İstatistik, Journal of the Turkish Statistical Association
2000, Vol. 3, No 1-2, pp. 1-12

# PREDICTION INTERVALS FOR A CHI DISTRIBUTION WITH A SCALE PARAMETER

Eisuke Hida
Institute of Mathematics
University of Tsukuba
Ibaraki 305-8571
Japan

**Abstract**

Suppose that $X_1, \dots, X_n$ are independent and identically distributed observable random variables with a chi-distribution $\chi_k(\theta)$ with $k$ degrees of freedom and a scale parameter $\sqrt{\theta}$. Let $Y$ be the unobserved random variable, with the same distribution $\chi_k(\theta)$, which is independent of the observable random vector $\mathbf{X} := (X_1, \dots, X_n)$. From the non-Bayesian and Bayesian points of view we construct a prediction interval of $Y$ based on $\mathbf{X}$. The coverage probabilities of the prediction intervals are numerically compared.

**Key Words**: Prediction interval, chi-distribution, non-Bayesian approach, Bayesian approach.

## 1. Introduction

In the theory of statistical prediction, various predictive procedures for unobserved random variable based on the observed data are considered (see, *e.g.* Guttman (1970), Takeuchi (1975), Akahira (1990), Geisser (1993), Takada (1996)). Recently Howlader and Hossain (1998) discussed the Bayesian prediction interval for the Maxwell distribution, that is, the chi-distribution with 3 degrees of freedom and a scale parameter, which was known as the distribution of the speed of a gas molecule. Suppose that $\mathbf{X} := (X_1, \dots, X_n)$ is an observable random vector, $Y$ is a random variable to be observed in future, and the joint distribution of $(\mathbf{X}, Y)$ depends on an unknown parameter $\theta$ in $\Omega$, where $\Omega$ is a parameter space. Let $\mathcal{Y}$ be a space representing the possible outcomes of $Y$. If for any $\alpha$ $(0 < \alpha < 1)$ there exists a subset $S_{\mathbf{X}}$ ( of $\mathcal{Y}$) based on $\mathbf{X}$ such that

$$P_\theta\{Y \in S_{\mathbf{X}}\} \geq 1 - \alpha, \qquad \text{for all } \theta \in \Omega,$$

then $S_{\mathbf{X}}$ is called a prediction region of $Y$ at confidence coefficient $1 - \alpha$. If $\mathcal{Y}$ is a subset of $\mathbf{R}^1$ and $S_{\mathbf{X}}$ is an interval $[a(\mathbf{X}), b(\mathbf{X})]$, then $S_{\mathbf{X}}$ is called a prediction interval of $Y$ at confidence coefficient $1-\alpha$.

Let $X_1, \dots, X_n$ be observable random variables with the chi-distribution $\chi_k(\theta)$ with $k$ degrees of freedom and an unknown scale parameter $\sqrt{\theta}$. Let $Y$ be the unobserved random variable, with the same distribution $\chi_k(\theta)$, which is independent of the random vector $\mathbf{X} = (X_1, \dots, X_n)$. We construct a prediction interval of $Y$ based on the complete sufficient statistic from the non-Bayesian and Bayesian points of view. We also compare the coverage probabilities of the prediction intervals numerically. It is noted that the Bayesian prediction interval obtained by Howlader and Hossain (1998) for the Maxwell distribution, *i.e.* the chi-distribution $\chi_3(\theta)$, is a special case of our results.

## 2. Non-Bayesian approach

First, the probability density function (or p.d.f. for short) of the chi-distribution $\chi_k(\theta)$ of $X_1$ is given by

$$f(x;\theta) = \frac{2}{\Gamma(\frac{k}{2})}\theta^{-\frac{k}{2}} e^{-\frac{x^2}{\theta}} x^{k-1} \quad \text{for } \ 0 < x < \infty, \tag{2.1}$$

where $k(> 2)$ is a known positive number and $\theta$ is an unknown positive number (see, *e.g.* Johnson, Kotz and Balakrishnan (1994), (1995)). Then we construct a prediction interval of $X_{n+1}$ based on $\mathbf{X}$ at confidence coefficient $1-\alpha$. Since the joint p.d.f. of $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(\boldsymbol{x};\theta) = \frac{2^n}{\{\Gamma(\frac{k}{2})\}^n}\theta^{-\frac{nk}{2}} \left\{ \exp\left( -\frac{1}{\theta}\sum_{i=1}^{n} x_i^2 \right) \right\} \prod_{i=1}^{n} x_i^{k-1}$$

the statistic $S := \sum_{i=1}^n X_i^2$ is complete and sufficient for $\theta$. Then $S$ is distributed according to the chi-square distribution $\chi^2_{nk}(\theta)$, whose p.d.f. is given by

$$f_S(s;\theta) = \frac{1}{\Gamma(\frac{nk}{2})}\theta^{-\frac{nk}{2}} e^{-\frac{s}{\theta}} s^{\frac{nk}{2}-1} \quad \text{for } \ 0 < s < \infty.$$

Let $Y := X_{n+1}$. Since $S$ and $X_{n+1}$ are independent, the joint p.d.f. of $(S, Y)$ is given by

$$f_{S,Y}(s, y;\theta) = \frac{2}{\Gamma(\frac{k}{2})\Gamma(\frac{nk}{2})}\theta^{-(\frac{k}{2}+\frac{nk}{2})} e^{-\frac{1}{\theta}(s+y^2)} s^{\frac{nk}{2}-1} y^{k-1}.$$

Hence the statistic $T := S + Y^2$ is complete and sufficient for $\theta$, and $T$ is distributed according to the chi-square distribution $\chi^2_{(n+1)k}(\theta)$ with $(n+1)k$ degrees of freedom and a scale parameter $\theta$. Then the conditional p.d.f. of $Y$ given $T = t$ is

$$f_{Y|T}(y|t) = \frac{f_{T,Y}(t,y;\theta)}{f_T(t;\theta)} \tag{2.2}$$
$$= \frac{2}{B(\frac{k}{2},\frac{nk}{2})}\left(\frac{y^2}{t}\right)^{(k-1)/2}\left(1-\frac{y^2}{t}\right)^{(nk/2)-1} t^{-1/2}$$

for $t > 0$ and $y > 0$, which is independent of $\theta$. This means that the prediction interval of $Y$ based on the sufficient statistic $T$ can be constructed independently of $\theta$. Let $Z := Y^2/T$. Then it is shown from (2.2) that the conditional distribution of $Z$ given $T = t$ is the beta distribution $Be(k/2, nk/2)$ with the p.d.f.

$$g_{Z|T}(z|t) = \frac{1}{B(\frac{k}{2},\frac{nk}{2})} z^{(k/2)-1}(1-z)^{(nk/2)-1} \quad \text{for} \quad 0 \le z \le 1, \tag{2.3}$$

which is independent of $t$, where $B(\cdot,\cdot)$ is the beta function. So, we simply denote $g_{Z|T}(z|t)$ by $g_Z(z)$. Hence any $\alpha$ $(0 < \alpha < 1)$, there exist $z_1$ and $z_2$ such that

$$P\{z_1 \le Z \le z_2\} = 1-\alpha.$$

Then we can obtain a prediction interval of $Y$ at confidence coefficient $1-\alpha$ as follows. For $k > 2$, the conditional p.d.f. $g_Z(z)$ is unimodal, we also take $(z_1, z_2)$ such that

$$\int_{z_1}^{z_2} g_Z(z)dz = 1-\alpha \tag{2.4}$$

and

$$g_Z(z_1) = g_Z(z_2) \tag{2.5}$$

simultaneously. From (2.4) we have

$$I_{z_2}\left(\frac{k}{2},\frac{nk}{2}\right) - I_{z_1}\left(\frac{k}{2},\frac{nk}{2}\right) = 1-\alpha, \tag{2.6}$$

where the incomplete beta function ratio

$$I_p(a,b) = \frac{1}{B(a,b)}\int_0^p z^{a-1}(1-z)^{b-1}dz \tag{2.7}$$

for $a > 0$ and $b > 0$. From (2.5) we have

$$\left(\frac{z_1}{z_2}\right)^{(k/2)-1} = \left(\frac{1-z_2}{1-z_1}\right)^{(nk/2)-1} \tag{2.8}$$

Letting $(z_1, z_2)$ be the simultaneous solution of (2.6) and (2.8). Since $Z = Y^2/T$ and $T = S + Y^2$, it follows that

$$P_\theta\left\{\sqrt{\frac{z_1 S}{1-z_1}} \le Y \le \sqrt{\frac{z_2 S}{1-z_2}}\right\} = 1-\alpha$$

for all $\theta$. Hence

$$\left[\sqrt{z_1 S/(1-z_1)},\ \ \sqrt{z_2 S/(1-z_2)}\right] \tag{2.9}$$

is a prediction interval of $Y$ based on the complete sufficient statistic $S$ through the beta distribution (2.3) at confidence coefficient $1-\alpha$.

## 3. Bayesian approach

For the Maxwell distribution, *i.e.*, the chi-distribution $\chi_3(\theta)$ with 3 degrees of freedom, Howlader and Hossain (1998) obtained the Bayesian prediction interval. In this section, in a similar way to Howlader and Hossain (1998), we construct the highest posterior density (HPD) prediction interval for the chi-distribution $\chi_k(\theta)$ with $k$ degrees of freedom. From (2.1) it follows that, given $\mathbf{X} = \boldsymbol{x}$, the likelihood function is

$$\begin{aligned} L(\boldsymbol{x};\theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &\propto \theta^{-nk/2} e^{-S/\theta}, \end{aligned}$$

where $S := \sum_{i=1}^{n} x_i^2$. Then we consider an asymptotically locally invariant prior, $p_H(\theta)$ proposed by Hartigan (1964), which can be derived from distributions satisfying $(d/d\theta)\log p_H(\theta) = -E[l_1 l_2]/E[l_2]$ where $l_i := (d^i/d\theta^i)\log f(x;\theta), (i=1,2)$, and if $E[l_1] = 0$, $E[l_1^2] + E[l_2] = 0$. In this case, the Hartigan prior can be shown to be $p_H(\theta) \propto \theta^{-2}$. Combining the likelihood function and the prior, the posterior p.d.f. of $\theta$ is

$$h_{\Theta|\mathbf{X}}(\theta|\boldsymbol{x}) = K\theta^{-(m+1)} e^{-S/\theta} \qquad \text{for} \quad \theta > 0,$$

where the normalizing constant $K = S^m/\Gamma(m)$ with $m := (nk+2)/2$. Then the conditional joint p.d.f. of an unobservable random variable $Y = X_{n+1}$ and $\theta$, given $\mathbf{X} = \boldsymbol{x}$, is

$$\begin{aligned} f_{Y,\Theta|\mathbf{X}}(y,\theta|\boldsymbol{x}) &= f_{Y|\Theta,\mathbf{X}}(y|\theta,\boldsymbol{x}) h_{\Theta|\mathbf{X}}(\theta|\boldsymbol{x}) \\ &= f_{Y|\Theta}(y|\theta) h_{\Theta|\mathbf{X}}(\theta|\boldsymbol{x}), \end{aligned}$$

since $Y$ is independent of $\mathbf{X}$. Thus, the predictive p.d.f. of $Y$ is

$$\begin{aligned} f_{Y|\mathbf{X}}(y|\boldsymbol{x}) &= \int_0^\infty f_{Y,\Theta|\mathbf{X}}(y,\theta|\boldsymbol{x}) d\theta \\ &= \frac{2S^m}{B(k/2,m)} \frac{y^{k-1}}{(y^2+S)^{m+(k/2)}} \end{aligned} \tag{3.1}$$

for $y > 0$, which is independent of $\boldsymbol{x}$. Since the predictive p.d.f. of $Y$ is unimodal, the HPD prediction interval $(h_1, h_2)$ of $Y$ at confidence coefficient $1-\alpha$ follows from the simultaneous solutions of the equations

$$P\{h_1 \le Y \le h_2\} = 1-\alpha, \tag{3.2}$$

$$f_{Y|\mathbf{X}}(h_1|\boldsymbol{x}) = f_{Y|\mathbf{X}}(h_2|\boldsymbol{x}). \tag{3.3}$$

Letting $v := y^2$ in (3.1), the equation (3.2) can be written as

$$\begin{aligned}&\int_{h_1^2}^{\infty} \frac{S^{-m}}{B(k/2,m)} v^{k/2-1}\left(1+\frac{v}{S}\right)^{-(m+(k/2))} dv \\ &- \int_{h_2^2}^{\infty} \frac{S^{-m}}{B(k/2,m)} v^{k/2-1}\left(1+\frac{v}{S}\right)^{-(m+(k/2))} dv = 1-\alpha.\end{aligned} \tag{3.4}$$

Letting $1/w := 1 + (v/S)$ in (3.4), we have

$$I_{p_1}(m, k/2) - I_{p_2}(m, k/2) = 1-\alpha \tag{3.5}$$

where $p_i = S/(h_i^2+S)$ and $I_{p_i}(k,l)$, as defined in (2.7), $i = 1, 2$. Also from (3.3), we have

$$\left(\frac{h_1}{h_2}\right)^{k-1} = \left(\frac{h_1^2+S}{h_2^2+S}\right)^{m+(k/2)} \tag{3.6}$$

Thus the HPD prediction interval

$$[h_1, \quad h_2] \tag{3.7}$$

of $Y$ at confidence coefficient $1-\alpha$ follows from the simultaneous solutions $h_1$ and $h_2$ of (3.5) and (3.6). And especially, if $k = 3$, then we obtain a prediction interval of Howlader and Hossain (1998).

Next, if we use a Jeffreys' prior $p_J(\theta) \propto \theta^{-1}$ which is proportional to the square root of Fisher's information and an improper uniform prior $p_U(\theta) = c$ $(0 < c < \infty)$, then we obtain the HPD prediction intervals, where $m = nk/2$ and $m = (nk-2)/2$, respectively.

## 4. Numerical evaluation

In this section we consider the following prediction intervals of $Y$ based on the complete sufficient statistic $S$:

($N_b$) The non-Bayesian prediction interval (2.9) through the beta distribution (2.3).

($B_H$) The Bayesian prediction interval (3.7) through $p_H(\theta)$.

($B_J$) The Bayesian prediction interval (3.7) through $p_J(\theta)$.

($B_U$) The Bayesian prediction interval (3.7) through $p_U(\theta)$.

For $k = 3$; $n = 10(5)30, 50, 100$; $\alpha = 0.05$; $\theta = 2$, we obtain the prediction intervals (see the Table 1). As is seen in the Table 1, the Bayesian prediction intervals ($B_H$) and ($B_J$), ($B_U$) are shorter than the non-Bayesian prediction intervals ($N_b$). A similar tendency to Table 1 is also seen for other values of $k$ and $\theta$.

| $n$ | $S$ | ($N_b$) | ($B_H$) | ($B_J$) | ($B_U$) |
|---|---|---|---|---|---|
| 10 | 30.1453 | [0.0722, 2.9690] | [0.3261, 2.9641] | [0.3350, 3.0717] | [0.3446, 3.1918] |
| 15 | 40.1561 | [0.0627, 2.7442] | [0.3185, 2.7780] | [0.3248, 2.8433] | [0.3314, 2.9135] |
| 20 | 60.9504 | [0.0642, 2.8998] | [0.3462, 2.9554] | [0.3515, 3.0069] | [0.3569, 3.0611] |
| 25 | 72.7962 | [0.0612, 2.8182] | [0.3423, 2.8838] | [0.3465, 2.9236] | [0.3509, 2.9651] |
| 30 | 97.7338 | [0.0637, 2.9696] | [0.3648, 3.0467] | [0.3686, 3.0815] | [0.3725, 3.1176] |
| 50 | 148.216 | [0.0588, 2.8111] | [0.3533, 2.8992] | [0.3556, 2.9189] | [0.3579, 2.9389] |
| 100 | 291.932 | [0.0569, 2.7738] | [0.3547, 2.8718] | [0.3559, 2.8814] | [0.3571, 2.8912] |

$k = 3$

| $n$ | $S$ | ($N_b$) | ($B_H$) | ($B_J$) | ($B_U$) |
|---|---|---|---|---|---|
| 10 | 45.8272 | [0.5526, 3.3452] | [0.7513, 3.3895] | [0.7649, 3.4617] | [0.7792, 3.5387] |
| 15 | 72.6349 | [0.5573, 3.3887] | [0.7879, 3.4656] | [0.7977, 3.5138] | [0.8079, 3.5641] |
| 20 | 99.3981 | [0.5591, 3.4081] | [0.8064, 3.5013] | [0.8141, 3.5374] | [0.8220, 3.5747] |
| 25 | 127.039 | [0.5621, 3.4313] | [0.8205, 3.5345] | [0.8268, 3.5635] | [0.8333, 3.5932] |
| 30 | 152.599 | [0.5601, 3.4231] | [0.8243, 3.5322] | [0.8296, 3.5563] | [0.8351, 3.5808] |
| 50 | 251.593 | [0.5527, 3.3850] | [0.8268, 3.5051] | [0.8300, 3.5193] | [0.8333, 3.5337] |
| 100 | 490.771 | [0.5425, 3.3286] | [0.8217, 3.4555] | [0.8233, 3.4625] | [0.8249, 3.4695] |

$k = 5$

Table 1: The non-Bayesian prediction interval ($N_b$), and the Bayesian prediction intervals ($B_H$), ($B_J$) and ($B_U$).

