A Robust Approach to Blind Source Separation Using Independent Component Analysis (ICA)

Roberto N. Padua and Warren I. Luzano[1]

received November, 2002; revised December, 2002

ABSTRACT

Independent Component Analysis (ICA) is a recent statistical technique developed in the early 1990's (Jutten et al. (1991), Cardoso (1994), Amari (1997), Hyvarinen (2000)) to deal with the problem of separating the independent components of a mixed signal. It has been recognized that the asymptotic behavior of an algorithm for ICA has to depend on the unknown probability density functions of each source. Efficient algorithms exist for fast estimation of an unknown density function and each algorithm requires the computation of score functions of the unknown sources. In such existing ICA algorithms, approximate models of the score functions are used based on moment expressions e.g. kurtosis, which are prone to outliers i.e. non-robust. This paper examines alternative robust procedures for approximating the score functions for ICA applications.

KEYWORDS AND PHRASES: independent component analysis, blind source separation, score function, robust methods, influence functions

1. INTRODUCTION

The simplest and often-used illustration for the blind source separation problem is the cocktail party illustration. Two microphones are placed at different distances from two speakers A and B. A listener observes and records the signals from the microphone A' and B'. Given A' and B', the problem is to recover the original signals A and B.

We note the observed signals A' and B' are each mixtures of the original signals A and B. Mathematically, let S be a vector (s x 1) of original signals, let A be an s x s matrix called a mixing matrix and let X be an s x 1 vector of observed signals. Then:

$$X = AS \tag{1}$$

The signal S and matrix A are both unknown. The blind source separation problem is to estimate the demixing matrix W:

$$W = A^{-1} \tag{2}$$

so that

$$S = WX \tag{3}$$

The matrix A can also be generalized to an m x s matrix where m ≠ s. Section 2 surveys the techniques employed in the literature (since 1995) for solving this blind source separation problem.

[1] Vice President for Research and International Affairs and Research Assistant respectively, Department of Mathematical Sciences, Mindanao Polytechnic State College, Lapasan, Cagayan de Oro City. Email Address: rnpadua@yahoo.com

A major issue in all independent component analysis (ICA) applications is the issue on the robustness of the criteria used to estimate the original source signals. Existing ICA algorithms approximate these criteria based on moment expressions, such as the kurtosis, which are prone to outliers. It has been demonstrated in several studies in the past (Huber, 1981) that outliers can distort the values of estimators as well as their distributions. This paper aims to propose a method for enhancing the insensitivity of moment expressions to outliers by the use of trimming techniques.

While independent component analysis was originally developed to deal with problems allied to the cocktail party problem, it is apparent that it has many other applications. ICA, for instance, was successfully applied to the problem of finding hidden factors in financial data (Oja (1998)); to noise reduction in natural images (Bell and Sejnowski (1997)); and to the problem of reproducing brain activity from electrical recordings of electroencephalograms (EEG) (Gonzales and Wintz (1997)).

This paper is organized as follows: Section 2 is a survey of existing ICA techniques; Section 3 proposes a class of robust estimators for kurtosis and negentropy measures; Section 4 provides the experimental results using the robust estimators proposed, and Section 5 gives the conclusions and recommendations.

2. SURVEY OF EXISTING ICA TECHNIQUES

Assume that we have n linear combinations x_1 , x_2 ,.... x_n of n independent components s_1 , s_2 ,.... s_n:

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \dots\dots\dots + a_{jn}s_n \qquad (4)$$

We assume that the s_i's have zero mean (and, thus, also the x). Equation (4) can be written in vector-matrix form as equation (2). Denote the j_{th} column of A by a_j, then the model can be written as :

$$x = \sum_{j=1}^{n} a_j s_j \qquad (5)$$

In ICA, the components s_i's are assumed to be statistically independent. Further, the distribution of s_i's are assumed to be non-Gaussian in order to ensure estimability of A.

