+t)-F(x-t)$ is nonincreasing in $|x|$ and $f \le \hat{f}$ on $[-a,a]$, it follows that

$$\int_{-\infty}^{\infty} \{F(x+t)-F(x-t)\}f(x)dx \le \int_{-\infty}^{\infty} \{F(x+t)-F(x-t)\}\hat{f}(x)dx. \tag{1.4}$$

Since f and $\hat{f}$ are even, it follows that

$$\int_{-\infty}^{\infty} \{F(x+t)-F(x-t)\}\hat{f}(x)dx = \int_{-\infty}^{\infty} \{F(x+t)-F(-x-t)\}\hat{f}(x)dx.$$

Hence

$$\begin{aligned} &\int_{-\infty}^{\infty} \{F(x+t)-F(x-t)\}\hat{f}(x)dx \\ &= \int_{-\infty}^{-t} \{F(x+t)-F(-x-t)\}\hat{f}(x)dx + \int_{-t}^{\infty} \{F(x+t)-F(-x-t)\}\hat{f}(x)dx \\ &= \int_{-t}^{\infty} \{F(x+t)-F(-x-t)\}\{\hat{f}(x)-\hat{f}(x+2t)\}dx. \end{aligned} \tag{1.5}$$

Since, for $x > -t$ we have $\hat{f}(x) - \hat{f}(x+2t) \geq 0$ and

$$F(x+t) - F(-x-t) \leq \hat{F}(x+t) - \hat{F}(-x-t), \tag{1.6}$$

it follows that

$$\begin{aligned}\int_{-t}^{\infty} \{F(x+t) - F(x-t)\}\{\hat{f}(x) - \hat{f}(x+2t)\}dx \\ \leq \int_{-t}^{\infty} \{F(x+t) - F(-x-t)\}\{\hat{f}(x) - \hat{f}(x+2t)\}dx.\end{aligned}$$

Therefore the assertion (i) follows from (1.4) and (1.5).

(ii) Let F be any element of $\mathcal{F}_g$. Since $f \geq \bar{f}$ on $[-b, b]$, it follows that

$$\int_{-\infty}^{\infty} \{F(x+t) - F(x-t)\}f(x)dx \geq \int_{-\infty}^{\infty} \{F(x+t) - F(x-t)\}\bar{f}(x)dx \tag{1.7}$$

Since $\bar{f}$ is even and unimodal, for $x > -t$ we have $\bar{f}(x) - \bar{f}(x+2t) \geq 0$ and

$$\bar{F}(x+t) - \bar{F}(-x-t) \leq F(x+t) - F(-x-t).$$

Hence it follows that

$$\begin{aligned}\int_{-\infty}^{\infty} \{F(x+t) - F(x-t)\}\bar{f}(x)dx \\ &= \int_{-t}^{\infty} \{F(x+t) - F(-x-t)\}\{\bar{f}(x) - \bar{f}(x+2t)\}dx \\ &\geq \int_{-t}^{\infty} \{\bar{F}(x+t) - \bar{F}(-x-t)\}\{\bar{f}(x) - \bar{f}(x+2t)\}dx \\ &= \int_{-\infty}^{\infty} \{\bar{F}(x+t) - \bar{F}(x-t)\}\bar{f}(x)dx\end{aligned}$$

This and (1.7) imply the assertion (ii). □

Corollary. The following hold:

(i) $$\sup_{F \in \mathcal{F}_g} \int_{-\infty}^{\infty} F(x+t)dF(x) = \int_{-\infty}^{\infty} \hat{F}(x+t)d\hat{F}(x)$$

(ii) $$\inf_{F \in \mathcal{F}_g} \int_{-\infty}^{\infty} F(x+t)dF(x) = \int_{-\infty}^{\infty} \bar{F}(x+t)d\bar{F}(x)$$

Proof. It is easy to see that for any F in $\mathcal{F}_g$

$$\int_{-\infty}^{\infty} F(x-t)dF(x) = 1 - \int_{-\infty}^{\infty} F(x+t)dF(x).$$

This implies that

$$\int_{-\infty}^{\infty} F(x+t)dF(x) = \frac{1}{2}\int_{-\infty}^{\infty}\{F(x+t) - F(x-t)\}dF(x) + \frac{1}{2}.$$

Therefore the corollary follows from Prposition. □

2 Main results

The following useful theorem is readily obtained from Prposition.