Next, for $k = 3, 4, \ldots, 10$ ; $n = 10(5)30, 50, 100$ ; $\alpha = 0.05, 0.10$ ; $\theta = 2, 0.5, 1, 3$, we obtain the coverage probabilities for the prediction intervals (see the Table 2). As is seen in the Table 2, the non-Bayesian prediction intervals ($N_b$) seem to be comparatively better than the Bayesian prediction interval ($B_H$), ($B_J$), ($B_U$) especially for $k = 3, 6, 9$.

| $n\backslash k$ | | 3 | 4 | 5 | 6 | 7 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| 10 | $(N_b)$ | 0.9505 | 0.9511 | 0.9484 | 0.9512 | 0.9502 | 0.9493 | 0.9497 |
| | $(B_H)$ | 0.9408 | 0.9454 | 0.9508 | 0.9464 | 0.9476 | 0.9598 | 0.9536 |
| | $(B_J)$ | 0.9502 | 0.9518 | 0.9544 | 0.9456 | 0.9510 | 0.9603 | 0.9544 |
| | $(B_U)$ | 0.9530 | 0.9583 | 0.9578 | 0.9432 | 0.9535 | 0.9599 | 0.9564 |
| 15 | $(N_b)$ | 0.9498 | 0.9503 | 0.9502 | 0.9496 | 0.9497 | 0.9518 | 0.9502 |
| | $(B_H)$ | 0.9431 | 0.9456 | 0.9481 | 0.9467 | 0.9489 | 0.9554 | 0.9510 |
| | $(B_J)$ | 0.9497 | 0.9501 | 0.9501 | 0.9486 | 0.9512 | 0.9559 | 0.9504 |
| | $(B_U)$ | 0.9555 | 0.9535 | 0.9527 | 0.9481 | 0.9515 | 0.9564 | 0.9518 |
| 20 | $(N_b)$ | 0.9498 | 0.9498 | 0.9500 | 0.9496 | 0.9488 | 0.9497 | 0.9485 |
| | $(B_H)$ | 0.9441 | 0.9472 | 0.9486 | 0.9449 | 0.9479 | 0.9515 | 0.9486 |
| | $(B_J)$ | 0.9489 | 0.9502 | 0.9501 | 0.9457 | 0.9492 | 0.9521 | 0.9497 |
| | $(B_U)$ | 0.9531 | 0.9538 | 0.9524 | 0.9469 | 0.9494 | 0.9531 | 0.9508 |
| 25 | $(N_b)$ | 0.9504 | 0.9513 | 0.9477 | 0.9490 | 0.9487 | 0.9496 | 0.9495 |
| | $(B_H)$ | 0.9470 | 0.9471 | 0.9457 | 0.9472 | 0.9486 | 0.9498 | 0.9484 |
| | $(B_J)$ | 0.9504 | 0.9501 | 0.9467 | 0.9485 | 0.9492 | 0.9507 | 0.9492 |
| | $(B_U)$ | 0.9530 | 0.9528 | 0.9485 | 0.9489 | 0.9499 | 0.9514 | 0.9501 |
| 30 | $(N_b)$ | 0.9505 | 0.9495 | 0.9506 | 0.9500 | 0.9507 | 0.9508 | 0.9501 |
| | $(B_H)$ | 0.9489 | 0.9474 | 0.9492 | 0.9498 | 0.9494 | 0.9517 | 0.9491 |
| | $(B_J)$ | 0.9517 | 0.9497 | 0.9518 | 0.9513 | 0.9510 | 0.9524 | 0.9498 |
| | $(B_U)$ | 0.9542 | 0.9519 | 0.9537 | 0.9521 | 0.9515 | 0.9526 | 0.9505 |
| 50 | $(N_b)$ | 0.9507 | 0.9495 | 0.9498 | 0.9496 | 0.9497 | 0.9498 | 0.9510 |
| | $(B_H)$ | 0.9484 | 0.9472 | 0.9478 | 0.9485 | 0.9482 | 0.9476 | 0.9511 |
| | $(B_J)$ | 0.9503 | 0.9487 | 0.9488 | 0.9492 | 0.9491 | 0.9485 | 0.9512 |
| | $(B_U)$ | 0.9524 | 0.9496 | 0.9500 | 0.9498 | 0.9495 | 0.9491 | 0.9512 |
| 100 | $(N_b)$ | 0.9503 | 0.9506 | 0.9499 | 0.9505 | 0.9490 | 0.9499 | 0.9496 |
| | $(B_H)$ | 0.9481 | 0.9502 | 0.9511 | 0.9495 | 0.9499 | 0.9497 | 0.9506 |
| | $(B_J)$ | 0.9489 | 0.9511 | 0.9515 | 0.9498 | 0.9503 | 0.9500 | 0.9511 |
| | $(B_U)$ | 0.9495 | 0.9515 | 0.9520 | 0.9502 | 0.9506 | 0.9502 | 0.9511 |

Table 2: Values of the coverage probabilities of the non-Bayesian prediction interval (2.9), *i.e.* $(N_b)$, and the Bayesian prediction intervals (3.7), *i.e.* $(B_H)$, $(B_J)$ and $(B_U)$ by simulation with 10000 iterations which are given in the box for each $k$ and each $n$ from the above to the bottom, where $\alpha = 0.05$ and $\theta = 2$.

| $n\backslash k$ | | 3 | 4 | 5 | 6 | 7 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| 10 | $(N_b)$ | 0.8984 | 0.9008 | 0.8996 | 0.8986 | 0.8999 | 0.9005 | 0.8995 |
| | $(B_H)$ | 0.8828 | 0.8966 | 0.9018 | 0.8975 | 0.9012 | 0.8920 | 0.9055 |
| | $(B_J)$ | 0.8976 | 0.9044 | 0.9072 | 0.9020 | 0.9017 | 0.8968 | 0.9064 |
| | $(B_U)$ | 0.9108 | 0.9122 | 0.9118 | 0.9063 | 0.9030 | 0.9009 | 0.9082 |
| 15 | $(N_b)$ | 0.8989 | 0.9007 | 0.9026 | 0.9008 | 0.8977 | 0.9009 | 0.9006 |
| | $(B_H)$ | 0.8904 | 0.8955 | 0.8978 | 0.8984 | 0.8983 | 0.8988 | 0.9032 |
| | $(B_J)$ | 0.8997 | 0.9017 | 0.9019 | 0.9016 | 0.8989 | 0.9009 | 0.9045 |
| | $(B_U)$ | 0.9086 | 0.9064 | 0.9062 | 0.9055 | 0.9006 | 0.9037 | 0.9062 |
| 20 | $(N_b)$ | 0.9009 | 0.9001 | 0.9001 | 0.8997 | 0.9013 | 0.9006 | 0.9016 |
| | $(B_H)$ | 0.8933 | 0.8957 | 0.8963 | 0.8948 | 0.9005 | 0.8935 | 0.9020 |
| | $(B_J)$ | 0.8997 | 0.8998 | 0.8996 | 0.8984 | 0.9013 | 0.8961 | 0.9038 |
| | $(B_U)$ | 0.9058 | 0.9043 | 0.9035 | 0.9013 | 0.9027 | 0.8979 | 0.9062 |
| 25 | $(N_b)$ | 0.9016 | 0.9009 | 0.9005 | 0.8997 | 0.9000 | 0.8999 | 0.9006 |
| | $(B_H)$ | 0.8973 | 0.8970 | 0.8975 | 0.8972 | 0.8977 | 0.8948 | 0.8980 |
| | $(B_J)$ | 0.9022 | 0.9010 | 0.9013 | 0.9000 | 0.8991 | 0.8955 | 0.8994 |
| | $(B_U)$ | 0.9073 | 0.9050 | 0.9042 | 0.9025 | 0.8920 | 0.8981 | 0.8999 |
| 30 | $(N_b)$ | 0.9047 | 0.8993 | 0.8987 | 0.8989 | 0.8999 | 0.9012 | 0.8999 |
| | $(B_H)$ | 0.8999 | 0.8970 | 0.8979 | 0.8989 | 0.8980 | 0.8935 | 0.9005 |
| | $(B_J)$ | 0.9044 | 0.8997 | 0.9010 | 0.9011 | 0.8990 | 0.8948 | 0.9014 |
| | $(B_U)$ | 0.9082 | 0.9029 | 0.9031 | 0.9032 | 0.9005 | 0.8946 | 0.9021 |
| 50 | $(N_b)$ | 0.9019 | 0.9012 | 0.9008 | 0.8975 | 0.8983 | 0.8998 | 0.8995 |
| | $(B_H)$ | 0.8982 | 0.8979 | 0.8981 | 0.8975 | 0.8969 | 0.8899 | 0.8998 |
| | $(B_J)$ | 0.9008 | 0.8996 | 0.8997 | 0.8984 | 0.8982 | 0.8892 | 0.9000 |
| | $(B_U)$ | 0.9032 | 0.9013 | 0.9015 | 0.8993 | 0.8993 | 0.8886 | 0.9011 |
| 100 | $(N_b)$ | 0.8992 | 0.9009 | 0.8997 | 0.8998 | 0.9000 | 0.8999 | 0.9005 |
| | $(B_H)$ | 0.8976 | 0.8999 | 0.8960 | 0.8993 | 0.9006 | 0.8882 | 0.8995 |
| | $(B_J)$ | 0.8994 | 0.9009 | 0.8968 | 0.8998 | 0.9013 | 0.8875 | 0.8999 |
| | $(B_U)$ | 0.9000 | 0.9014 | 0.8979 | 0.9001 | 0.9014 | 0.8852 | 0.9006 |

Table 2 (continued): Values of the coverage probabilities of the non-Bayesian prediction interval (2.9), *i.e.* $(N_b)$, and the Bayesian prediction intervals (3.7), *i.e.* $(B_H)$, $(B_J)$ and $(B_U)$ by simulation with 10000 iterations which are given in the box for each $k$ and each $n$ from the above to the bottom, where $\alpha = 0.10$ and $\theta = 2$.

| $n\backslash k$ | | 3 | 4 | 5 | 6 | 7 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| 10 | $(N_b)$ | 0.9505 | 0.9506 | 0.9483 | 0.9492 | 0.9509 | 0.9492 | 0.9500 |
| | $(B_H)$ | 0.9453 | 0.9432 | 0.9532 | 0.9529 | 0.9487 | 0.9491 | 0.9525 |
| | $(B_J)$ | 0.9519 | 0.9502 | 0.9548 | 0.9538 | 0.9520 | 0.9506 | 0.9534 |
| | $(B_U)$ | 0.9589 | 0.9555 | 0.9588 | 0.9542 | 0.9548 | 0.9523 | 0.9551 |
| 15 | $(N_b)$ | 0.9501 | 0.9501 | 0.9487 | 0.9506 | 0.9498 | 0.9496 | 0.9499 |
| | $(B_H)$ | 0.9429 | 0.9438 | 0.9496 | 0.9518 | 0.9481 | 0.9466 | 0.9499 |
| | $(B_J)$ | 0.9480 | 0.9498 | 0.9524 | 0.9530 | 0.9508 | 0.9485 | 0.9505 |
| | $(B_U)$ | 0.9539 | 0.9535 | 0.9550 | 0.9541 | 0.9523 | 0.9495 | 0.9516 |
| 20 | $(N_b)$ | 0.9496 | 0.9491 | 0.9500 | 0.9497 | 0.9500 | 0.9511 | 0.9510 |
| | $(B_H)$ | 0.9435 | 0.9448 | 0.9484 | 0.9485 | 0.9465 | 0.9469 | 0.9497 |
| | $(B_J)$ | 0.9491 | 0.9477 | 0.9511 | 0.9492 | 0.9476 | 0.9480 | 0.9500 |
| | $(B_U)$ | 0.9533 | 0.9508 | 0.9525 | 0.9512 | 0.9498 | 0.9488 | 0.9511 |
| 25 | $(N_b)$ | 0.9504 | 0.9492 | 0.9498 | 0.9492 | 0.9495 | 0.9495 | 0.9503 |
| | $(B_H)$ | 0.9449 | 0.9463 | 0.9488 | 0.9484 | 0.9466 | 0.9463 | 0.9502 |
| | $(B_J)$ | 0.9490 | 0.9493 | 0.9498 | 0.9501 | 0.9475 | 0.9467 | 0.9505 |
| | $(B_U)$ | 0.9532 | 0.9516 | 0.9519 | 0.9514 | 0.9496 | 0.9474 | 0.9512 |
| 30 | $(N_b)$ | 0.9497 | 0.9495 | 0.9500 | 0.9497 | 0.9490 | 0.9495 | 0.9515 |
| | $(B_H)$ | 0.9452 | 0.9451 | 0.9486 | 0.9489 | 0.9480 | 0.9463 | 0.9503 |
| | $(B_J)$ | 0.9489 | 0.9472 | 0.9504 | 0.9499 | 0.9487 | 0.9477 | 0.9518 |
| | $(B_U)$ | 0.9518 | 0.9494 | 0.9523 | 0.9517 | 0.9506 | 0.9478 | 0.9523 |
| 50 | $(N_b)$ | 0.9512 | 0.9495 | 0.9501 | 0.9495 | 0.9494 | 0.9487 | 0.9487 |
| | $(B_H)$ | 0.9488 | 0.9472 | 0.9503 | 0.9491 | 0.9496 | 0.9463 | 0.9478 |
| | $(B_J)$ | 0.9502 | 0.9488 | 0.9510 | 0.9503 | 0.9502 | 0.9468 | 0.9485 |
| | $(B_U)$ | 0.9520 | 0.9507 | 0.9523 | 0.9511 | 0.9505 | 0.9469 | 0.9490 |
| 100 | $(N_b)$ | 0.9502 | 0.9508 | 0.9507 | 0.9502 | 0.9497 | 0.9502 | 0.9492 |
| | $(B_H)$ | 0.9467 | 0.9512 | 0.9485 | 0.9504 | 0.9495 | 0.9484 | 0.9487 |
| | $(B_J)$ | 0.9476 | 0.9518 | 0.9488 | 0.9509 | 0.9499 | 0.9484 | 0.9490 |
| | $(B_U)$ | 0.9489 | 0.9523 | 0.9494 | 0.9513 | 0.9504 | 0.9485 | 0.9491 |

Table 2 (continued): Values of the coverage probabilities of the non-Bayesian prediction interval (2.9), *i.e.* $(N_b)$, and the Bayesian prediction intervals (3.7), *i.e.* $(B_H)$, $(B_J)$ and $(B_U)$ by simulation with 10000 iterations which are given in the box for each $k$ and each $n$ from the above to the bottom, where $\alpha = 0.05$ and $\theta = 0.5$.