The basic premise to the framework of ICA is the concept of non-Gaussianity. According to the Central Limit Theorem, the sum of properly normalized independent random variables has a distribution that is closer to a normal distribution than the individual summands. Thus, the sum of the independent variables has a distribution closer to a normal distribution than any one of the two variables. This simple observation leads to the general approaches currently being used in ICA. Let

$$y = w^T x = \sum_j w_j x_j \qquad (6)$$

$$z = A^T w \tag{7}$$

Equations (6) and (7) leads to:

$$y = w^T x = w^T As = z^T s \tag{8}$$

It follows that y, which is a linear combination of the unknown independent signals s. Since a sum of two or more independent variables is more Gaussian than any of its components, it follows that y is more Gaussian than any of the s_i's.

Essentially, therefore, the problem reduces to finding an appropriate measure of non-Gaussianity and then maximizing this measure. The classical measure of non-Gaussianity is kurtosis or the fourth-order cumulant defined by:

$$\mathrm{Kurt}(y) = E(y^4) - 3[E(y^2)]^2 \tag{9}$$

If we assume that y has been normalized so that it has a unit variance; Equation (9) simplifies to : $\mathrm{kurt}(y) = E(y^4) - 3$. For a normal random variable, kurt(y) = 0 but for most (although not all) non-Gaussian random variables, kurt(y) 0. Since the kurtosis of a random variable can be positive or negative, the typical measure of non-Gaussianity is :

$$\mathrm{NEG} = |\,\mathrm{kurt}(y)\,| \tag{10}$$

The ICA procedure now consists of maximizing NEG.

Example : Let S = 2 so that ;

$$x_1 = a_{11}s_1 + a_{12}s_2$$
$$x_2 = a_{21}s_1 + a_{22}s_2$$

We seek s_1 and s_2 which maximizes :

$$|\,\mathrm{kurt}(y)\,| = \left| z_1{}^4\,\mathrm{kurt}(s_1) + z_2{}^4\,\mathrm{kurt}(s_2) \right|$$

subject to : $\mathrm{var}(y) = z_1{}^2\,\mathrm{var}(s_1) + z_2{}^2\,\mathrm{var}(s_2)$

These equations can be simplified further by noting that

$$\mathrm{var}(s_1) = \mathrm{var}(s_2) = 1$$

so that : $\mathrm{Max}\,|\,\mathrm{kurt}(y)\,| = \left| z_1{}^4\mathrm{kurt}(s_1) + z_2{}^4\mathrm{kurt}(s_2) \right|$

subject to $z_1{}^2 + z_2{}^2 = 1$

It is easy to show that the solutions are { (1,0), (-1,0), (0,1), (0,-1) }.which corresponds to the fact that the solutions pick out one of the components s_1 or s_2.

A non-Gaussianity measure which is more justified in statistical theory is negentropy or negative entropy. Let H(y) be the entropy of a random variable y, then:

$$\begin{aligned} H(y) &= -\int f(y)\log f(y)\,dy, && \text{if y is continous} \\ &= -\sum P(Y = a_i)\log P(y = a_i), && \text{if y is discrete.} \end{aligned} \qquad (10)$$

The more unpredictable the random variable is, the larger is its entropy. The Gaussian variable has the largest entropy among all random variables of equal variance. Thus, a non-Gaussianity measure that is zero for Gaussian random variable and always non-negative is:

$$J(y) \cong H(y_{\text{gaussian}}) - H(y) \qquad (11)$$

where $H(y_{\text{gaussian}})$ denotes the entropy of a Gaussian random variable. Estimating negentropy using (11) would require an estimate of the pdf of y. It will be more expedient to find approximations to J(y) than to find, possibly, non-parametric estimates of f(y).One such estimate is given by:

$$J(y) \cong \frac{1}{12}E(y^3)^2 + \frac{1}{48}\text{kurt}(y)^2 \qquad (12)$$

Equation (12) suffers for the same non-robustness properties inherent with kurtosis.