Theorem 1. Let X and Y be independent random variables distributed with a common F in $\mathcal{F}_g$. Then the distribution of $|X-Y|$ is stochastically smallest and largest under $\hat{F}$ and $\bar{F}$, respectively, i.e.,

(i) $$\sup_{F\in\mathcal{F}_g} P_{F\times F}(|X-Y| \le t) = P_{\hat{F}\times\hat{F}}(|X-Y| \le t), \quad t \ge 0,$$

(ii) $$\inf_{F\in\mathcal{F}_g} P_{F\times F}(|X-Y| \le t) = P_{\bar{F}\times\bar{F}}(|X-Y| \le t), \quad t \ge 0.$$

Proof. For any F in $\mathcal{F}_g$ we have

$$P_{F\times F}(|X-Y| \le t) = \int_{-\infty}^{\infty}\{F(y+t) - F(y-t)\}dF(y)$$

where f is a density of F. Therefore the theorem follows from Theorem 1. □

Next, we consider the case that X and Y may be contaminated. Define the ε-contaminated class $\mathcal{P}_{g,\varepsilon}$ of $\mathcal{F}_g$ as

(2.1) $$\mathcal{P}_{g,\varepsilon} = \{H = (1-\varepsilon)F + \varepsilon K \ : \ F \in \mathcal{F}_g, \ K \in \mathcal{M}\},$$

where $\mathcal{M}$ is the set of all distributions on R. The following is an extension of Theorem 1.

Theorem 2. Let X and Y be independent random variables distributed with a common H in $\mathcal{P}_{g,\varepsilon}$. Then

(i) $$\sup_{H\in\mathcal{P}_{g,\varepsilon}} P_{H\times H}(|X-Y| \le t)$$
$$= (1-\varepsilon)^2 P_{\hat{F}\times\hat{F}}(|X-Y| \le t) + 2\varepsilon(1-\varepsilon)\hat{F}(|X| \le t) + \varepsilon^2, \quad t \ge 0,$$

(ii) $$\inf_{H\in\mathcal{P}_{g,\varepsilon}} P_{H\times H}(|X-Y| \le t)$$
$$= (1-\varepsilon)^2 P_{\bar{F}\times\bar{F}}(|X-Y| \le t), \quad 0 \le t < \infty,$$

where $\hat{F}$ and $\bar{F}$ are given by (1.2) and (1.3), and Δ_0 is the point mass distribution at 0 .

Proof. (i) For any $H = (1-\varepsilon)F + \epsilon K$ in $\mathcal{P}_{g,\varepsilon}$ we have

$$\begin{aligned} & P_{H\times H}(|X-Y| \le t) \\ & \quad = \ (1-\varepsilon)^2 P_{F\times F}(|X-Y| \le t) + 2\epsilon(1-\varepsilon) P_{F\times K}(|X-Y| \le t) + \varepsilon^2 P_{K\times K}(|X-Y| \le t). \end{aligned}$$

It is easy to see that

$$P_{F\times K}(|X-Y| \le t) \le P_{\hat{F}\times\Delta_0}(|X-Y| \le t) = \hat{F}(|X| \le t)$$

and

$$P_{K\times K}(|X-Y| \le t) \le 1 = P_{\Delta_0\times\Delta_0}(|X-Y| \le t).$$

Therefore the assertion (i) follows from (i) of Theorem 1.