| $n\backslash k$ | | 3 | 4 | 5 | 6 | 7 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| 10 | $(N_b)$ | 0.9505 | 0.9509 | 0.9496 | 0.9501 | 0.9506 | 0.9493 | 0.9487 |
| | $(B_H)$ | 0.9398 | 0.9443 | 0.9450 | 0.9414 | 0.9494 | 0.9497 | 0.9504 |
| | $(B_J)$ | 0.9504 | 0.9511 | 0.9505 | 0.9470 | 0.9473 | 0.9513 | 0.9522 |
| | $(B_U)$ | 0.9583 | 0.9564 | 0.9552 | 0.9511 | 0.9508 | 0.9531 | 0.9535 |
| 15 | $(N_b)$ | 0.9498 | 0.9500 | 0.9512 | 0.9499 | 0.9501 | 0.9505 | 0.9499 |
| | $(B_H)$ | 0.9427 | 0.9439 | 0.9458 | 0.9424 | 0.9479 | 0.9489 | 0.9476 |
| | $(B_J)$ | 0.9497 | 0.9497 | 0.9499 | 0.9462 | 0.9467 | 0.9504 | 0.9483 |
| | $(B_U)$ | 0.9557 | 0.9543 | 0.9512 | 0.9499 | 0.9487 | 0.9517 | 0.9496 |
| 20 | $(N_b)$ | 0.9501 | 0.9505 | 0.9494 | 0.9498 | 0.9502 | 0.9474 | 0.9483 |
| | $(B_H)$ | 0.9449 | 0.9469 | 0.9471 | 0.9443 | 0.9484 | 0.9458 | 0.9482 |
| | $(B_J)$ | 0.9498 | 0.9499 | 0.9501 | 0.9465 | 0.9473 | 0.9474 | 0.9493 |
| | $(B_U)$ | 0.9536 | 0.9528 | 0.9528 | 0.9478 | 0.9479 | 0.9478 | 0.9485 |
| 25 | $(N_b)$ | 0.9500 | 0.9510 | 0.9487 | 0.9497 | 0.9514 | 0.9502 | 0.9487 |
| | $(B_H)$ | 0.9455 | 0.9477 | 0.9470 | 0.9436 | 0.9513 | 0.9477 | 0.9470 |
| | $(B_J)$ | 0.9493 | 0.9503 | 0.9492 | 0.9466 | 0.9476 | 0.9484 | 0.9481 |
| | $(B_U)$ | 0.9527 | 0.9528 | 0.9516 | 0.9487 | 0.9482 | 0.9496 | 0.9488 |
| 30 | $(N_b)$ | 0.9506 | 0.9496 | 0.9501 | 0.9512 | 0.9504 | 0.9480 | 0.9513 |
| | $(B_H)$ | 0.9450 | 0.9473 | 0.9488 | 0.9469 | 0.9489 | 0.9457 | 0.9487 |
| | $(B_J)$ | 0.9478 | 0.9495 | 0.9503 | 0.9488 | 0.9466 | 0.9465 | 0.9508 |
| | $(B_U)$ | 0.9504 | 0.9519 | 0.9517 | 0.9505 | 0.9475 | 0.9468 | 0.9491 |
| 50 | $(N_b)$ | 0.9489 | 0.9499 | 0.9505 | 0.9494 | 0.9502 | 0.9500 | 0.9501 |
| | $(B_H)$ | 0.9473 | 0.9501 | 0.9483 | 0.9487 | 0.9493 | 0.9481 | 0.9492 |
| | $(B_J)$ | 0.9491 | 0.9512 | 0.9492 | 0.9496 | 0.9462 | 0.9488 | 0.9501 |
| | $(B_U)$ | 0.9510 | 0.9526 | 0.9503 | 0.9507 | 0.9461 | 0.9494 | 0.9506 |
| 100 | $(N_b)$ | 0.9496 | 0.9508 | 0.9504 | 0.9514 | 0.9496 | 0.9508 | 0.9503 |
| | $(B_H)$ | 0.9494 | 0.9491 | 0.9489 | 0.9515 | 0.9492 | 0.9500 | 0.9508 |
| | $(B_J)$ | 0.9503 | 0.9499 | 0.9493 | 0.9520 | 0.9487 | 0.9504 | 0.9511 |
| | $(B_U)$ | 0.9508 | 0.9505 | 0.9499 | 0.9525 | 0.9484 | 0.9505 | 0.9512 |

Table 2 (continued): Values of the coverage probabilities of the non-Bayesian prediction interval (2.9), *i.e.* $(N_b)$, and the Bayesian prediction intervals (3.7), *i.e.* $(B_H)$, $(B_J)$ and $(B_U)$ by simulation with 10000 iterations which are given in the box for each $k$ and each $n$ from the above to the bottom, where $\alpha = 0.05$ and $\theta = 1$.

| $n\backslash k$ | | 3 | 4 | 5 | 6 | 7 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| 10 | $(N_b)$ | 0.9518 | 0.9500 | 0.9505 | 0.9502 | 0.9495 | 0.9494 | 0.9494 |
| | $(B_H)$ | 0.9433 | 0.9439 | 0.9522 | 0.9417 | 0.9501 | 0.9450 | 0.920 |
| | $(B_J)$ | 0.9525 | 0.9518 | 0.9547 | 0.9418 | 0.9541 | 0.9474 | 0.9453 |
| | $(B_U)$ | 0.9608 | 0.9581 | 0.9585 | 0.9429 | 0.9555 | 0.9501 | 0.9483 |
| 15 | $(N_b)$ | 0.9508 | 0.9488 | 0.9492 | 0.9498 | 0.9507 | 0.9492 | 0.9506 |
| | $(B_H)$ | 0.9444 | 0.9436 | 0.9483 | 0.9429 | 0.9489 | 0.9472 | 0.9440 |
| | $(B_J)$ | 0.9503 | 0.9479 | 0.9505 | 0.9433 | 0.9519 | 0.9485 | 0.9457 |
| | $(B_U)$ | 0.9557 | 0.9516 | 0.9542 | 0.9422 | 0.9529 | 0.9496 | 0.9479 |
| 20 | $(N_b)$ | 0.9513 | 0.9493 | 0.9484 | 0.9503 | 0.9487 | 0.9500 | 0.9490 |
| | $(B_H)$ | 0.9463 | 0.9471 | 0.9468 | 0.9440 | 0.9479 | 0.9429 | 0.9432 |
| | $(B_J)$ | 0.9512 | 0.9505 | 0.9485 | 0.9447 | 0.9496 | 0.9444 | 0.9441 |
| | $(B_U)$ | 0.9559 | 0.9537 | 0.9510 | 0.9448 | 0.9512 | 0.9467 | 0.9469 |
| 25 | $(N_b)$ | 0.9504 | 0.9485 | 0.9509 | 0.9499 | 0.9491 | 0.9507 | 0.9505 |
| | $(B_H)$ | 0.9476 | 0.9464 | 0.9477 | 0.9462 | 0.9470 | 0.9424 | 0.9453 |
| | $(B_J)$ | 0.9514 | 0.9491 | 0.9499 | 0.9469 | 0.9515 | 0.9442 | 0.9472 |
| | $(B_U)$ | 0.9542 | 0.9512 | 0.9514 | 0.9484 | 0.9532 | 0.9459 | 0.9479 |
| 30 | $(N_b)$ | 0.9500 | 0.9502 | 0.9496 | 0.9505 | 0.9505 | 0.9493 | 0.9494 |
| | $(B_H)$ | 0.9472 | 0.9471 | 0.9482 | 0.9485 | 0.9494 | 0.9405 | 0.9451 |
| | $(B_J)$ | 0.9501 | 0.9494 | 0.9501 | 0.9492 | 0.9507 | 0.9420 | 0.9470 |
| | $(B_U)$ | 0.9534 | 0.9515 | 0.9514 | 0.9496 | 0.9514 | 0.9437 | 0.9487 |
| 50 | $(N_b)$ | 0.9503 | 0.9509 | 0.9515 | 0.9501 | 0.9501 | 0.9508 | 0.9497 |
| | $(B_H)$ | 0.9476 | 0.9497 | 0.9515 | 0.9502 | 0.9489 | 0.9445 | 0.9485 |
| | $(B_J)$ | 0.9493 | 0.9507 | 0.9530 | 0.9508 | 0.9497 | 0.9430 | 0.9486 |
| | $(B_U)$ | 0.9510 | 0.9521 | 0.9536 | 0.9515 | 0.9501 | 0.9459 | 0.9493 |
| 100 | $(N_b)$ | 0.9504 | 0.9504 | 0.9498 | 0.9485 | 0.9496 | 0.9499 | 0.9501 |
| | $(B_H)$ | 0.9493 | 0.9488 | 0.9478 | 0.9493 | 0.9492 | 0.9442 | 0.9504 |
| | $(B_J)$ | 0.9499 | 0.9495 | 0.9484 | 0.9499 | 0.9496 | 0.9447 | 0.9505 |
| | $(B_U)$ | 0.9509 | 0.9503 | 0.9488 | 0.9502 | 0.9505 | 0.9450 | 0.9509 |

Table 2 (continued): Values of the coverage probabilities of the non-Bayesian prediction interval (2.9), *i.e.* $(N_b)$, and the Bayesian prediction intervals (3.7), *i.e.* $(B_H)$, $(B_J)$ and $(B_U)$ by simulation with 10000 iterations which are given in the box for each $k$ and each $n$ from the above to the bottom, where $\alpha = 0.05$ and $\theta = 3$.

## Acknowledgements

The author wishes to thank Professor M. Akahira for his valuable advices and encouragements.

## REFERENCES

[1] Akahira, M. (1990). *Theory of Statistical Prediction.* Lecture Note at the Middle East Technical University, Ankara.

[2] Geisser, S. (1993). *Predictive Inference: An Introduction.* Chapman & Hall, New York.

[3] Guttman, I. (1970). *Statistical Tolerance Regions: Classical and Bayesian.* Griffin, London.

[4] Hartigan, J. (1964). Invariant prior distributions. *Ann. Math. Statist.* **35**, 836–845.

[5] Howlader, H. A., Hossain, A. (1998). Bayesian prediction intervals for Maxwell parameters. *Metron* **56**, 97-106.

[6] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions Volume 1* (2nd ed.), Wiley, New York.

[7] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions Volume 2* (2nd ed.), Wiley, New York.

[8] Takada, Y. (1996). Statistical properties of prediction intervals. *Sugaku Expositions* **9**, 153–168.

[9] Takeuchi, K. (1975). *Statistical Prediction Theory.* (In Japanese), Baifukan, Tokyo.

## ÖZET

$X_1, X_2, ..., X_n$ bağımsız ve ayni $k$ serbestlik derecesine ve $\sqrt{\theta}$ karma parametresine sahip olan $\chi_k(\theta)$ dağılımından elde edilmiş bir örneklem olsun. Varşayalım ki, $Y$ ayni $\chi_k(\theta)$ dağılımından olan ve $\mathbf{X} = (X_1, X_2, ..., X_n)$ örnekleminden bağımsız olan bir rasgele değişkendir. Bayes ve Bayes olmayan bakış açılarından $Y$ için $\dot{\mathbf{X}}$ 'e dayalı öngörü güven aralıkralı kuruluyor.

**İstatistik, Journal of the Turkish Statistical Association**
**2000, Vol. 3, No 1-2, pp. 13-20**

# A NOTE ON THE PROBABILITY BOUND FOR m-DEPENDENT RANDOM VARIABLES

İhsan Karabulut
Department of Statistics, Ankara University, Ankara

**Abstract**

An upper bound for probabilities of sum of non-identical and m-dependent random variables is proposed. A theorem and application extensions are provided.

**Key Words:**Truncation, large deviation probabilities, convexity.

## 1. Introduction

Bounds on the large deviation probabilities for the sum of random variables (r.v.'s) are important tools for both practical and theoretical purposes. The important bounds are the well-known kornerstones of Tchebychef and Markov bounds. Bennet(1962) and Hoeffding(1963) and the some references therein give good account on the development of the subject. Also, Petrov(1995) devotes a short chapter on the probability inequalities for sums of independent r.v.'s. Nagaev(1965) obtains a large deviation inequality for the sum of independent and identically distributed r.v.'s under the existence of moments of order larger than two. Kurtz(1972) gives a number of inequalities with the bounds constructed by means of some specified functions. He obtained bounds for weigthed sum of independent but not necessarily identical random variables. Fuk and Nagaev(1971) proposed probability inequalities for sums of non-identically distributed independent rv.'s. Probability inequalities for dependent random variables do not have long history as much as for the independent r.v.'s. Partly because the Tchebychef or Markov inequalities can be used in case of sum of dependent r.v.'s via using the some order of expectations. For example, Shi and Shao(1988) used Marcinkiewitz's result to each component after decomposing sum of m-dependent and identically distributed random variables. Tikhomirow(1980), uses one of his results on the central limit theorem of weakly dependent stationary sequence of r.v.'s including m-dependent r.v.'s.

A number of probability bounds for sums of bounded and dependent r.v.'s are given by Hoeffding(1963) with their uses in applications. Recently, Matula(1997) have obtained probability and moment bounds for negatively associated random variables. In this paper, a result by Hoeffding(1963) for large deviation probability of on sums of bounded and m-dependent r.v.'s(see, subsection 5.5.d. of ) is extended to the m-dependent r.v.'s by combining another result by Fuk and Nagaev(1971) .

## 2. Preparation

In this section some notation will be fixed and explained the method for obtaining the probability bound on sums of m-dependent r.v.'s given in the next section. The notation will be kept the same as those of Fuk and Nagaev(1971) and Hoeffding(1963).

Let $X_1, X_2, \ldots, X_n$ be any sequence of m-dependent r.v.'s that is any two vectors of the form $(X_{a-p}, X_{a-p+1}, \ldots, X_{a-1}, X_a)$ and $(X_b, X_{b+1}, \ldots, X_{b+q-1}, X_{b+q})$ are independent for $b-a > m$. Throughout the paper prescibed real numbers $x, y_1, y_2, \ldots, y_n$ will be all positive. Truncation will play an important role in the development of the probability bounds for the sums of m-dependent r.v.'s. Here, the truncation behaves as a trade-off tool between having all order of the moments greater than 2 without any pain and slow decreasing large deviation probabilities namely the summands in the first term of the right hand side of (2.1) below. A truncation of the r.v.'s $X_1, X_2, \ldots, X_n$ is defined as

$$Y_i = \begin{cases} X_i & , X_i \le y_i \\ 0 & , X_i > y_i \end{cases}$$

Define $S = \sum_{i=1}^n X_i$ and $S^T = \sum_{i=1}^n Y_i$. It is known that for any sequence of r.v.'s and a positive constant $x$

$$\begin{aligned} \{(X_1, X_2, \ldots X_n) : S \ge x\} &= \{S^T \ge x, X_i \le y_i, i = 1, 2, \ldots, n\} \\ &\cup \ \{S \ge x, X_i > y_i \text{ at least one } i = 1, 2, \ldots, n\}. \end{aligned}$$

Therefore, we have

$$P(S \ge x) \le \sum_{i=1}^{n} P(X_i \ge y_i) + P(S^T \ge x). \tag{2.1}$$

The exponential bounds for the large deviation probabilities are preferred in applications mainly due to their easy manipulations for obtaining tight bounds. In summary, for constant $h > 0$, $x \in R$ and $Ee^{hS} < \infty$ we can write that

$$P(S \ge x) \le e^{-hx} E e^{hS}. \tag{2.2}$$

In case of independence of r.v.'s in $S$ , the expectatation on the right hand side of (2.2) can be written as the product of $Ee^{hX_i}$. The problem arises when the independence assumption is violated.

A remedy for the arisen problem is to decompose the sum of any m-dependent r.v.'s into partial sums $S_j$ of independent r.v.'s which are not necessarily independent(see 5.5.d. of Hoeffding(1963)). These sums are written as follows:

$$S_j = X_j + X_{j+(m+1)} + X_{j+2(m+1)} + \ldots + X_{j+(m+1)(n_j-1)}.$$

where, $n_j$ is the integer value defined as $n_j = [\frac{n-j+(m+1)}{m+1}]$ The summands above are independent because of the m-dependence of the r.v.'s. Therefore, we write

$$S = S_1 + S_2 + \ldots + S_{m+1}.$$

Hereafter, the r.v. $X_i$, its truncation $Y_i$ and the preassigned truncation constant $y_i$ will be denoted as $X_{ij}$, $Y_{ij}$ and $y_{ij}$ respectively if $X_i$ belongs to the sum $S_j$.

As a provision for the next steps choose positive constants $p_j = n_j/n$ which satisfies $\sum_{j=1}^{m+1} p_j = 1$, then $\bar{X} = S/n$ can be written as

$$\bar{X} = \frac{p_1}{n_1} S_1 + \frac{p_2}{n_2} S_2 + \ldots + \frac{p_{m+1}}{n_{m+1}} S_{m+1}. \tag{2.3}$$

Before stating the main result of this manuscript some more notation will be introduced : Let, $F_{ij}$ denote the distribution function of the r.v. $X_i$, which is the summand in the partial sum $S_j$. The truncated moments and their sums are

$$\mu_j = \sum_{i=1}^{n_j} \int_{-\infty}^{y_{ij}} x dF_{ij}(x), B_j^2 = \sum_{i=1}^{n_j} \int_{-\infty}^{y_{ij}} x^2 dF_{ij}(x)$$

$$A_{j,t} = \sum_{i=1}^{n_j} \int_{-y_{ij}}^{0} |x|^t dF_{ij}(x)$$

Also, define $y_j = \max(y_{1j}, y_{2j}, y_{3j}, \ldots, y_{n_j j})$ for each $j = 1, 2, \ldots, m+1$

Though, the results are stated with the moments truncated at the $y_{ij}$'s, they are valid for full moments without truncation as much as their existence are not under question.