A class of estimators proposed by Hyvarinen (1998) based on the maximum-entropy principle is given by:

$$J(y) \cong \sum_{i=1}^{p} k_i \left\{ E(G_i(y)) - E(G_i(v)) \right\}^2 \qquad (13)$$

where k_i's are constants, v is a normalized Gaussian variable and the functions G_i's are non-quadratic functions.

Equation (13) can be used to construct a measure of non-gaussianity that : a) is zero for a Gaussian random variable and (b) always non-negative. In particular, more robust estimates can be obtained by using:

$$\begin{aligned} G_1(u) &= \frac{1}{a_1}\log\cosh a_1 u \qquad \text{or} \\ G_2(u) &= -\exp\left(-u^2/2\right) \end{aligned} \qquad (14)$$

where $1 \le a_1 \le 2$. If $G(y) = y^4$ is used, then we are led back to the kurtosis-based contrast function.

3.A CLASS OF ROBUST ESTIMATORS FOR ICA

We propose to investigate the properties of two "robustified" versions of the kurtosis and negentropy measures in this section. The robustified versions are:

$$\mathrm{kurt}(y) = \left[\frac{1}{1-2\beta}\int_{F^{-1}(\beta)}^{F^{-1}(1-\beta)} \left(y^4\right) dF(y)\right] - 3 \qquad (15)$$

and

$$J(y)_\beta = \frac{1}{12}\left[\frac{1}{(1-2\beta}\int_{F^{-1}(\beta)}^{F^{-1}(1-\beta)} \left(y^3\right) dF(y)\right]^2 + \frac{1}{48}\mathrm{kurt}(y)_\beta{}^2 \qquad (16)$$

which correspond to trimming 100%x 2β extreme observations and then averaging the respective 3rd or 4th powers of the remaining observations. The sample counterparts of (15 and (16) are:

$$\mathrm{kurt}(y) = \left[\frac{1}{1-2\beta}\sum_{y=Y[N\beta]}^{Y_{N[1-\beta]}} \left(y^4\right)\right] - 3$$

and

$$J(y)_\beta = \frac{1}{12}\left[\frac{1}{n(1-2\beta)}\sum_{y=Y_{[N\beta]}}^{Y_{N[1-\beta]}} \left(y^3\right)\right]^2 + \frac{1}{48}\mathrm{kurt}(y)_\beta{}^2$$

The influence function of the functional in Equation (15) is given by: (see Appendix for derivation)

$$\begin{aligned}(1-2\beta)\mathrm{IF}_{T_{2\beta}}(y) &= Y_\beta{}^4 - \omega_{2\beta}(F) \quad , \quad Y < Y_\beta \\ &= Y^4 - \omega_{2\beta}(F) \quad , \quad Y_\beta \le Y \le Y_{1-\beta} \\ &= Y_{1-\beta}{}^4 - \omega_{2\beta}(F) \quad , \quad Y_{1-\beta} < Y\end{aligned}$$

where:

$$\begin{aligned}\omega_{2\beta}(F) &= (1-2\beta)T_{2\beta}(F) + \beta Y_\beta{}^4 + \beta Y_{1-\beta}{}^4 \; ; \\ Y_\beta &= \beta^{th} \text{ quantile of } y \; ; \\ Y_{1-\beta} &= (1-\beta)^{th} \text{ quantile of } y \; ;\end{aligned}$$

and

$$T_{2\beta}(F) = \frac{1}{1-2\beta}\int_{Y_\beta}^{Y_{1-\beta}} \left(y^4\right) dF(y)$$

The influence function of equation (16) will have the same form as (17) except that the middle part will have contributions due to the trimmed third moment. The influence function shows that the estimators will remain bounded (robust) and in fact,

$$\sqrt{n}\left(T_{2\beta}(F_n) - T_{2\beta}(F)\right) \xrightarrow{d} N(0, V) \qquad \text{where}$$
$$V = E\left(IF^2_{\beta,F}(x)\right) \qquad \text{(Stiegler, 1969)}$$

This result also shows that Estimators (15) and (16) are strongly consistent (converges in probability to $T_{2\beta}$ and to $J(y)_{\beta}$.).