(ii) Let Δ_n be the point mass distribution at n. Then, for any $H = (1-\varepsilon)F + \varepsilon K$ in $\mathcal{P}_{g,\varepsilon}$ we have

$$P_{F\times K}(|X-Y| \le t) \ge \lim_{n\to\infty} P_{\bar{F}\times K_n}(|X-Y| \le t) = 0$$

and

$$P_{K\times K}(|X-Y| \le t) \ge \lim_{n\to\infty} P_{K_n\times K_n}(|X-Y| \le t) = 0.$$

Therefore the assertion (ii) follows from (ii) of Theorem 2 . $\square$.

Remark 1. Let $\mathcal{F}_g^*$ be the set of all continuous densities f such that $f \le g$. The assertions (i) of Proposition, Corollary and Theorem 1 also hold for this broader $\mathcal{F}_g^*$. These results were proved in a different way by Ando and Kimura (2003). In this case, we note that the symmetry and unimodality of f are not required in the definition of $\mathcal{F}_g^*$. Let $\mathcal{P}_{g,\varepsilon}^*$ be the set defined by replacing $\mathcal{F}_g$ with $\mathcal{F}_g^*$ in the definition (2.1) of $\mathcal{P}_{g,\varepsilon}$. We can see that the assertion (i) of Theorem 2 also holds for $\mathcal{P}_{g,\varepsilon}^*$

3 The explosion and implosion bias of an Q-estimate

As an application of the previous results, we consider the explosion and implosion biases of an Q-estimate for robust scale estimation. To give a robust scale estimation model we need to define a certain class of distributions. Let F_0 be a specified distribution with an even unimodal continuous density f_0. For some given constants c and γ ($0 < \gamma < 1$ and $c > 1-\gamma$) we consider the (c,γ)-symmetric unimodal neighborhood $\mathcal{F}_{c,\gamma}(F_0)$ of F_0, defined as the set of all continuous distributions F which has an even unimodal densitiy f and satisfies $f \le (\frac{c}{1-\gamma})f_0$. Further we consider the γ-contamination neighborhood $\mathcal{P}_{c,\gamma}(F_0)$ of $\mathcal{F}_{c,\gamma}(F_0)$, that is

$$\mathcal{P}_{c,\gamma}(F_0) = \{H = (1-\gamma)F + \gamma K \ : \ F \in \mathcal{F}_{c,\gamma}(F_0), \ K \in \mathcal{M}\}. \tag{3.1}$$

The robust scale model we consider here is given as follows: Let $X_1, \cdots, X_n$ be independent and identically distributed random variables with H. We assume that H belongs to the class

$$\mathcal{P}_{c,\gamma}(F_{\mu,s}) = \{H : H(x) = (1-\gamma)F(\frac{x-\mu}{s}) + \gamma K(x),\ x \in R,\ F \in \mathcal{F}_{c,\gamma}(F_0), K \in M\}, \tag{3.2}$$

where μ is an unknown location parameter and $s > 0$ is an unknown scale parameter to be estimated. For more details of the model (3.2), see Martin and Zamar (1993) and Ando and Kimura (2003). Among various robust estimates for scale s, we consider an Q_n-estimate given by the kth order statistic

$$Q_n = d \cdot \{|X_i - X_j|;\ i < j\ \}_{(k)}, \tag{3.3}$$

where d is a constant factor and $k = \binom{h}{2} \approx \binom{n}{2}/4$, where $h = [n/2]+1$. The Q_n was proposed as an alternative to MAD_n by Rousseeuw and Croux (1993) and it has 50% breakdown point and higher efficiency than MAD_n. Since Q_n is location and scale equivariant, we derive the explosion and implosion biases of Q_n over $\mathcal{P}_{c,\gamma}(F_0)$ ($F_0 = F_{0,1}$). The explosion bias $B_Q^+(c,\gamma)$ and implosion bias $B_Q^-(c,\gamma)$ of the Q-estimate T over $\mathcal{P}_{c,\gamma}(F_0)$ are defined by

$$B_Q^+(c,\gamma) = \sup\{Q(H) : H \in \mathcal{P}_{c,\gamma}(F_0)\}, \tag{3.4}$$

$$B_Q^-(c,\gamma) = \inf\{Q(H) : H \in \mathcal{P}_{c,\gamma}(F_0)\}. \tag{3.5}$$

The asymptotic version of Q_n is given by

$$Q(H) = dG_H^{-1}(\frac{1}{4}) = dL_H^{-1}(\frac{5}{8}). \tag{3.6}$$

where G_H and L_H are the distributions of $|X-Y|$ and $X-Y$, respectively. Note that L_H is symmetric about the origin.