## 3. A Crude Bound for the Sums of Random Variables

The Tchebychef or Markov inequalities can be used safely for the sums of any type of dependent r.v.'s as it is noted in the introduction at the expense of assuming the existence of all degree of moments and product moments as much as they are needed. That limits the utilization of these inequalities. In the following, a probability bound for the sums of m-dependent random variables will be given in which the order limitations of the existence of moments partly removed at the expense of its crudeness.

Although being a crude bound it may be helpful in some instances especially where it is not known much about the existence of any order expectations greater than two and more product moments. The result that we present can be extended to other cases considered by Fuk and Nagaev(1971).

**Remark.** With some effort it seems to be possible to obtain a more refined exponential probability bound for m-dependent r.v.'s possibly by adding the stationarity assumption on the sequence of r.v.'s using the method provided by Tikhomirov(1980). In this case one should work with moment generating functions of truncated r.v.'s instead of using characteristic functions.

**Theorem.** Let, $X_1, X_2, \ldots, X_n$ be a sequence of m-dependent random variables, for $t \geq 2$ and constants $0 < \alpha < 1$, $\beta = 1 - \alpha$. For each $j = 1, 2, \ldots, m+1$ if

$$\max\left(\frac{tn_j}{y_j}, \frac{n_j}{y_j}\ln(\frac{n_j y_j^{t-1}\beta x}{A_{j,t}} + 1)\right) \leq \frac{\alpha x n_j^2}{e^t B_j^2}$$

then

$$\begin{aligned} P(\bar{X} \geq x) &\leq \textstyle\sum_{j=1}^{m+1}\sum_{i=1}^{n_j} P(X_{ij} \geq y_{ij}) \\ &+ \textstyle\sum_{j=1}^{m+1} p_j \exp\left\{\beta\frac{n_j x}{y_j} - ((1-\frac{\alpha}{2})\frac{n_j x}{y_j} - \frac{n_j\mu_j}{y_j})\ln(\frac{n_j y_j^{t-1}\beta x}{A_{j,t}} + 1)\right\} (3.1) \end{aligned}$$

or

$$\begin{aligned} P(\bar{X} \geq x) &\leq \textstyle\sum_{j=1}^{m+1}\sum_{i=1}^{n_j} P(X_{ij} \geq y_{ij}) \\ &+ \textstyle\sum_{j=1}^{m+1} p_j \exp\left\{(\beta - \frac{\alpha t}{2})\frac{n_j x}{y_j} - (\beta\frac{n_j x}{y_j} - \frac{n_j\mu_j}{y_j})\ln(\frac{n_j y_j^{t-1}\beta x}{A_{j,t}} + 1)\right\} (3.2) \end{aligned}$$

otherwise

$$P(\bar{X} \geq x) \leq \sum_{j=1}^{m+1}\sum_{i=1}^{n_j} P(X_{ij} \geq y_{ij}) + \sum_{j=1}^{m+1} p_j \exp\left\{-\alpha x n_j^2(\frac{\frac{\alpha x}{2} - \mu_j}{e^t B_j^2})\right\} \qquad (3.3)$$

The inequalities above may be found more meaningful in some cases if arithmetic mean $\bar{X}$ is replaced by $S$. In this case $x$ will be replaced by $n.x$ in the terms in second sum the each inequalities.

The following corollary is a natural extension of the theorem as for Corollary 4. of Fuk and Nagaev(1971). It simplifies the use of the theorem.

**Corollary.** *If $EX_i = 0$ and $E|X_i|^t < \infty$ for $i = 1, 2, \ldots, n$, $t \geq 2$, $\beta = t/(t+2)$ and $y_j = y_{ij} = \ldots = y_{n_j} = \beta x n_j$ for $j = 1, 2, \ldots, m+1$. then from the last two inequalities*

$$\begin{aligned} P(\bar{X} \geq x) &\leq \textstyle\sum_{j=1}^{m+1} \frac{A_{j,t}c_1}{(n_j x)^t} + \sum_{j=1}^{m+1} p_j (\frac{A_{j,t}}{(n_j \beta x)^t + A_{j,t}}) \exp\left\{-\frac{c_2 x^2 n_j^2}{B_j^2}\right\} \\ &\leq \textstyle\sum_{j=1}^{m+1} \frac{A_{j,t}c_1}{(n_j x)^t} + \sum_{j=1}^{m+1} p_j \exp\left\{-\frac{c_2 x^2 n_j^2}{B_j^2}\right\} \end{aligned}$$

*where $c_1 = \beta^{-t}$ and $c_2 = 2(t+2)^{-2} e^{-t}$.*

The proof follows the following route:

The second term in the right hand side of (2.1) will be obtained for $\bar{X}$ in the first step. Then, the first term is obtained in any plausable way; especially, this term can be obtained by the use of Markov inequality for each r.v. $X_i$ under the existence of specified order of moments as it is done in the proof of the Corollary.

Note that right hand side of (2.2) is bounded for the arithmetic mean of the truncated r.v.'s. Hence, for $h > 0$, and any real number $x > 0$, we can use Hoeffding(1963)'s argument on decomposing $\bar{X}$, namely (2.3). Because of the convexity of the exponential function, it is obtained that

$$\begin{aligned} P(\bar{X}^T \geq x) &\leq e^{-hx} E e^{h \frac{1}{n} \sum_{j=1}^{m+1} p_j S_j^T} \\ &\leq e^{-hx} \sum_{j=1}^{m+1} p_j E e^{\frac{h}{n_j} S_j^T} \end{aligned}$$

Each expectation in the sum $S_j^T$ now can be written for each $j$ as

$$E e^{\frac{h}{n_j} S_j^T} = \Pi_{i=1}^{n_j} E e^{\frac{h}{n_j} Y_{ij}}$$

In the lines of Fuk and Nagaev(1971), using the upper bounds $hy_j < t$ and $hy_j \geq t$ separately, where $t \geq 2$, then combining these in (2.2) we get

$$E \exp\left\{\frac{h}{n_j} S_j^T - hx\right\}$$

$$\leq \exp\left\{\left(\frac{1}{2} e^t B_j^2 (\frac{h}{n_j})^2 - \alpha h x\right) + \left(\frac{e^{\frac{h}{n_j} y_j} - 1 - \frac{h}{n_j} y_j}{y_j^t} A_{j,t} - \beta h x\right) + h\mu_j\right\}$$

To minimize the right hand side of the inequality with respect to h, each part in parentheses of the exponent is considered separately, regarding the $hy_j < t$ and $hy_j \geq t$ cases. As a result, the minimizing h values are found as

$$h_1 = \frac{\alpha x n_j^2}{e^t B_j^2}$$

and

$$h_2 = \max\left(\frac{n_j t}{y_j}, \frac{n_j}{y_j}\ln(\frac{n_j y_j^{t-1}\beta x}{A_{j,t}} + 1)\right)$$

If $h_2 \leq h_1$ then using the same arguments for (32) and (33) of Fuk and Nagaev(1971) it is obtained that

$$Ee^{\frac{h}{n_j}S_j^T - hx} \leq n_j\beta\frac{x}{y_j} - ((1-\frac{\alpha}{2})x - \mu_j)h_2 \tag{3.4}$$

$$\leq (\beta - \frac{\alpha t}{2})\frac{n_j x}{y_j} - (\beta x - \mu_j)h_2. \tag{3.5}$$

By replacing $h_2 = \frac{n_j}{y_j}\ln(\frac{n_j y_j^{t-1}\beta x}{A_{j,t}} + 1)$ for each $j$ in (3.4) we obtain the second term in the right hand side of the inequality (3.1). Similarly (3.2) is obtained if $h_2 = \frac{n_j t}{y_j}$ is replaced. Otherwise, pluging in $h_1$ for each $j$ in (3.5) we end up with the second term on the right hand side of the inequality (3.3). Then the proof is completed.

It is also possible to obtain bilateral versions of the given inequalities above by noting that

$$P(|\bar{X}| \geq x) = P(\bar{X} \geq x) + P(-\bar{X} \geq x).$$

Hence, after replacing $x, y_j$, $B_j^2$ and $\mu_j$ by $-x, -y_j$

$$B_j^2 = \sum_{i=1}^{n_j}\int_{-y_{ij}}^{\infty} x^2 dF_{ij}(x),\ \ \mu_j = -\sum_{i=1}^{n_j}\int_{-y_{ij}}^{\infty} x dF_{ij}(x)$$

respectively and defining

$$B_{n_j}^2 = \sum_{i=1}^{n_j}\int_{-\infty}^{\infty} x^2 dF_{ij}(x),\ \ A_{n_{ij}} = \int_{-\infty}^{\infty}|x|^t dF_{ij}(x)$$

in the proof above the inequality given in corollary can be written as

$$P(|\bar{X}| \geq x) \leq \sum_{j=1}^{m+1}\sum_{i=1}^{n_j} P(|X_{ij}| \geq y_{ij}) + 2\sum_{j=1}^{m+1} p_j \exp\left\{-\frac{c_2 x^2 n_j^2}{B_{n_j}^2}\right\}$$

where $c_2$ is defined as previously.

## 4. Application

Hoeffding(1963)'s idea of finding probability bounds for bounded and m-de- pendent r.v.'s has been used together with that of Fuk and Nagaev(1971) to obtain the results presented in the previous section. The same method applies to the Theorem 4.

of Fuk and Nagaev(1971) or Lemma 2.3 of Petrov(1995). So, we obtain the following bilateral inequality for sum of m-dependent r.v.'s:

$$P(|S| \geq x) \leq \sum_{j=1}^{m+1} \sum_{i=1}^{n_j} P(|X_{ij}| \geq y_{ij}) + 2 \sum_{j=1}^{m+1} p_j \exp\left\{ \frac{x}{y_j} - \frac{x}{y_j} \ln(\frac{x y_j}{B_{n_j}^2} + 1) \right\}. \quad (4.1)$$

This inequality allows us to obtain a bound for $E|S|^k$ under the existence of **k** th moment of the each r.v. which are m-dependent and non-identical. To reach this end, we will use two facts. Those are:

$$E|Y|^k = \int_0^\infty P(|Y| \geq y) k y^{k-1} dy$$

under the existence of $k$th moment of the any r.v. and for $k \geq 2$,

$$\left( \sum_{i=1}^{n_j} E X_{ij}^2 \right)^{k/2} \leq n_j^{\frac{k}{2}-1} \sum_{i=1}^{n_j} E|X_{ij}|^k.$$

Multiplying the both side of the inequality (4.1) by $kx^{k-1}$ then integrating with respect to x after choosing appropriate positive constants $r_{ij}$ that are $r_{ij} > k/2$ and $y_{ij} = x/r_{ij}$ and determining $r_j = \min\{y_{ij} : i = 1, 2, \ldots, n_j\}$, $j = 1, 2, \ldots, m+1$. we get

$$E|S|^k \leq \sum_{j=1}^{m+1} \sum_{i=1}^{n_j} r_{ij}^k E|X_{ij}|^k + \sum_{j=1}^{m+1} \frac{k(r_j n_j)^{k/2} e^{r_j}}{n} \sum_{i=1}^{n_j} E|X_{ij}|^k Beta(k/2, r_j - k/2).$$

Here, $Beta(k/2, r_j - k/2)$ denotes the beta function $\int_0^1 x^{\frac{k}{2}-1}(1-x)^{r_j - \frac{k}{2}-1} dx$.

Further simplifications can be made in the last inequality.

## References

[1] Bennet, G. (1962), Probability Inequalities for the Sum of Independent Random Variables, *J. Amer. Statist. Assoc.*, **57**, 33-45.

[2] Fuk, D. Kh., Nagaev, S. V. (1971), Probability Inequalities for Sums of Independent Random Variables, *Theory Probab. Appl.*, **16**, 643-660.

[3] Hoeffding, W. (1963), Probability Inequalities for Sums of Bounded Random Variables, *J. Amer. Statist. Assoc.*, **58**, 13-30.

[4] Kurtz, Thomas G. (1972), Inequalities for the Law of Large Numbers, *Ann. Math. Statist.*, **43**, 1874-1883.

[5] Matula, P. (1997), Probability and Moment Bounds for Sums of Negatively Associated Random Variables, *Theor. Probab. Math. Statist.*, **55**, 135-141.

[6] Nagaev, S. V. (1965), Some Limit Theorems for Large Deviations, *Theor. Probab. Appl.*, **10**, 214-235.

[7] Petrov, V. V. (1995), *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*, Oxford Science Publications, Clarendon Press, Oxford.

[8] Shi, X. and Shao, J. (1988), Resampling Estimation When Observations are m-dependent, *Commun. Statist.-Theory Meth.*, **17**(11), 3923-3934.

[9] Tikhomirov, A., N. (1980), On the Convergence Rate in the Central Limit Theorem for Weakly Dependent Random Variables, *The Theor. Probab. Appl.*, **25**, 790-809.

## ÖZET

m- bağımlı olup aynı dağılımlı olmayan rasgele değişkenlerin toplamları için bir olasılık üst sınırı verilmiştir.

İstatistik, Journal of the Turkish Statistical Association
2000, Vol. 3, No 1-2, pp. 21-28

# A TOOL TO MONITOR PROCESSES

Donald S. Holmes

Stochos, Inc. 14 N. College Street,
Schenectady, N.Y. 12305, U.S.A.

A. Erhan Mergen

Rochester Institute of Technology, College of Business
Decision Sciences, 107 Lomb Memorial Dr.
Rochester, N.Y. 14623, U.S.A.

**Abstract**

The purpose of this paper is to present a method which will enable the manager to make comparisons of the quality performance of a set of product lines so that the management's attention is concentrated where it is most needed. The comparison is usually difficult because the different lines may produce different parts with different properties or with the same properties but different specifications. The method proposed in this paper uses three Z values, which are described below (For other management tools, see, for example, Holmes (1986), and Holmes and Mergen (1989). An example of the use of the proposed measures is also discussed.

## 1. Introduction

Suppose a plant has two product lines; say product A and product B. Assume that three quality characteristics for product A and two for product B are being checked. SPC (Statistical Process Control) personnel may have separate charts to monitor each characteristic of each product line. Management, however, needs to know how the plant is doing overall. The charts used by the SPC people at the operational level are too detailed for management and comparisons of different characteristics and/or the product lines, which have different units of measures, are difficult. This paper presents three measures, which will allow management to evaluate quality performance in three areas:

1. Conformance to Nominal (process target value)
2. Conformance to tolerance (process width)
3. Ability to maintain statistical control (process stability).

By using the three values proposed in this paper, management could answer such questions as: "Is the process centered on nominal with respect to quality characteristics 1, 2, and 3?", "Is the process capable of satisfying the specification limits set for the characteristics 1, 2, and 3?", "Is the process in statistical control with respect to characteristics 1, 2, and 3?"

## 2. Proposed model

The three Z values proposed in this paper provide the following information about:

1. *How close the average is to the nominal of the specifications, Z-Nominal ($Z_N$)*
2. *How the width of the process compares with the desired width, Z-Sigma ($Z_S$) – where actual width is defined as six standard deviations and desired width is defined as upper specification limit (USL) minus lower specification limit (LSL).*
3. *The state of statistical control (i.e., stability) of the process, Z-Control ($Z_C$).*

***1. Z-Nominal ($Z_N$):***

$Z_N$ is used to show the condition of the process center relative to the process nominal. $Z_N$ is defined as follows:

$$Z_N = \frac{(\text{Average - Nominal})}{\frac{s}{\sqrt{n}}} \qquad (1)$$

where average is the process average in a given time period, s is the process standard deviation in a given time period, and n is the number of observations in a given time period.

This measure is similar to the T-rate system developed and used by the Bell Laboratories (see Hodley (1981)). Note that we are using Z rather than t since we assume that n will be 30 or more for summary data. The values for average, s, and n can be obtained from descriptive statistics output for a product line for each time period. If the absolute value of $Z_N$ for a product line for a certain characteristic is greater than 3, then this indicates that the process for that product line for the characteristic is not centered on nominal during that time period.