In the case of a location parameter, it is known that the trimmed mean has a breakdown point of $\varepsilon^* = \beta$. Since the original observations, in the present application are ordered and are the basis for the computation of the trimmed kurtosis and trimmed negentropy measures, it follows that both statistical measures will inherit the same breakdown point. Thus, these induced estimators will be insensitive to outliers provided that no more than 100β% are on either side of the sample.

Apart from robustness considerations, these induced estimators will also display strong nonparametric efficiency property, namely, their asymptotic efficiency relative to their untrimmed counterparts never drops below $(1\text{-}2\beta)^2$. (Staudte, p.105, 1990).

Instead of fixing the trimming proportion β *ab initio*, one can also use an adoptive trimming procedure as follows: From an observed data set calculate the sample median $(\tilde{x})$ and a measure of dispersion, say the average median absolute deviation (MAD). Form the interval $\left(\tilde{x} \pm 2\text{MAD}\right)$. Trim off all observations falling outside the interval. Such trimming procedures had been extensively studied in Tukey's (1972) Princeton Robustness study.

4. EXPERIMENTAL RESULTS

We mixed the original signals utilizing the mixing matrix A below:

$$A = \begin{bmatrix} -0.3210 & -1.2316 & 0.9442 & -1.0181 \\ 1.2366 & 1.0556 & -2.1204 & -0.1824 \\ -0.6313 & -0.1132 & -0.6447 & 1.5210 \\ -2.3532 & 0.3792 & -0.7043 & -0.0384 \end{bmatrix}$$

4.1 Source Signal Generation

The source signals were generated utilizing the hyperbolic sine function. Specifically, to generate super-Gaussian signals, we take the hyperbolic sine of normally-distributed random numbers. To generate sub-Gaussian signals, we take the inverse-hyperbolic sine of these random numbers.

We generated a total of n=300 source signals. In this data set, we contaminated the true source signals by signals coming from the Cauchy distribution. About 10% of the source signals come from the Cauchy distribution. The algorithm for doing so is given below:

Algorithm:

(1.) Generate u from N(0,1).
(2.) For super-Gaussian signals, take the hyperbolic sine of u.
For sub-Gaussian signals, take the inverse hyperbolic sine of u
(3.) Repeat (1) and (2) for n=270 times.
(4.) Generate Cauchy-distributed signals 30 times (10% of size 300).
(5.) Randomly put it to the original source signals (now total sample size is 300)

4.2 Mixing

Let $(s_1(t), s_2(t), s_3(t), s_4(t)) = S(t)$ be the source signals generated from (4.1), we obtained our observed signals by performing the following matrix multiplication:

$$X(t) = A\,S(t).$$

Thus, we also generated 300 observed signals and are pre-whitened in the succeeding steps.

The data are pre-whitened as follows:

4.3 Pre-Whitening

1. Trimming. We arranged the data from lowest to highest and trimmed off the 5% lowest and 5% highest observations. The remaining observations will be subjected to further treatment.

2. Centering the data. We subtracted the mean $m = E(x)$ from each of the data points so as to make a zero-mean variable. Let $\tilde{x} = x - m$.

3. Whitening. We made the covariance matrix of the new standardized variables equal to I, i.e. $E\left(\tilde{x}\tilde{x}^T\right)=I$. Let $E\left(xx^T\right)=S$. By the eigenvalue decomposition procedure, we obtain:

$$S=PDP^T$$

where P is an orthogonal matrix of eigenvector of S and $D=\text{diag}\left(\lambda_1,...,\lambda_n\right)$. Let $\tilde{X}=PD^{-\frac{1}{2}}P^TX'$.

The fast ICA algorithm of Hyvarinen (2000) is used for actual processing of the whitened data set. For the estimation of one independent component, the algorithm is given as follows:

(a.) Choose an initial (e.g. random) weight vector w.

(b.) Let $w^T=E\left(x\,g\left(w^Tx\right)\right)-E\left(g'\left(w^Tx\right)w\right)$.

(c.) Let $w_{new}=\dfrac{w^+}{\left\|w^+\right\|}$.