As in Section 1, we define the densities $\hat{f}$ and $\bar{f}$ by

$$\hat{f}(x) = \begin{cases} (\frac{c}{1-\gamma})f_0(x) & \text{if} \quad |x| \le a, \\ 0 & \text{if} \quad |x| > a, \end{cases} \tag{3.7}$$

$$\bar{f}(x) = \begin{cases} (\frac{c}{1-\gamma})f_0(b) & \text{if} \quad |x| \le b, \\ (\frac{c}{1-\gamma})f_0(x) & \text{if} \quad |x| > b, \end{cases} \tag{3.8}$$

where a and b are the constants such that

$$a = F_0^{-1}(\frac{c-\gamma+1}{2c}) \qquad \text{and} \qquad F_0(b) - bf_0(b) = \frac{2c+\gamma-1}{2c}.$$

The explosion and implosion biases of the Q-estimate are obtained by next two theorems. Although the thorems can be readily derived using Theorem 2 and G_H, we give the proofs following Rousseeuw and Croux (1993) .

Theorem 3. Let F_0 have an even unimodal continuous density f_0. Then

$$B_Q^+(c,\gamma) = \begin{cases} d(\bar{F}^{*2})^{-1}(\frac{5-8\gamma+4\gamma^2}{8(1-\gamma)^2}) & \text{if} \quad 0 \le \gamma < \frac{1}{2}, \\ 0 & \text{if} \quad \gamma \ge \frac{1}{2}, \end{cases} \tag{3.9}$$

where $\bar{F}^{*2}$ denotes the convolution of $\bar{F}$.

Proof. For any $H = (1-\gamma)F + \gamma K$ in $\mathcal{P}_{c,\gamma}(F_0)$, $Q(H)$ is the smallest solution of

$$\int_{-\infty}^{\infty} H(y + d^{-1}Q(H))dH(y) \ge \frac{5}{8}. \tag{3.10}$$

Let X be distributed with F, and let Y_1 and Y_2 be distributed with K. Then (3.10) is expressed as

$$\begin{aligned}(1-\gamma)^2 F^{*2}(d^{-1}Q(H)) + \gamma(1-\gamma)\{1 + P(|X - Y_1| \le d^{-1}Q(H))\} \\ +\gamma^2 P((Y_1 - Y_2) \le d^{-1}Q(H)) \ge \frac{5}{8}.\end{aligned} \tag{3.11}$$

Each term in (3.11) is increasing in $Q(H)$. Hence, $Q(H)$ is maximized when $F^{*2}(d^{-1}Q(H)), P(|X - Y_1| \le d^{-1}Q(H))$ and $P((Y_1 - Y_2) \le d^{-1}Q(H))$ are minimized. To do this, by Theorem 1 we need to take $F = \bar{F}$. Let $K = K_n$ be the normal distribution $N(n, n^2)$ with mean n and variance n^2. Then, under $\bar{H}_n = (1-\gamma)\bar{F} + \gamma K_n$, as $n \to \infty$ we have $P(|X - Y_1| \le d^{-1}Q(H)) \to 0$ and $P((Y_1 - Y_2) \le d^{-1}Q(H)) \to \frac{1}{2}$. Substituing the lower bounds 0 and $\frac{1}{2}$ of $P(|X - Y_1| \le d^{-1}Q(H))$ and $P((Y_1 - Y_2) \le d^{-1}Q(H))$ into (3.11), we obtain Theorem 3. $\square$.