***2. Z-Sigma ($Z_S$):***

$Z_S$ is used to show condition of the actual process width of a product line relative to the desired width (i.e., tolerance (T), which is upper specification limit minus lower specification limit).

$$Z_s = \frac{s - s'}{s(s)} \qquad (2)$$

where s is the process standard deviation in a given time period, s' is the desired standard deviation, and s(s) is the standard deviation of s values. If we take sample size of n and if T/8 is the desired standard deviation then $Z_S$ becomes as defined in equation (3)

$$= \frac{\left(s - \left(\frac{T}{8}\right)\right)}{\left(\frac{\left(\frac{T}{8}\right)}{\sqrt{2n}}\right)} \qquad (3)$$

If $Z_S$ is between ∀3, there is no statistical evidence that the specifications are not being met to a satisfactory level. If the $Z_S$ value for a product line for a certain characteristic is less than –3 then the data indicates that the process is capable of satisfying the relevant specification limits during that time period. If $Z_S$ is greater than +3, it means the process in question is too wide to meet the specification limits.

***3. Z-Control ($Z_C$):***

$Z_C$ is used to show the condition of the process relative to statistical control. One approach to this issue involves Mean Square Successive Difference (MSSD), which is defined as follows:

$$MSSD = \frac{1}{n-1}\sum_{i=1}^{n-1}\left(X_{i+1} - X_i\right)^2 \qquad (4)$$

where $X_i$'s are the individual observations and n is the number of observations in a given time period.

Dividing the MSSD by 2, $\frac{MSSD}{2}$, provides an estimate for potential variance if the process is in control. A comparison of $\frac{MSSD}{2}$ to the variance calculated the usual way may be used to test for randomness (see Dixon and Massey (1969) and Holmes and Mergen (1995)). If n is greater than 20 and the population is normal then

$$Z_C = \left[1 - \frac{MSSD}{2s^2}\right] / \sqrt{\frac{(n-2)}{(n-1)(n+1)}} \qquad (5)$$

is approximately normally distributed with a mean value of zero and a standard deviation of one. If $Z_C$ is greater than +3, it indicates lack of control due to trends; if $Z_C$ is less than –3, it indicates lack of control due to cycles. $Z_C$ values inside ∀3 reflect "random" variation, i.e., no evidence that process is not in control.

Again by comparing $Z_C$ values for different product lines, one can quickly check which product lines are in control (i.e., stable).

## 3. Example

Suppose the plant that we mentioned in the introduction has two product lines: product A and B. The nominal values and the specification limits for each of the different quality characteristics of these two product lines are given in Table 1 below.

| | PRODUCT A | | | PRODUCT B | |
|---|---|---|---|---|---|
| | Chr.1 | Chr.2 | Chr.3 | Chr.1 | Chr.2 |
| NOM | 50.00 | 3.00 | 17.50 | 30.00 | 5.00 |
| USL | 52.00 | 5.00 | 18.50 | 35.00 | 7.00 |
| LSL | 48.00 | 1.00 | 16.50 | 25.00 | 3.00 |

Table 1. Nominal values and specifications for product A and B.

Four monthly averages, standard deviations, and the three Z values for each product line for each characteristic are given in Table 2.

MONTH

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| <u>PRODUCT A</u> | | | | |
| <u>Chr.1</u> | | | | |
| AVE. | 50.50 | 49.60 | 49.96 | 49.97 |
| STD.DEV. | 0.50 | 0.48 | 0.27 | 0.25 |
| $Z_N$ | 7.07 | -5.89 | -1.05 | -0.85 |
| $Z_S$ | 0.00 | -0.40 | -4.60 | -5.80 |
| $Z_C$ | 1.75 | 0.10 | -2.25 | -1.53 |
| <u>Chr.2</u> | | | | |
| AVE. | 3.98 | 3.50 | 3.99 | 4.00 |
| STD.DEV. | 0.71 | 0.45 | 0.49 | 0.79 |
| $Z_N$ | 9.76 | 7.85 | 14.28 | 8.95 |
| $Z_S$ | 4.20 | -1.00 | -0.20 | 5.80 |
| $Z_C$ | -0.63 | 0.30 | -0.47 | 1.01 |
| <u>Chr.3</u> | | | | |
| AVE. | 17.81 | 17.60 | 17.64 | 18.00 |
| STD.DEV. | 0.60 | 0.62 | 0.59 | 0.80 |
| $Z_N$ | 3.65 | 1.14 | 1.68 | 4.42 |
| $Z_S$ | 14.00 | 14.80 | 13.60 | 22.00 |

| | | | | |
|---|---|---|---|---|
| $Z_C$ | 0.49 | -0.65 | -0.40 | 3.43 |
| PRODUCT B | | | | |
| Chr.1 | | | | |
| AVE. | 32.00 | 31.84 | 29.99 | 30.00 |
| STD.DEV. | 1.60 | 1.70 | 1.45 | 1.31 |
| $Z_N$ | 8.84 | 3.49 | -0.05 | 0.00 |
| $Z_S$ | 2.80 | 3.60 | 1.60 | 0.48 |
| $Z_C$ | 0.27 | -2.54 | 1.64 | 0.99 |
| Chr.2 | | | | |
| AVE. | 5.95 | 4.57 | 4.97 | 5.00 |
| STD.DEV. | 0.98 | 0.99 | 0.69 | 0.61 |
| $Z_N$ | 6.85 | -3.07 | -0.30 | 0.00 |
| $Z_S$ | 9.60 | 9.80 | 3.80 | 2.20 |
| $Z_C$ | -0.28 | -0.24 | 1.53 | -1.59 |

n = 50 for each month.

Table 2. Averages, Std. dev.'s and 3 Z values for product A and B.

Several different analyses can be done with the Z values given in Table 2. $Z_N$ values in the Table for all three characteristics of product A and two characteristics of product B over the four-month period show how the different characteristics of product A and B are doing with respect to their nominal values. As can be seen from Table 2 above, during the first month characteristics of product A and B all have large positive $Z_N$ values. That means, in both product lines process averages of different characteristics are all very much above their nominal. In the second month, the process average for characteristic 3 of product A is closer to its nominal, but still not centered on the nominal (the average is 1.14 standard deviations of the averages above the nominal). As far as product B is concerned, we see some improvement in bringing the averages for characteristics 1 and 2 closer to their nominal.

Similar analysis could be made for the $Z_S$ and $Z_C$ values of products A and B. For example, $Z_S$ values for characteristic 1 of product A are less than –3 during month three and four, indicating that the process is capable of satisfying the relevant specification limits during this time period. However, $Z_S$ values for characteristic 3 of product A are all greater than +3, implying that the process for this characteristic is not capable of meeting the required specifications.

$Z_C$ values, on the other hand, for characteristics 1 and 2 of product A and characteristics 1 and 2 of product B are all within ∀3, indicating that those processes were in control during this four month period. Characteristic 3 of product A, however, has a $Z_C$ value of 3.43 in the fourth month, which implies that the process for that characteristic is out-of-control in that month.

Another application of these Z values would be charting them over time, like a trend chart, to see the patterns in the process management results.

## 4. Conclusion

The three Z values presented in this paper are simple tools for management to make quality comparisons of different product lines, characteristics, etc. Some of the advantages, among many, of these Z values can be listed as:

- Uniform scale so different characteristics can be compared easily (weight, length, diameter, temperature, etc.)
- Ease of interpretation of the Z values
- Ability of comparison of the Z values even without charting them
- Ease of making them a part of any SPC system.

## References

[1] Dixon, W.J. and Massey, F.J. (1969) Introduction to Statistical Analysis, third edition, New York; McGraw-Hill.

[2] Hodley, B. (1981) The quality measurement plan (QMP). The Bell System Technical Journal, 60 (2), 215-273.

[3] Holmes, D.S. (1986) A quality portfolio management chart. Quality, December, 67.

[4] Holmes, D.S. and Mergen, A.E. (1989) A managerial tool for process capability analysis. Quality and Reliability Engineering, International, 5 (2), 143-147.

[5] Holmes, D.S. and Mergen, A.E. (1995) An alternative method to test for randomness of a process. Quality and Reliability Engineering, International, 11(3), 171-174.

## ÖZET

Bu çalışmada idareciye idarenin kalite konusunda ihtiyaç duyduğu noktalarda değişik üretim bantlarındakı kalite performanslarını yapmak için bir metod sunulmuştur. Farklı bantlar, farklı ürünler veya farklı özellikte ayni ürünler üretilebileceyinden ürünlerde kalite karşılaştırmaları genel olarak zordur. Bu makalede önerilen metod makalede tanımlandığı şekilde üç *Z* değerini kullanmıstır (diger yöntemler için örneğin Holmes (1986) ve Holmes ve Mergen (1988)' e bakınız). Ayrıca bir örnek üzerinde önerilen metod tartışılmıştır.

İstatistik, Journal of the Turkish Statistical Association
2000, Vol. 3, No 1-2, pp. 29-39

# ESTIMATION OF MULTIVARIATE PROBABILITY DENSITY FUNCTION WITH KERNEL FUNCTIONS

S.Gökçe Cula and Ö.Toktamış
Hacettepe University, Department of Statistics

### Abstract

In this study, multivariate kernel density estimation has been investigated. Also, the applicability of mutivariate kernel density function, estimation of two variable probability density function whose geometric presentation is possible has been shown by using the earthquake data in Marmara region.

**Key words:** Multivariate density estimation, bandwidth choice, cross-validation, biased validation, bootstrap.

## 1. Introduction

A nonparametric modelling process in multivariate case is more complicated than one in univariate case, in order to determine the structure in data sets studies related to nonparametric probability density estimation are less. Therefore, in recent years it has needed that this issue has been taken account more frequently, in this study kernel density estimation method, which was first studied by Rosenblatt (1956) and Parzen (1962), which has extensive application field in univariate case and whose matematical properties can be investigated very well is studied.

The univariate kernel density estimation has one bandwidth parameter. The specification of more bandwidth parameters than one is required for multivariate density estimation. Also, it has been faced with difficulties in geometric presentation in multivariate case. The multivariate density estimation is the genèralization of univariate case.

The kernel estimation at a given point in one variable case defined as an weighted mean which is calculated by overlapping the mean point of the kernel function with the given point taking account the other observations with weights obtained according to the kernel function and the bandwidth is extended to multivariate case (Toktamiş, 1995).

## 2. Multivariate kernel density estimation

Let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$, denote a $d$-variate random sample having density $f$ , the component of be and the component of $\mathbf{X}_i$ vector be $\mathbf{X}_{i=} (X_{i1}, X_{i2,...,} X_{id})$' and the component of $\mathbf{x}$ vector be $\mathbf{x} = (x_1, x_2, ..., x_n)'$ and $\mathbf{x} \in R^d$. Also, $\int$ notation is shortland for $\int ... \int_{R^d}$, $d\mathbf{x}$ is shortland for $dx_1, dx_2, ..., dx_d$ and the $d \times d$ identity matrix is denoted by $I$. Under these notations, the $d$ dimensional kernel density estimator is given as follows:

$$\hat{f}(\mathbf{x}; \mathbf{H}) = \frac{1}{n\,|\mathbf{H}|^{1/2}} \sum_{i=1}^{n} K_d\left(\frac{\mathbf{x} - \mathbf{X}_i}{\mathbf{H}^{1/2}}\right) \tag{2.1}$$

Where $\mathbf{H}$ is a symetric positive definite $d \times d$ matrix called the bandwidth matrix and $K$ is a $d$-variate kernel functions satisfying $\int_{-\infty}^{\infty} K_d(\mathbf{x})\, d\mathbf{x} = 1$. $K_d$ is usually chosen to be a $d$-variate probability density function. Nevertheless, in order to conctruct a multivariate kernel function from a multivariate kernel function, two common techniques have been followed. One of them is to use (2.2) called product kernel function as a kernel function and the other one is to use (2.3) called spherical kernel function as a kernel function. These equations are given as follows:

$$K_d(\mathbf{x}) = \prod_{i=1}^{d} K(x_i) \tag{2.2}$$

$$K_d(\mathbf{x}) = \frac{K\left\{\left(\mathbf{x}'\mathbf{x}\right)^{1/2}\right\}}{\int K\left\{\left(\mathbf{x}'\mathbf{x}\right)^{1/2}\right\} d\mathbf{x}} \tag{2.3}$$

Another choice which is widely used is to use symmetric unimodal probability density function. The most widely used function for this purpose is standard $d$-variable normal density function which is given as follows:

$$K_d(\mathbf{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\mathbf{x}'\mathbf{x}\right) \tag{2.4}$$

The kernel estimator which is given by (2.1) requires specification of the bandwidth matrix $\mathbf{H}$, which has $\frac{d(d+1)}{2}$ distinct entries. As dimension increases it is getting more difficult to control calculations in which $\mathbf{H}$ matrix is used. In order to simplify $\mathbf{H}$ matrix some restrictions are proposed. Therefore, three situation are considered. The simplest situation corresponds to the restriction $\mathbf{H} \in S$ which means that $\mathbf{H} = h^2 I$ $(h > 0)$ . This restriction, which is to use one constant $h$ bandwidth, means that the amount of smoothing in each direction is the same. This is suitable if the scales of all variables are roughly the same. So this selection can be done only if each variable is standardized to be on a common scale (Simonoff, 1996). Another restriction is to take $\mathbf{H} \in D$, $\mathbf{H} = diag\left(h_1^2, h_2^2, ..., h_d^2\right)$ . This restriction allows different amounts of smoothing in each coordinate direction. This approach is the practical version of restriction $\mathbf{H} \in S$. Let $F$ denote the class of symmetric, positive definite $d \times d$ matrices, $D$ is the subclass of diagonal positive definite $d \times d$ matrices, $D \subseteq F$. When the smoothing in different directions from the direction of coordinates are required, the full bandwidth matrix, $\mathbf{H} \in F$, would be appropriate. In this case as the number of different elements of $\mathbf{H}$ matrix increases that is the number of parameters to be estimated increases. This means that in case of $\mathbf{H} \in F$ kernel density estimation becomes more complicated. Jones and Wand has been done a detailed study related to the three different choices given above, of bandwidth matrix $\mathbf{H}$ which will be used to estimate the bivariate density function (Wand and Jones, 1993).

Under $\mathbf{H} \in D$ , $\mathbf{H} = diag\,(h_1^2, h_2^2, ..., h_d^2)$, multivariate kernel density estimation can be written as follows (Wand and Jones, 1995):

$$\hat{f}\,(x;h) = \frac{1}{n}\left(\prod_{j=1}^{d} h_j\right)^{-1} \sum_{i=1}^{n} K_d\left(\frac{x_1 - X_{i1}}{h_1}, \frac{x_2 - X_{i2}}{h_2}, ..., \frac{x_d - X_{id}}{h_d}\right) \qquad (2.5)$$

In 1996, the equation given above (2.5) was simplyfied by Cacoullos under $\mathbf{H} \in S$ , $S = \{h^2\mathbf{I} : (h > 0)\}$ . This simplyfied form is defined in (2.6):

$$\hat{f}\,(x;h) = \frac{1}{nh^d} \sum_{i=1}^{n} K_d\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right) \qquad (2.6)$$

This selection of bandwidth means that the amount of smoothing is the same in every direction. By using product kernel function as kernel function $K_d$, (2.6) can be rewritten as (Sain et al., 1994):

$$\hat{f}\,(x;h) = \frac{1}{nh^d} \sum_{i=1}^{n} \left\{\prod_{j=1}^{d} K_d\left(\frac{x_j - X_{ij}}{h}\right)\right\} \qquad (2.7)$$

The use of only one bandwidth parameter $h$ in (2.6) shows that scalling the kernel function which is placed at each observation is the same in different coordinate directions. If spread of the data points on one coordinate axis is wider than the other one, then it is necessary to use a different bandwidth for every variable. But, in this situation, it is very difficult to obtain optimal bandwidth from mean integrated squared error, MISE, related to kernel estimation. Because, to make MISE minimum for every $h$ requires very complicated calculations. In the majority of multivariate statistical processes, the data need to be standardized in order to make disapear the difference among the ranges of variables (Wand and Jones, 1993). If the standardization on the variables of multivariate kernel density estimation is carried out, then the equation (2.6) which includes one smoothing parameter is used (Silverman, 1986).

### 3. Asymptotic MISE approximations

In the majority of studies related to the density estimation, the comment about estimation performance is made by measuring the closeness of the estimator to its target value. Rosenblatt stated that the use of MISE which is widely used for the kernel density estimation and easily followed criteria is preferred due to its matematically simpler (Rosenblatt, 1956).