(d.) If $w_{old}\bullet w_{new}\approx 1$ stop. Else go to (b).

To estimate several independent components, we need to run the one-unit fast ICA algorithm using several units with weight vectors $w_1, w_2, ..., w_n$. We decorrelate the outputs $w_1{}^Tx, w_2{}^Tx, ..., w_n{}^Tx$ by the Gram-Schmidt process after each iteration. Specifically, let w_p be the output vector in the p^th iteration, then:

(a.) Let $w_{p+1}=w_{p+1}-\sum_{j=1}^{p} w_{p+j}{}^T w_j w_j$.

(b.) $w_{p+1}=\dfrac{w_{p+1}}{\sqrt{w_{p+1}{}^T\,w_{p+1}}}$.

A MATLAB implementation of the fast ICA algorithm is utilized from the World Wide Web.[2]

[2] *WWW address: http://www.cis.hut.fi/projects/ica/fastica/

4.4 Performance Criterion

The performance criterion used in this study is error measure proposed by Amari et. al. (1996).

$$E = \sum_i \left[\sum_j \frac{|p_{ij}|}{\max_k |p_{ik}|} - 1 \right] + \sum_j \left[\sum_i \frac{|p_{ij}|}{\max_k |p_{kj}|} - 1 \right]$$

where Pij is the (i, j) element of the global matrix **P = WA. P** is close to a permutation of the scaled identity matrix . A value of zero indicates perfect match between the found solution and the sources. The greater the value of the above measure, the poorer the performance.

4.5 Results

Fig. 1 shows the hypothetical observed (and pre-whitened) signals. Note that the sample size have been reduced to 270 sample size, as a result from trimming done in **4.3.** These signals were fed to the FastICA program.

Fig. 1a is the product of the first column of the mixing matrix **A** and the first row of the source matrix S (the 1st original signal). Fig. 1b is the product of the second column of the mixing matrix **A** and the second row of the source signal matrix (the 2nd original signal). The remaining figures (1c and 1d) are the products of the remaining 3rd and 4th column of **A** and the corresponding 3rd and 4th rows of the source matrix S.

Fig. 2 shows the output form FastICA. We note that the performance index in this experiment is zero. This means that the global matrix G=AW is a permutation matrix, and so

$$\mathrm{Y} = \mathbf{WX} = (\mathbf{WA})\,\mathbf{S} = \mathbf{I}\,\mathbf{S} = \mathbf{S},$$

which shows perfect separation.

FIGURE 1
(The original signals)