Theorem 4. Let F_0 have an even unimodal continuous density f_0. Then

$$B_Q^-(c,\gamma) = \begin{cases} Q(\hat{F}) & \text{if} \quad 0 \le \gamma < \frac{1}{2}, \\ 0 & \text{if} \quad \gamma \ge \frac{1}{2}, \end{cases} \tag{3.12}$$

where $Q(\hat{F})$ satisfies the equation

$$(1-\gamma)^2 \hat{F}^{*2}(d^{-1}Q(\hat{F})) + 2\gamma(1-\gamma)\hat{F}(Q(\hat{F})) + \gamma^2 = \frac{5}{8}, \tag{3.13}$$

and $\hat{F}^{*2}$ denotes the convolution of $\hat{F}$.

Proof. By (3.11), $Q(H)$ is minimized when $F^{*2}(d^{-1}Q(H))$, $P(|X - Y_1| \le d^{-1}Q(H))$ and $P((Y_1 - Y_2) \le d^{-1}Q(H))$ are maximized. Hence, by Theorem 1 we choose $F = \hat{F}$ and $K = \Delta_0$, where $K = \Delta_0$ is the poit mass distribution at the zero. Then $P(|X - Y_1| \le d^{-1}Q(\hat{H})) = 2\hat{F}(Q(\hat{H}) - 1$ and $P((Y_1 - Y_2) \le d^{-1}Q(\hat{H})) = 1$. Therefore we obtain Theorem 4. $\square$

Remark 2. Let $\mathcal{P}^*_{c,\gamma}(F_0)$ be the set of defined by replacing $\mathcal{F}_{c,\gamma}$ with $\mathcal{F}^*_{c,\gamma}$, that is

$$\mathcal{P}^*_{c,\gamma}(F_0) = \{H = (1-\gamma)F + \gamma K \ : \ F \in \mathcal{F}^*_{c,\gamma}, \ K \in \mathcal{M}\}.$$

As shown in Ando, Kakiuchi and Kimura (2006), $\mathcal{P}^*_{c,\gamma}(F_0)$ is the (c,γ)-contamination neighborhood of F_0 intorduced by Ando and Kimura (2003). We should notice that Throem 4 holds for $\mathcal{P}^*_{c,\gamma}(F_0)$.

Acknowledgments. The author thanks Dr.Masakazu Ando and Dr.Itsuro Kakiuchi for their helpful discussions. This research was supported by Nanzan Pache Research Subsidy I-A-2, 2006.

References

[1] Ando,M. and Kimura,M.(2003). A characterization of the neighborhoods defined by certain special capacities and their applications to bias-robustness of estimates. *J. Statist. Plann. Inference*, **116**, 61-90.

[2] Ando,M. and Kimura,M. (2004). The maximum asymptotic bias of S-estimates for regression over the neighborhoods defined by certain special capacities. *J. Multivariate Anal.*, **90**, 407-425.

[3] Ando,M. and Kimura,M.(2005). On the maximum asymptotic bias of robust regression estimates over certain contamination neighborhoods. Technical Report No. 2004-07, Nanzan Academic Sosiety, Mathematical Sciences and Information Engineering.

[4] Ando,M., Kakiuchi,I. and Kimura,M.(2006). Robust nonparametric confidence intervals and tests for the median in the presence of (c,γ)-contamination. Technical Report No. 2005-01, Nanzan Academic Sosiety, Mathematical Sciences and Information Engineering.

[5] Martin,R.D. and Zamar,R.(1993). Bias robust estimationof scale. *Ann. Statist.*, **21**, 991-1017.

[6] Rousseeuw, P.J. and Croux,C.(1993). Alternatives to the median absolute deviation. *J. Amer. Statist. Assoc.*, **88**, 1273-1283.

Miyoshi Kimura
Department of Information Systems
and Mathematical Sciences
Nanzan University
27 Seirei-cho, Seto, Aichi, 489-0863
JAPAN.
E-mail: kimura@ms.nanzan-u.ac.jp