An asymptotic approach for MISE of multivariate kernel density estimation can be obtained in a similar manner to the univariate kernel density estimation. While this approach is obtained, some assumptions as to density function $f$, kernel function $K_d$ and bandwidth matrix $\mathbf{H}$ are given as follows:

i ) Each entry of $Hf(.)$ is piecewise continuous and square integrable;

ii ) $\mathbf{H}=\mathbf{H_n}$ , is a sequence of bandwidth matrices such that $n^{-1}\left|\mathbf{H}\right|^{-1/2}$ and all entries of $\mathbf{H}$ approach zero as $n\to\infty$ ;

iii ) $K_d$ is a bounded, compactly supported $d$-variate kernel satisfying

$$\int_{-\infty}^{\infty} K_d(\mathbf{z})\,d\mathbf{z}=\mathbf{1}\ ,\quad \int_{-\infty}^{\infty}\mathbf{z}K_d(\mathbf{z})\,d\mathbf{z}=\mathbf{0}\ \ \text{ve}\ \int_{-\infty}^{\infty}\mathbf{z}'\mathbf{z}K_d(\mathbf{z})\,d\mathbf{z}=\mu_2(K_d)\,I \qquad (3.1)$$

where, $\mu_2(K_d)=\int_{-\infty}^{\infty} z_i^2K_d(\mathbf{z})\,d\mathbf{z}$ is independent of i.

Under the above assumptions, the asymptotic mean integrated square error of a multivariate kernel density estimator, AMISE, can be obtained as follows:

$$AMISE\left\{\hat{f}(\mathbf{x};\mathbf{H})\right\}=\frac{1}{n}\left|\mathbf{H}\right|^{-1/2}\int_{-\infty}^{\infty}K_d(\mathbf{z})^2\,d\mathbf{z}+\frac{1}{4}\mu_2(K_d)^2\int_{-\infty}^{\infty}tr^2\left\{\mathbf{H}H_f(\mathbf{x})\right\}d\mathbf{x} \qquad (3.2)$$

(Wand and Jones, 1995). Generally, an explicit expression for the AMISE optimal bandwidth matrix of the multivariate kernel density estimator is not available and the numerical value for this quantity can only be obtained by simulation. The most important problem in (3.2), is how the multivariate integrals in (3.2) are evaluated. Nevertheless, it can be obtained for AMISE simpler formula under $\mathbf{H}\in D$ and $\mathbf{H}\in S$ . For example, in the case where $\mathbf{H}=h^2\mathbf{I}$, AMISE of the multivariate kernel density estimator can be obtained as follows (Wand and Jones, 1995):

$$AMISE\left\{\hat{f}(\mathbf{x};\mathbf{H})\right\}=\frac{1}{nh^d}\int_{-\infty}^{\infty}K_d(\mathbf{z})^2\,d\mathbf{z}+\frac{1}{4}h^4\mu_2(K_d)^2\int_{-\infty}^{\infty}\left\{\nabla^2 f(\mathbf{x})\right\}^2d\mathbf{x} \qquad (3.3)$$

where $\nabla^2 f(\mathbf{x})=\sum_{i=1}^{d}\left(\frac{\partial^2}{\partial x_i^2}\right)f(\mathbf{x})$ .

When a specific error criteria, for example AMISE, is fixed at a predetermined value, the required sample size increases rapidly with the number of dimension. The study which is related to this issue was carried out by Scott and Wand (Scott and Wand, 1991). For example, when density function $f$ and kernel function $K_d$ are taken normally distributed with mean zero and variance $I_D$ , the sample sizes necessary to achieve the given AMISE=0.393 have been obtained and it has been given in Table 1 (Simonoff, 1996).

| Dimension | Required Sample Size, $n$ |
|---|---|
| 1 | 4 |
| 5 | 480 |
| 7 | 5382 |
| 10 | 299149 |

Table 1. Sample size required for each dimension to achieve the given AMISE=0.393

## 4. Bandwidth selection

Kernel density estimators are affected from the bandwidth h very much. When the bandwidth $h$ is chosen very small, the variance of estimator increases while the bias of estimator decreases. When bandwidth $h$ is chosen very big, bias increases while variance decreases. Therefore, we have to choose such a bandwidth that the optimal bandwidth is obtained. Some error criteria are used to obtained optimal bandwidth $h$. The optimal bandwidth which makes AMISE minimum for multivariate kernel density estimation given in (3.3) is:

$$h_{AMISE} = \left\{ \frac{d \int\limits_{-\infty}^{\infty} K_d(\mathbf{z})^2 \, d\mathbf{z}}{n\mu_2 (K_d)^2 \int\limits_{-\infty}^{\infty} \left\{ \nabla^2 f(\mathbf{x}) \right\}^2 d\mathbf{x}} \right\}^{1/(d+4)} \tag{4.1}$$

As seen from (4.1) The optimal bandwidth which makes AMISE minimum depends on second derivatives of an unknown density $f$. As this (4.1) involves an unknown function, the optimal bandwidth can not be obtained by using this equation. That's why other methods a few of which are mentioned below are suggested. Some of these methods are given as folows.

### 4.1. Choice of bandwidth for a standard distribution

The aim of this method is to find an expression instead of $\int\limits_{-\infty}^{\infty} \left\{ \nabla^2 f(\mathbf{x}) \right\}^2 d\mathbf{x}$ in (4.1). For this purpose the density function which is known is taken and the optimal bandwidth $h$ is obtained. For example if standard $d$-variate normal density function is used instead of unknown density function $f$, then the integral which was taken place in (4.1) becomes:

$$\int_{-\infty}^{\infty} \left\{\nabla^2 f(\mathbf{x})\right\}^2 d\mathbf{x} = \left(2\sqrt{\pi}\right)^{-d}\left(\frac{1}{2}d+\frac{1}{4}d^2\right)$$

(Silverman, 1986). By substituting $\int_{-\infty}^{\infty} \left\{\nabla^2 f(\mathbf{x})\right\}^2 d\mathbf{x}$ into (4.1) the optimal bandwidth is obtained.

### 4.2. Data-driven methods for bandwidth selections

The least squared cross-validation method, LSCV, for the choice of bandwidth matrix $\mathbf{H}$ is exactly a data driven method. Rudemo and Bowman both separately from each other suggested this method for the choice of bandwidth of kernel estimator in a univariate density function (Rudemo, 1982; Bowman, 1984). This method is generalized to obtain the bandwidth matrix $\mathbf{H}$ in multivariate case. As a result of generalization, least squared cross-validation function (LSCV($\mathbf{H}$)) is found as follows:

$$LSCV(\mathbf{H}) = \int \hat{f}(\mathbf{x};\mathbf{H})^2\, d\mathbf{x} - 2n^{-1}\sum_{i=1}^{n} \hat{f}_{-i}(\mathbf{X}_i;\mathbf{H}) \tag{4.2}$$

where $\hat{f}_{-i}(\mathbf{X}_i;\mathbf{H})$ is the kernel estimator based on the sample with $\mathbf{X}_i$ deleted. Here, the main purpose is to obtain the optimal bandwidth matrix $\mathbf{H}$ which minimizes the expression given by (4.2) and to use this in the multivariate kernel density estimation. Under the restriction $\mathbf{H} \in D$ if the standard normal density function is used instead of kernel function $K_d$, then (4.2) becomes:

$$\begin{aligned} LSCV(h_1,h_2,...,h_d) &= \frac{1}{\left(2\sqrt{\pi}\right)^d n h_1 h_2 ... h_d} + \\ &+ \frac{1}{\left(2\sqrt{\pi}\right)^d n^2 h_1 h_2 ... h_d} \sum_{i=1}^{n}\sum_{j\neq i} \lfloor \exp\left\{-\frac{1}{4}\sum_{k=1}^{d}\left(\frac{x_{ik}-x_{jk}}{h_k}\right)^2\right\} - \\ &- \left(2\times 2^{d/2}\right)\exp\left\{-\frac{1}{2}\sum_{k=1}^{d}\left(\frac{x_{ik}-x_{jk}}{h_k}\right)^2\right\}\rfloor \end{aligned} \tag{4.3}$$

As seen from (4.3) the function LSCV($h_1,h_2,...,h_d$) is a data driven one. The optimal bandwidth which makes (4.3) minimum are obtained and used in kernel density estimation. The least square cross-validation function can have more than one minimum. The studies have shown that the use of the bandwidth which has the largest local minimum is appropriate.

For multivariate case, Sain and his colleques generalized biased cross-validation method, BCV, which is developed by Scott ve Terrell (1987) to obtained the bandwidth $h$ in a univariate kernel density estimation. Sain and his colleagues used the estimation $\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i}^{iv}(x_i)$ which has smaller bias instead of $\int f''(x)^2\,dx$ (Sain et al., 1994). They used standard normal density function instead of kernel function $K_d$ under the restriction $\mathbf{H} \in D$ and they found biased cross-validation function, (BCV$(h_1, h_2, ..., h_d)$), for multivariate product kernel estimation as follows :

$$\begin{aligned} BCV(h_1,h_2,...,h_d) \; = \; & \frac{1}{\left(2\sqrt{\pi}\right)^d n h_1 h_2 ... h_d} + \\ & + \frac{1}{4n(n-1)h_1 h_2 ... h_d}\sum_{i=1}^{n}\sum_{j\neq i}\lfloor\left\{\sum_{k=1}^{d}\left(\frac{x_{ik}-x_{jk}}{h_k}\right)^2\right\}^2 - \\ & -(2d+4)\left\{\sum_{k=1}^{d}\left(\frac{x_{ik}-x_{jk}}{h_k}\right)^2\right\} + \left(d^2+2d\right)\rfloor \times \\ & \times\prod_{k=1}^{d}\Phi\left(\frac{x_{ik}-x_{jk}}{h_k}\right) \end{aligned} \tag{4.4}$$

where $\Phi$ is standard normal density function. To obtain the optimal bandwidth, they found the bandwidth which minimizes (4.4) (Sain et al., 1994).

They also developed bootstrap method which is used by Taylor (1989) to find the bandwidth $h$ in a univariate case for multivariate density estimation. MISE for multivariate case is:

$$MISE(\mathbf{h}) = \int_{R^d} E^*\left\{\hat{f}^*(\mathbf{x}) - \hat{f}(\mathbf{x})\right\}^2 d\mathbf{x}$$

where $\hat{f}(\mathbf{x})$ is the multivariate kernel estimator, $\hat{f}^*(\mathbf{x})$ is a multivariate kernel estimator calculated with data sample from $\hat{f}(\mathbf{x})$, and the expectation, $E^*$, is taken with respect to the density $\hat{f}(\mathbf{x})$. Under restriction $\mathbf{H} \in D$ by using the standard normal kernel function for $K_d$, bootstrap function, (B$(h_1, h_2, ..., h_d)$), for multivariate product kernel estimation as follows :

$$\begin{aligned} B(h_1,h_2,...,h_d) \; = \; & \frac{1}{\left(2\sqrt{\pi}\right)^d n h_1 h_2 ... h_d} + \\ & + \frac{1}{\left(2\sqrt{\pi}\right)^d n^2 h_1 h_2 ... h_d}\sum_{i=1}^{n}\sum_{j\neq i}\lfloor\frac{n-1}{2^{d/2}}\exp\left\{-\frac{1}{8}\sum_{k=1}^{d}\left(\frac{x_{ik}-x_{jk}}{h_k}\right)^2\right\} + \\ & + \exp\left\{-\frac{1}{4}\sum_{k=1}^{d}\left(\frac{x_{ik}-x_{jk}}{h_k}\right)^2\right\} - \frac{2\times 2^{d/2}}{3^{d/2}}\exp\left\{-\frac{1}{6}\sum_{k=1}^{d}\left(\frac{x_{ik}-x_{jk}}{h_k}\right)^2\right\} \end{aligned} \tag{4.5}$$

(Sain et al., 1994). The bandwidths which minimize this function are taken and these values are used to obtain multivariate kernel density estimation.

## 5. Application

In this study firstly the computer programs have been written to obtain the values of kernel estimates of multivariate density function since no standard computer programs available. Secondly some computer programs have been written for some methods (LSCV, BCV, B) which have developed in order to obtain the bandwidths. The programs have been coded by Delphi 3 for methods (Cula, 1998).

In this application, the data are consist of 255 earthquakes with magnitude at least 4 on Richter scale occurring between 1900-1999 years in Marmara region. This data was used to obtain kernel estimation of bivariate probability density function. Here, bivariate standard normal density function is used as a kernel function. As two variables were measured by using the same scale, the raw data were used without making any standardisation. In this simulation firstly, the bandwidth $h$ was increased by 0.01 between 0 and 2 and then the values of the functions $\mathrm{BCV}(h_1, h_2, ..., h_d)$, $\mathrm{LSCV}(h_1, h_2, ..., h_d)$ and $\mathrm{B}(h_1, h_2, ..., h_d)$ were found and their distribution was obtained.

The bandwidth $h$ which makes the function $\mathrm{BCV}(h_1, h_2, ..., h_d)$ minimum was obtained as 1.36 and the value of the function $\mathrm{BCV}(h_1, h_2, ..., h_d)$ which corresponds to this value was obtained as 0.02234. The graph of the function $\mathrm{BCV}(h_1, h_2, ..., h_d)$ which corresponds to the bandwidths were given in Figure 1.

The bandwidth $h$ which makes the function $\mathrm{BCV}(h_1, h_2, ..., h_d)$ minimum is substituted into the bivariate kernel density estimator, obtaining the value of estimation. Figure 2 gives surface plot and contour plot of the kernel estimate for the earthquake data.

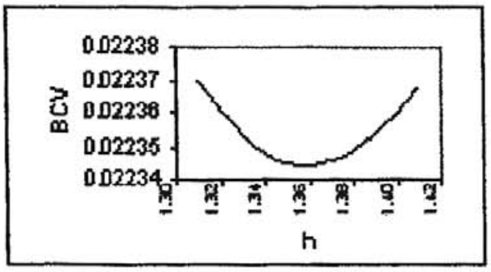


Figure 1: The graph of the function $\mathrm{BCV}(h_1, h_2, ..., h_d)$ for the bandwidth's value between 1.30-1.42 for the earthquake data set

Optimal bandwidth $h$ value couldn't be obtained from cross-validation and bootstrap methods. The graph related to this is given in Figure 3.

As seen from Figure 3, as $h$ increases the value of the function $\mathrm{LSCV}(h_1, h_2, ..., h_d)$ also increases and as$h$ increases the value of the function $\mathrm{B}(h_1, h_2, ..., h_d)$ decreases. However, minimum value couldn't be obtained for both of the functions, in other words optimal bandwidth couldn't be found.

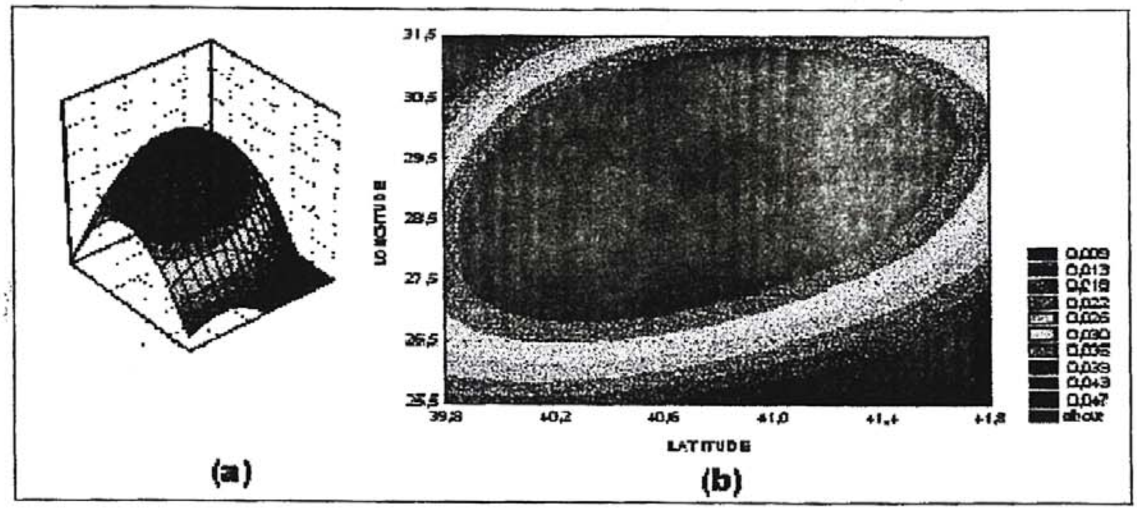


Figure 2: The graphs of the bivariate kernel density estimation values related to the earthquake data for the Marmara region when $h$=1,36 a)Surface Plot b)Contour Plot

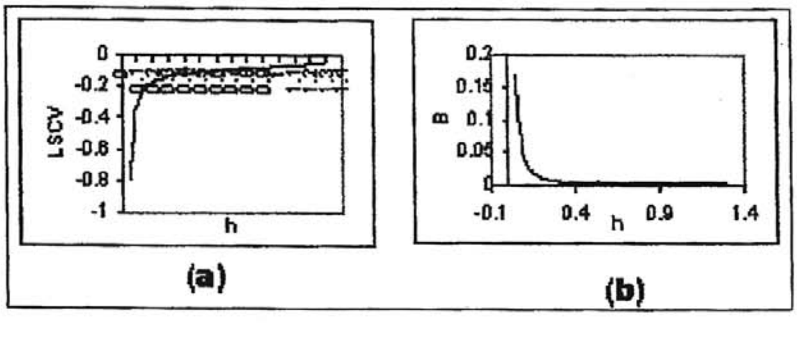


Figure 3: The graph of a) the function LSCV$(h_1, h_2, ..., h_d)$, b) the function B$(h_1, h_2, ..., h_d)$ for the bandwidth's value between 0.1-1.4 for the earthquake data set

## 6. Conclusion

In the application of earthquake data the bandwidth value which was obtained by using the method BCV has been found as 1.36. It couldn't be obtained the bandwidth value for both of the methods LSCV and B. By putting the bandwidth which is obtained from method LSCV into the bivariate kernel density function, the estimation values have been calculated and the graphs related to these values have been drawn (Figure 2). According to the data of 255 earthquakes with magnitude at least 4 on the Richter scale occurring between 1900-1999 years in Marmara region, the density related to the observation between longitude 40.30-41.20 and between latitude 27.75-30.40 is found the highest. It can be said that Istanbul, Izmit, Yalova cities which fall into these coordinates have higher probability of occurring earthquake with magnitute at least 4 on the Richter scale in Marmara region than the other places in Marmara region.