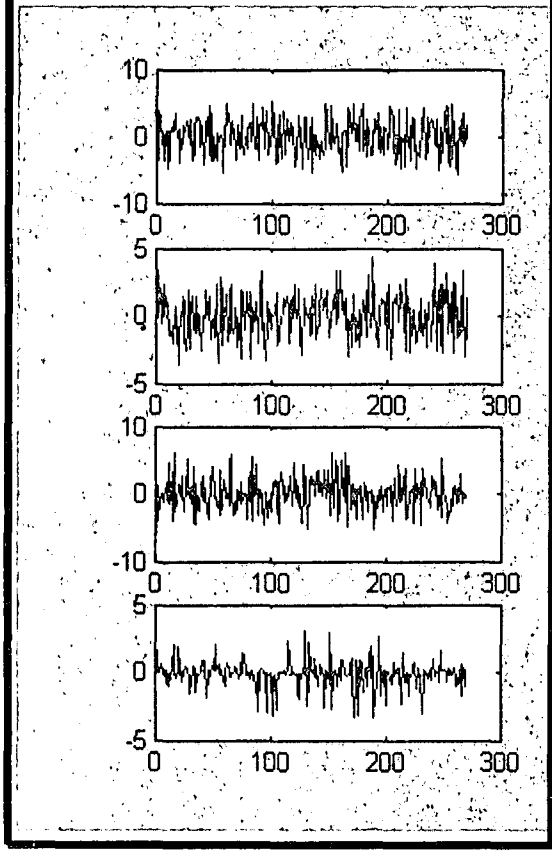
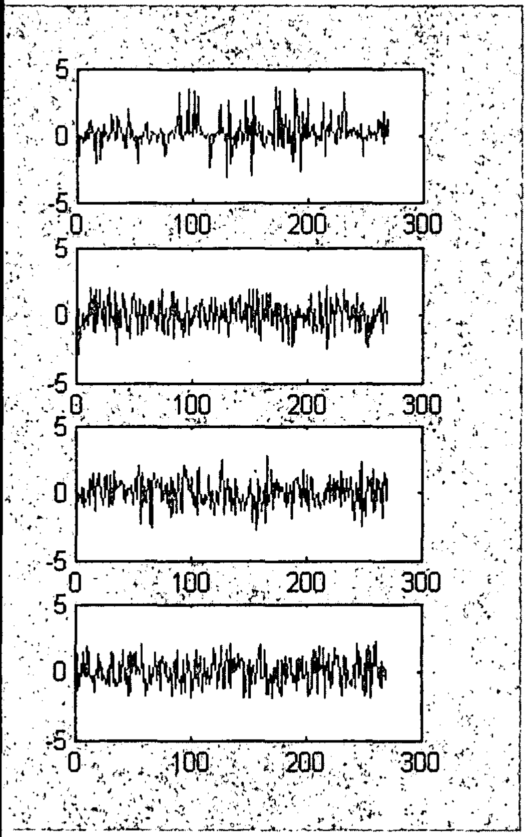


FIGURE 2
(The recovered signals)



ACKNOWLEDGMENT

The author would like to thank the editors of The Philippine Statistician and the anonymous referee for their helpful and constructive comments and suggestions.

References

S. AMARI, S. CHOI AND A. CICHOCKI. "Flexible Independent Component Analysis" *Journal of VLSI Signal Processing*, 10: pp. 1-17, 2000

A. BELL AND T. SEJNOWSKI. "An Information Maximization Approach to Blind Separation and Blind Deconvolution" *Neural Computation*, 7:1129, 1995

A. BENVÉNISTE, M. METIVIER AND P. PRIOCURET. *Adaptive Algorithm and Stochastic Approximations.* Springer-Verlag, 1990

J.F. CARDOSO AND A. SOULOUMIAC. "Blind Beamforming for Non-Gaussian Signals" *IEEE Proceedings_F*, 140(b): 362-370, 1993

S. CHOI, "Adaptive Blind Signal Separation Algorithm" *Neural Networks for Signal Processing*, 8: pp. 13-22, 1998

A. HYVARINEN AND E. OJA. "Independent Component Analysis: Algorithm and Applications" *Neural Networks*, 13:pp. 411-430, 2000

HUBER, P.J. "Minimax Aspects of Bounded Influence Regression" *Journal of the American Statistical Association*, 78: pp. 68-70, 1983

A. MANSOUR, C. JUTTEN *et. al.* "Kurtosis: Definition and Properties" *FUSION '98, Proceedings*, pp. 40-46, 1998

R. STAUDTE AND S. SHEATHER. *Robust Estimation and Testing.* Wiley Series, 1990

R. STIEGLER. "L-Estimation of Location Parameter: Asymptotics and Monte Carlo" *Journal of the American Statistical Association*, 1969

Appendix 1. Derivation of the Influence Function of the Trimmed Kurtosis

The derivation presented here applies equally well to trimmed moment estimates of the moments of our unknown probability distribution. Let F be the unknown probability distribution of a random variable x. Let:

$$\begin{aligned} x_{1-\beta} &= F^{-1}(1-\beta) \\ \mu_\beta &= T_\beta(F) = \frac{1}{1-\beta}\int_0^{x(1-\beta)} \left(y^4\right) dF(y) \end{aligned}$$