## References

[1] Bowman, A. W. (1984) An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, 71, 2, 353-360.

[2] Cula, S.G. (1998) Çok Değişkenli Olasılık Yoğunluk Fonksiyonunun Çekirdek Fonksiyonlari ile Kestirimi, Doktora Tezi, H.Ü., Fen Faktültesi, Ankara, 1212.

[3] Parzen, E. (1962) On the estimation of a probability density function and the mode, *Ann. Math Statistics*, 33, 1065-1076. Royal Stat. Soc. Ser B 34, 385-392.

[4] Rosenblatt, M. (1956) Remarks on some non-parametric estimates a density function, *Annals Math. Statist.*, 27, 832-837.

[5] Rudemo, M. (1982) Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* 9, 65-78.

[6] Sain, R.S., Scott,D.W. and Baggerly, K.A. (1994) Cross-validation of multivariate densities, *Journal of the American Statistical Association*, vol.89, No.427, 807-817.

[7] Scott, D. W. and Terrell, G. R. (1987) Biased and unbiased cross-validation in density estimation, *Journal of the American Statistical Association*, Vol.82, No.400, 1131-1146.

[8] Scott, W. D. and Wand, M. P. (1991) Feasibility of multivariate density estimates, *Biometrika*, 78, 1, 197-205.

[9] Silverman, B. W. (1986) Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.

[10] Simonoff, J. S. (1996) Smoothing Methods in Statistics, Verlag, New York.

[11] Taylor, C. C. (1989) Bootstrap choice of the smoothing parameter in kernel density estimation, *Biometrika*, 76, 4, 705-712.

[12] Toktamış, Ö. (1995) Olasılık yoğunluk fonksiyonunun çekirdek kestirimi üzerine bir çalişma, Hacettepe Fen ve Mühendislik Bilimleri Dergisi, 16, 145-163.

[13] Wand, M. P. and Jones, M. C. (1993) Comparison of smoothing parameterization in bivariate kernel density estimation, *Journal of the American Statistical Association*, vol.88, No.422, 520-528.

[14] Wand, M. P. and Jones, M. C. (1995) Kernel Smoothing, Chapman and Hall, New York.

## ÖZET

Bu çalışmada, çok değişkenli olasılık yoğunluk fonksiyonunun çekirdek kestirim yöntemi incelenmiştir. Ayrıca çok değişkenli olasılık yoğunluk fonksiyonunun uygulanabilirliği, geometrik gösterimlere olanak sağlayan iki değişkenli olasılık yoğunluk fonksiyonunun kestirimi Marmara Bölgesi için elde edilen deprem verileri kullanılarak yapılan uygulama ile gösterilmiştir.

İstatistik, Journal of the Turkish Statistical Association
2000, Vol. 3, No 1-2, pp. 40-46

# MULTIPARAMETER ESTIMATION IN TRUNCATED POWER SERIES

N. Sanjari Farsipour and A. Jahedi
Department of Statistics Shiraz University Shiraz, Iran

### Abstract

Let $X_1, ..., X_p$ be $(p > 1)$ independent random variables, where each $X_i$ has a distribution belonging to one parameter truncated power series distribution. The problem is to estimate simultaneously the unknown parameters under the weighted error loss function.

**Key words:** Power series, truncated Poisson distribution.

## 1. Introduction

Recently, there has been considerable interest in simultaneous estimation of parameters from several independent distributions other than Gaussian. Since celebrated work of James and Stein (1961), numerous results have been obtained for the simultaneous estimation problems under several plausible loss. Hwang (1982) obtained improved estimators in discrete exponential families, Tsui (1986) showed robustsness of Clevenson-Zidek (1975) type estimators when the underlying distributions belong to much larger class of distributions. Recently, Dey, Ghosh and Srinivasan (1987) considered the loss developed by Stein for simultaneous estimation of $p$ independent gamma scale parameters and their reciprocals.

This paper is devoted to simultaneous estimation of parameters of a truncated distribution under a squared error loss function given as

$$L(\theta_i, \delta_i) = \sum_{i=1}^{n} \left( \frac{\delta_i}{\theta_i} - 1 \right)^2 . \tag{1.1}$$

Let $X = (X_1, ..., X_p)$ where $X_1, ..., X_p$ are $p$ independent random variables, each $X_i$ having probability function.

$$P_{\theta_i}(x_i) = g_i(\theta_i) t_i(x_i) \theta_i^{x_i} \quad x_i = a_i,\ a_i + 1, ... \tag{1.2}$$

where $a_i$ is a nonzero positive integer and $g_i(\theta_i)$ is a normalizing constant, given as

$$g_i^{(-1)}(\theta_i) = \sum_{x_i=a_i}^{\infty} t_i(x_i)\theta_i^{x_i} \quad \theta_i > 0, \, i = 1, 2, ..., p. \tag{1.3}$$

Such a distribution will be refered to as a power series distribution truncated at left. Special cases include the truncated Poisson and the truncated negative binomial distributions. The problem is to simltaneously estimation of $\theta = (\theta_1, ..., \theta_p)$ under the loss (1.1). For the loss (1.1), the best multiple estimator of $\theta$ (which is also the best unbiased estimator) is given by $\delta(X) = \left(\delta_1^0(X), ..., \delta_p^0(X)\right)$ where

$$\delta_i^0(x) = \begin{cases} \frac{t_i(x_i-1)}{t_i(x_i)} & x_i = a_i + 1, a_i + 2, ... \\ 0 & otherwise. \end{cases} \tag{1.4}$$

For proving the unbiasedness of $\delta_i^0(X)$ note that

$$E\left(\delta_i^0(X)\right) = E\left(\frac{t_i(X_i - 1)}{t_i(X_i)}\right)$$

$$= \sum_{x_i=a_i}^{\infty} \frac{t_i(x_i - 1)}{t_i(x_i)} f_{\theta_i}(x_i)$$

$$= \sum_{x_i=a_i}^{\infty} \frac{t_i(x_i - 1)}{t_i(x_i)} f_{\theta_i}(x_i) \frac{1}{\sum\limits_{x_i=a_i}^{\infty} t_i(x_i)(\theta_i)^{x_i}} t_i(x_i)\theta_i^{x_i}$$

$$= \frac{\sum\limits_{x_i=a_i}^{\infty} t_i(x_i - 1)(\theta_i)^{x_i}}{\sum\limits_{x_i=a_i}^{\infty} t_i(x_i)(\theta_i)^{x_i}}$$

$$= \frac{t_i(a_i - 1)(\theta_i)^{x_i} + \theta \sum\limits_{x_i=a_i+1}^{\infty} t_i(x_i - 1)(\theta_i)^{x_i}}{\sum\limits_{x_i=a_i}^{\infty} t_i(x_i)(\theta_i)^{x_i}}$$

and $t_i(a_i - 1)$ is defined as zero then $E(\delta_i^0(X)) = \theta_i$. It follows from Brown and Hwang (1982) and Ghosh and Yang (1988) that $\delta_i^0$ is admissible for $p = 1$. Thus for $p > 2$, we propose the rival estimator of $\theta$ as $\delta(X) = \delta^0(X) + \phi(X)$ where $\phi(x) = (\phi_1(x), ..., \phi_p(x))$. Also, assume that $\phi_i(x) = 0$ if $x_i < a_i + 1$, $i = 1, ..., p$.

In section 2 a difference inequality involving the risk differences of imoroved estimators from the unbiased estimator is obtained and solved. In section 3 we obtain two classes of dominating estimators.

## 2. Obtaining the difference inequality

Let $x = (x_1, ..., x_p)$ be a vector of observations of the random vector $X = (X_1, ..., X_p)$, where the $X_i$'s, $i = 1, ..., p$ are mutually independent random variables with probability function $P_{\theta_i}(x_i)$ as given in (1.2). Then for any real-valued function $\phi_i(x)$ where $E\,|\phi_i(X)|$ is finite, and $\phi_i(x) = 0$ if $x_i < a_i + 1$ the following identity (Hwang 1982) holds

$$E(\theta_i^{-1}\phi_i(X)) = E(\phi_i(X + e_i))/\delta_i^0(X + e_i) \tag{2.1}$$

**Theorem 2.1.** Let $\phi_i$ and $t_i$ be defined previously then we have the following equality

$$E(\theta_i^{-2}\phi_i^2(X)) = E(\frac{\phi_i^2(X + 2e_i)t_i(X + 2)}{t_i(X_i)}) \tag{2.2}$$

**Proof.**

$$E(\theta_i^{-2}\phi_i^2(X)) = \sum_{x_i=a_i}^{\infty} \theta_i^{-2}\phi_i^2(x)\left[\frac{t_i(x_i)\theta_i{}^{x_i}}{\sum\limits_{x_i=a_i}^{\infty} t_i(x_i)\theta_i{}^{x_i}}\right]$$

$$= \sum_{x_i=a_i}^{\infty} \phi_i^2(x)\frac{t_i(x_i)\theta_i{}^{x_i-2}}{\sum\limits_{x_i=a_i}^{\infty} t_i(x_i)(\theta_i)^{x_i}}$$

$$= \sum_{y_i=a_i-2}^{\infty} \phi_i^2(y + 2e_i)\frac{t_i(y_i + 2e_i)\theta_i{}^{y_i}}{\sum\limits_{y_i=a_i-2}^{\infty} t_i(y_i + 2)(\theta_i)^{y_i+2}}$$

$$= \sum_{y_i=a_i-2}^{\infty} \frac{\phi_i^2(y + 2e_i)t_i(y_i + 2e_i)}{t_i(y_i)}\frac{t_i(y_i)\theta_i{}^{y_i}}{\sum\limits_{y_i=a_i-2}^{\infty} t_i(y_i + 2)(\theta_i)^{y_i+2}}$$

$$= E\left[\frac{\phi_i^2(X_i + 2e_i)t_i(X_i + 2e_i)}{t_i(X_i)}\right]$$

We know $t(a_i - 2) = t(a_i - 1) = 0$ and $\phi(x_i) = 0$ when $x_i < a_i + 1$ and we are done.

**Theorem 2.2.** Let $\phi_i$ and $t_i$ and $\delta$ be as defined previously then we have the following equality

$$E(\theta_i^{-2}\phi_i^2(X)\delta_i^0(X)) = E(\frac{\phi_i(X_i + 2e_i)t_i(X_i + e_i)}{t_i(X_i)}) \tag{2.3}$$

**Proof.**

$$E(\theta_i^{-2}\phi_i^2(X)\delta_i^0(X)) = \sum_{x_i=a_i}^{\infty} \theta_i^{-2}\phi_i(x)\delta_i^0(x)\frac{t_i(x_i)\theta_i^{x_i}}{\sum\limits_{x_i=a_i}^{\infty} t_i(x_i)\theta_i^{x_i}}$$

$$= \sum_{x_i=a_i}^{\infty} \phi_i(x)\delta_i^0(x)\frac{t_i(x_i)\theta_i^{x_i-2}}{\sum\limits_{x_i=a_i}^{\infty} t_i(x_i)(\theta_i)^{x_i}}$$

$$= \sum_{y_i=a_i-2}^{\infty} \phi_i(y+2e_i)\delta_i^0(y+2)\frac{t_i(y_i+2e_i)\theta_i^{y_i}}{\sum\limits_{y_i=a_i-2}^{\infty} t_i(y_i+2)(\theta_i)^{y_i+2}}$$

$$= \sum_{y_i=a_i-2}^{\infty} \frac{\phi_i(y+2e_i)t_i(y_i+2e_i)\delta_i^0(y+2)}{t_i(y_i)}\frac{t_i(y_i)\theta_i^{y_i}}{\sum\limits_{y_i=a_i-2}^{\infty} t_i(y_i+2)(\theta_i)^{y_i+2}}$$

$$= E(\frac{\phi_i(X_i+2e_i)t_i(X_i+e_i)}{t_i(X_i)})$$

which completes the proof. Now suppose that $\delta(x) = \delta^0(x) + \phi(x)$ is an estimator of $\theta$, where $\phi(x) = (\phi_1(x), ..., \phi_p(x))$ and $\phi_i^{'}s$ satisfy (2.1), (2.2), (2.3). The following theorem gives an unbiased estimator of risk difference of $\delta(x)$ and $\delta^o(x)$.

**Theorem 2.3.** If $\Delta(\theta) = R(\delta^*, \theta) - R(\delta^0, \theta) = E_\theta\Delta(X)$ is the risk difference, then the unbiased estimator of the risk difference is given as

$$\Delta(x) = \sum_{i=1}^{p} [\psi_i^2(x+2e_i)\frac{\delta_i^0(x+2)}{\delta_i^0(x+1)} + 2\psi_i(x+2e_i)\frac{\delta_i^0(x+2)}{\delta_i^0(x+1)} - 2\psi_i(x+e_i)]$$

so that $\psi_i(x) = \frac{\phi_i(x)}{\delta_i^0(x)}$, $i = 1, ..., p$.

**Proof.**

$$\begin{aligned}\Delta(\theta) &= E_\theta\Delta(X) \\ &= R(\delta^*, \theta) - R(\delta^0, \theta) \\ &= E(L(\delta^*, \theta) - L(\delta^0, \theta)) \\ &= E[\sum_{i=1}^{p}(\frac{\delta^*(X)}{\theta_i} - 1)^2 - (\frac{\delta^0(X)}{\theta_i} - 1)^2]\end{aligned}$$

$$= E[\sum_{i=1}^{p}(\frac{\delta^0(X) + \phi_i(X)}{\theta_i} - 1)^2 - (\frac{\delta^0(X)}{\theta_i} - 1)^2]$$

$$= E[\sum_{i=1}^{p}(\frac{\phi_i^2(X)}{\theta_i^2} - 2\frac{\phi_i(X)}{\theta_i} + 2\frac{\delta_i^0(X)\phi_i(X)}{\theta^2})]$$

$$= \sum_{i=1}^{p}[E(\phi_i^2(X)\theta_i^{-2}) - 2E(\phi_i(X)\theta_i^{-1}) + 2E(\delta_i^0(X)\phi_i(X)\theta^{-2})]$$

$$= \sum_{i=1}^{p} E[\frac{\phi_i^2(X + 2e_i)t_i(X_i + 2)}{t_i(X_i)} - 2\frac{\phi_i(X + e_i)}{\delta_i^0(X + e_i)} + 2\frac{\phi_i(X + 2e_i)t_i(X_i + 1)}{t(X_i)}]$$

$$= E[\sum_{i=1}^{p}\frac{\phi_i^2(X + 2e_i)}{\delta_i^0(X + 2)\delta_i^0(X + 1)} + 2\frac{\phi_i(X + 2e_i)}{\delta_i^0(X + 1)} - \frac{\phi_i(X + e_i)}{\delta_i^0(X + 1)}]$$

$$= E[\sum_{i=1}^{p}[\frac{\phi_i^2(X + 2e_i)\delta_i(X + 2)}{\delta_i^{0^2}(X + 2)\delta_i^0(X + 1)} + 2\frac{\phi_i(X + 2e_i)\delta_i^2(X + 2)}{\delta_i^0(X + 2)\delta_i^0(X + 1)} - 2\frac{\phi_i^2(X + e_i)}{\delta_i^{0^2}(X + 1)}]]$$

So

$$\Delta(x) = \psi_i^2(x + 2e_i) \cdot \frac{\delta_i^0(x + 2)}{\delta_i^0(x + 1)} + 2\psi_i(x + 2e_i) \cdot \frac{\delta_i^0(x + 2)}{\delta_i^0(x + 1)} - 2\psi_i(x + 2e_i)$$

which completes the proof.