We derive the influence function of a one-sided trimmed estimator. Let

$$F_{x,\varepsilon}(y) = (1-\varepsilon)\,F(y) + \varepsilon\,\Delta_x(y)$$

where $0<\varepsilon<1$ and $\Delta_x(\bullet)$ is the probability distribution that puts a mass of one (1) at x. Let :

$$\begin{aligned} g(\varepsilon) &= T_\beta\,(F_{x,\varepsilon}) = \frac{1}{1-\beta}\int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} \left(y^4\right) d\,F(y) \\ &= \left(\frac{1}{1-\beta}\int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} \left(y^4\right) d\,F(y)\right) + \left(\varepsilon\int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} \left(y^4\right) d\,(\Delta_x - F)(y)\right) \end{aligned}$$

The derivative with respect to ε of this quantity is:

$$\begin{aligned} g'(\varepsilon) &= \frac{F_{x,\varepsilon}^{-1}(1-\beta)}{1-\beta}\, f\!\left(F_{x,\varepsilon}^{-1}(1-\beta)\right)\frac{\partial}{\partial\varepsilon}\left[F_{x,\varepsilon}^{-1}(1-\beta)\right] \\ &\quad + \int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} \left(y^4\right) d\,(\Delta_x - F)(y) \\ &\quad + \varepsilon\frac{\partial}{\partial\varepsilon}\int_0^{F_{x,\varepsilon}^{-1}(1-\beta)} \left(\frac{y^4}{1-\beta}\right) d\,(\Delta_x - F)(y) \end{aligned}$$

The influence function $\mathrm{IF}_{T_{\beta,F}}(x) = \lim_{\varepsilon \to 0} g'(\varepsilon)$ so:

$$\mathrm{IF}_{T_{\beta,F}}(x) = \left(\frac{F_{x,\varepsilon}{}^{-1}(1-\beta)}{1-\beta} f(x_{1-\beta})\, \mathrm{IF}_{1-\beta}(x) \right) + \left(\int_0^{F^{-1}(1-\beta)} \left(y^4\right) d\,(\Delta_x - F)(y) \right)$$

$$= \left(\frac{x_{1-\beta}}{1-\beta} f(x_{1-\beta})\, \mathrm{IF}_{1-\beta}(x) \right) + \left(\frac{x^4}{1-\beta} I\{x \le x_{1-\beta}\} \right)$$

Substituting the influence function of the $(1-\beta)^{th}$ quantile yields:

$$\mathrm{IF}_{T_{\beta,F}}(x) = \begin{cases} \dfrac{x^4 - \beta x_{1-\beta}{}^4}{1-\beta} , & 0 \le x \le x_{1-\beta} \\ x_{1-\beta}{}^4 - \mu_\beta , & x_{1-\beta} < x \end{cases}$$

Appendix 2 Asymptotic Normality Derivation

Let $G = \Delta_x - F$ represents a distribution close to F (in the sense of the supremum distance). Expand the functional T(F) in an Edgeworth-like expression as:

$$T(G) = T(F) + \int IF_{T,F}(x)\, d\,(G - F)\,(x) + R \qquad \text{(a)}$$

where R is a remainder term. Consider now the sample counterpart of Equation (a):

$$T(F_N) = T(F) + \int IF_{T,F}(x)\, d\,(F_N - F)\,(x) + R_N \qquad \text{(b)}$$

where we assume that $n^{1/\ 2}R_N \xrightarrow{p} 0$ in probability. Thus,

$$\left(\sqrt{n}\left[(T(F_N)) - T(F)\right] - \int IF_{T,F}(x)\, d\,(F_N - F)\right) \longrightarrow 0 \text{ in probability}$$

or

$$\sqrt{n}\left[(T(F_N)) - T(F)\right] \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF_{T,F}(x_i) \qquad \text{(c)}$$

The right-hand side of © is a sum of iid random variables for which the Central Limit Theorem, so:

$$\left(\sqrt{n}\left[(T(F_N)) - T(F)\right]\right) \xrightarrow{d} N\,(0, V) \qquad \text{(d)}$$

where:

$$V = E\left(IF_{T,F}(x)^2\right).$$

The result of the trimmed-kurtosis can now be easily applied to Equation (d).