## 3. Classes of improved estimators

In this section we will obtain two classes of dominating estimators. The following theorem gives a class of shirinkage estimators when the dimension is more than 3.

**Theorem 3.1:** Consider the rival estimator $\delta(x) = \delta^0(x)(1 + \psi(x))$ where $\psi(x) = (\psi_1(x), \psi_2(x), ..., \psi_p(x))$ with

$$\psi_i(x) = \frac{c(x)x_i}{b + S_1}, i = 1, ..., p \tag{3.1}$$

and $S_1 = \sum_{j=1}^{p} x_j^2$. Suppose further the following conditions hold:

i) $C(x)$ is nondecreasing in each coordinate

ii) $0 < C(x) \leq \min(p - 3, \frac{\sqrt{b}}{2})$

iii) $b \geq 4p$.

Then for $p \geq 4, \delta(x)$ will dominate $\delta^0(x)$ in terms of risk given $\psi_i(x) \geq (9/32) . \frac{\delta_i^0(x+1)}{\delta_i^0(x)}$

**Proof.** Assign to the Dey and Chung (1991) we know $|\psi_i(x)| < \frac{c(x)}{2\sqrt{b}} \leq (1/4)$ then

$$\Delta(x) = \sum_{i=1}^{p} [\psi_i^2(x+2e_i)\frac{\delta_i^0(x+2)}{\delta_i^0(x+1)} + 2\psi_i(x+2e_i)\frac{\delta_i^0(x+2)}{\delta_i^0(x+1)} - 2\psi_i(x+e_i)]$$

So if $A_i = \frac{\delta_i^0(x+2)}{\delta_i^0(x+1)}$ then

$$\Delta(x) \leq \sum_{i=1}^{p} ((1/16)A_i + 2(1/4)A_i - 2\psi_i(x+e_i))$$

$$\Delta(x) \leq \sum_{i=1}^{p}((9/16)A_i - 2\psi_i(x+e_i))$$

For $\Delta(x) < 0$ it is necessary that $(9/16)A_i - 2\psi_i(x+e_i) \leq 0$

$$\psi_i(x+e_i) > (9/32)A_i = (9/32)\frac{\delta_i^0(x+2)}{\delta_i^0(x+1)}$$

then

$$\psi_i(x) \geq (9/32)A_i = (9/32)\frac{\delta_i^0(x+1)}{\delta_i^0(x)}$$

and

$$\psi_i(x) = \frac{c(x)x_i}{b+S_i} > (9/32)\frac{\delta_i^0(x+1)}{\delta_i^0(x)}$$

and

$$c(x) > (9/32)\frac{\delta_i^0(x+1)}{x_i\delta_i^0(x)}(b+s_1)$$

So $\psi_i(x) < 1/4$ then

$$\frac{c(x)x_i}{b+S_1} < 1/4$$

or

$$(9/32)\frac{\delta_i^0(x+1)}{x_i\delta_i^0(x)}(b+S_1) < C(x) < (1/4)\frac{b+S_1}{x_i}$$

**Example 3.1.** Let $X_i$'s be independently distributed with probability function

$$P_\theta(x_i) = \frac{e^{-\theta_i}\theta_i^{x_i}}{x_i!}g(\theta_i)$$

$$= \frac{e^{-\theta_i}\theta_i^{x_i}}{x_i!}\frac{1}{\sum_{i=1}^{\infty}\frac{e^{-\theta_i}\theta_i^{x_i}}{x_i!}} \qquad x_i = a_i+1, a_i+2, ...$$

So that $\delta_i^0(x) = \frac{t_i(x_i-1)}{t_i(x_i)} = x_i$ we know that $|\psi_i(x)| < 1/4$ and $\psi_i(x) > (9/32)\frac{\delta_i^0(x+1)}{\delta_i^0(x)}$ then

$$(9/32)\frac{\delta_i^0(x+1)}{\delta_i^0(x)} < \psi_i(x) < 1/4$$

or

$$(9/32)\frac{x_i+1}{x_i} < \psi_i(x) < 1/4$$

or

$$\frac{c(x)x_i}{b+S_1} > (9/32)\frac{x_i+1}{x_i}$$

or

$$C(x) > (9/32)\frac{x_i+1}{x_i^2}(b+S_1)$$

or $\psi_i(x) < 1/4$ then $\frac{c(x)x_i}{b+S_1} < 1/4$ or $C(x) < (1/4)\frac{b+S_1}{x_i}$.

## References

[1] Clevenson, M.L. and Zidek, J.V (1975). Simultaneous estimation of the independent poisson laws. *J. Amer. Statist. Assoc.*, 70, 698-705.

[2] James, W. and Stein, C. (1961). Estimation with quadratic loss. Proceeding Fourth Berkley symposium on Mathematics, Statistics and Probability, 1, 316-379, University of California Press.

[3] Hwang, J.T. (1982). Improving upon standard estimators in discrete exponential families with application to Poisson binomial laws. *Ann. Statist.*,10, 857-867.

[4] Tsui, K.W. (1986). Further developments on the robustness of Clevenson-Zidek type means estimator. *J. Amer. Statist. Assoc.*, 81,176-180.

[5] Dey, D.K. and Chung, Y. (1991). Multiparameter estimation in truncated power series distributions under the stein's loss. *Commun. Statist. Theory Meth.*, 20(1), 309-326.

## ÖZET

Kesilmiş kuvvet serisi dağılımından elde edilmiş örneklemin yardımı ile ağırlandırılmış hata fonksiyonları kullanılarak parametrelerin tahmini yapılıyor.

İstatistik, Journal of the Turkish Statistical Association
2000, Vol. 3, No 1-2, pp. 47-57

# THE INTERACTION BETWEEN ECONOMETRICS, INFORMATION SYSTEMS AND STATISTICAL INFRASTRUCTURES:
## Anticipation and Comparative Analysis in a Decisional Structure

Orhan Güvenen
Institute of World Systems, Economics and Strategic Research
Bilkent University, Ankara, Turkey

## ABSTRACT

The aim of this paper is to show how econometrics, information systems, and statistical infrastructures are independent, and can be viewed together as a means to optimally impact the decisional structure in any given society. This paper also aims to formulate some ideas on econometrics towards the coming years as a tool for analyzing international interdependency.

With the advent of the information age, tremendous amounts of highly disaggregated information flows have had an increased influence over the social sciences both theoretically and empirically, as well as over the decision making structures. This necessitates a reexamination of the existing economic theories in order to accurately reflect today's realities. In this sense, econometrics can be used as the primary tool in terms of the quantitative aspect for increasing the value added to economic theory and the social sciences as a whole.

Additionally, this reexamination should take advantage of the advances in information systems and in statistical infrastructures. To the extent that information is received in a timely and thorough manner, and is utilized through statistical analysis, decisional structures will succeed in achieving their targets in a rapidly changing world environment.

In conclusion, this paper sets out to show that the intense flow of information and the improvement in quantification techniques will have a strong impact on the social sciences. Furthermore, there will be a corresponding revision of the theoretical aspect of the social sciences due to the strong link between the empirical and theoretical aspects of the social sciences. This will be essential because the large new flows of disaggregated information and wide use of quantification techniques and data processing, assisted by information technology, will render the existing theories insufficient. The reformulation of the theories will be based on the increased information available, and will be based on an interdisciplinary approach. This will necessitate the reform of the existing theories and quantification techniques.

## 1. Introduction

Over the past few decades, the world has been experiencing a phenomenon which has seen the proliferation of information at exponential rates. Thus, this era has been appropriately termed the "information age", and has been likened in its significance to the Industrial Revolution in terms of its impact on the entire *modus operendi* of the global system.

The challenge at hand is how to utilize or design information systems that will appropriately and efficiently optimize the availability of information in a manner that will most favorably impact the decisional structure of policy makers. The most effective means for doing so is through the extensive coordination of econometrics, information systems, and statistical infrastructures.

As each phase of technological advancement propels the reverberations of change throughout the global system, the existing static models for installing and accessing information systems, which allow decision makers to know their alternatives and to make decisions on the most up to date information, become insufficient.

The aim of this paper is to review the interaction between econometrics, information systems, and statistical infrastructures and to demonstrate the need for further examination of this interaction so as to create a dynamic model of information processing, analyzing and accessibility. The ultimate goal is to produce an optimal coordination of the aforementioned components so that decision makers can function most productively in the global system to the benefit of all the countries, institutions, and individuals within this system.

This paper also aims to formulate some ideas on econometrics towards the coming years as a tool for analyzing international interdependency. Finally, one additional aim of this paper is to demonstrate how the existing theories will be rendered insufficient due to the large flows of disaggregated information and the wide use of quantification techniques and data processing, assisted by information technology. This will necessitate the reform of the existing theories and quantification techniques in the social sciences.

## 2. Some Comments on the Specification Methodology of Social and Economic Phenomena

The specification of social and economic phenomena presents numerous and substantial difficulties arising from biases in the specification methodology and from measurement problems.

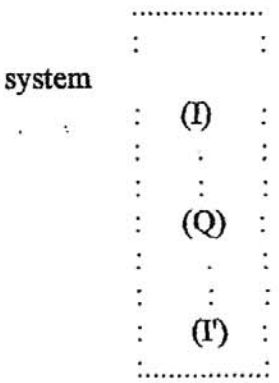
In particular, the uniqueness of social and economic phenomena, in the sense of taking place at a particular point in history, geography, and society, makes it difficult to undertake controlled experiments as in the natural sciences. The problem is compounded further due to the close relationship between human behavior and the existing social and economic phenomena.

Individuals, or decision makers, are affected by the existing social and economic phenomena when making decisions. Once decisions are made and acted upon, the social and economic phenomena that form the basis of future decisions change. Additionally, the continuous structural changes taking place due to the dynamic nature of social and economic phenomena add to the difficulties in making generalizations.

To alleviate these problems, the analysis and quantification of any social and economic phenomena requires the understanding of the global structure which is of one of a chaotic, complex, and stochastic nature.

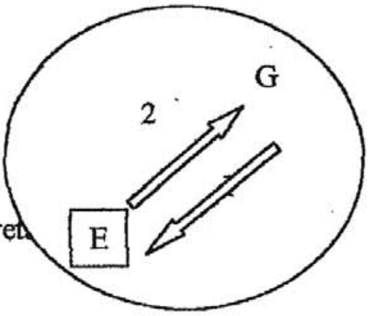
As shown in Diagrams 1 and 2, the nature of specification methodology may limit our understanding of social and economic phenomena. As Diagram 1 implies, a certain information set is collected as is determined by our own perceptions of reality or by the paradigm in which we find ourselves. Our biases in quantification and in quantitative interpretation put further restrictions on our understanding of social and economic phenomena.

**Diagram 1**



: I : An Information set
: Q : Quantification
: I' : Quantitative Interpret

**Diagram 2**



G : Global

E:Economics

Within the existing mathematical tools currently being utilized , the global structure remains indefinable. What actually occurs then in analyzing and quantifying social and economic phenomena is a subset approach which is a partial analysis of the phenomena disconnected form the global structure (Diagrams 3 and 4)

**Diagram 3**

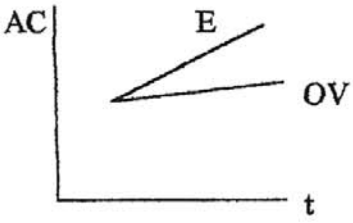
(X) analogy (X')

Stochastic phenomenon Model

**Diagram 4**

Specification——Estimation — Use

AC : Alternative cost



E: Impact of E type specification variable

OV: Impact of other variables

At the first stage, analysis and understanding of the global structure is processed within an interdisciplinary approach ; then , via this information , the elaboration of any specification at the structural level occurs. This approach would lead to a different specification, which represents less alternative costs in the medium and long terms, compared to a specification, which limits itself to the structural level.

However, even if our specification methodology is free of the problems mentioned above, certain specification problems arising from the lack of data or from mis-measurement problems exist because the specification methodology is determined independently of the data collection process. Thus , a social scientist has to rely on a data base which answers to the needs of governments and businesses. That is , some data required by the specification methodology "may not be available in published form , or may not exist at all" (Johnson,1991). The best example of this is expectation data. If expectations play an important role in our specifications, then the lack of data on expectations poses a great difficulty.

In this respect, this problem might be overcome to some extent if the data base creation process, or if institutions, pay attention to the needs of different specification methodologies as well as to the needs of the governments and businesses.

## 3. The Link Between Economic Theory, Information Theory, Econometrics , Information Systems , and Statistical Infrastructures

In this section , the roles of econometrics , information system and statistical infrastructures in particular are reviewed. This section demonstrates why the link between these three areas is so important to structural efficiency in the decision making process , and therefore to the global system.

### 3.1. Information Theory

The decision making process depends upon receiving timely and accurate information. Furthermore , the decision making process is only optimized depending upon the value of the information circulating throughout the organization and to the decision makers. Information theory is a framework approach that allows quantification measures to be applied to information systems so as to quantify the information in a meaningful way for the decision makers. The emphasis of information theory is on facilitating the information receiving the users, as opposed to data processing, which is concerned with processing data almost irrespective of the users ability to use that data.

### 3.2. Econometrics and Economic Theory

The economic issues that are studied naturally relate to that area within the social sciences that is concerned with the description and analysis of production, distribution, and consumption of goods and services within a society. While econometrics has traditionally been limited to the field of economics, it can in fact be considered as a quantification method for all of the social sciences. The validity of this claim derives from the fact that social sciences are concerned primarily with human society or with individuals as members of that society. All social sciences use statistical data to assess or project models measuring various aspects of such societies. Statistics scientifically reinforce the theories or models of society that are being studied, and in this sense they provide a system of measurement used to understand the status of societies.

Thus, the embetterment of society, which necessitates the analysis of social data, requires that a quantitative methodology be employed to determine trends and the effectiveness of policy. Specifically, econometrics is constantly used as the means to bring a value added to not only economics, but to the whole of social sciences by adding the irreplaceable value of scientific measurement.

Indeed, the role of econometrics is more important in social sciences today than ever. Given the advent of the information age, tremendous amounts of highly disaggregated information flows have had an increased influence over the social sciences both theoretically and empirically as well as over the decision making structures.

With this dramatic increase in disaggregated information, the absorption of the relevant information can be done most efficiently when the information has been quantified. This has resulted in an increase in the level of quantification in the social sciences , and thereby a change in the structure of social sciences. This quantification process faces two fundamental and related issues.

The first one is rapid changes in the world economy, which brings up the question of the ability of detecting structural changes. Relationships estimated by econometrics are time , data and location specific. If there is a continuous change in the structural aspects , then not only do we need to alter our model specifications , but also revise our data base. If we use econometrics without paying attention to these structural changes, then the coefficients of the models will be misleading us rather than guiding us.

This brings up to the second question on the methodology of econometrics. Today, there are three main approaches to econometric methodology, the Bayesians, Classical, and the Agnostics. The pioneering works in the Bayesian tradition include Zellner (1971) and Leamer (1978). The Classical who dominate their field estimate their parameters on the basis of economic theory and proceed with tests such as 't' , F and x2 statistics to support the validity of their theorizing. However, as Johnston (1991) mentions , "charges of data mining and other abuses have confused this tribe and they are presently in some disarray." On the other hand, Agnostics, led by Sims (1980) , argue that economic theory is no help in

...specifying the form of relationships. This approach relies on vector autoregressions (VAR's) methods. However, this method has many problems arising from collinearity and too many coefficients, and suffers from empiricism. This requires, then, a reevaluation of the tools, theories and estimations to be used by decision makers in the coming decades. International macro modeling is an example of such a search. Its importance is heightened by the need for better international policy coordination.

Specifically, though, several proposals for reconsidering international macroeconomic modeling can be made. These might include 1) the need for further theoretical and empirical studies on the interdependence of the world economy, transmission of fluctuations, and internationally coordinated policies, 2) the impact of structural changes on the capacity of international macroeconomic models in exploring and forecasting the economic phenomena; and a closer link between micro and macro economic considerations, 3) more emphasis on medium-term modeling, 4) the identification of institutional, national, regional, behavioral specification, and some economic behavioral changes, 5) more international coherence on the trade of goods, services, and financial and information flows, 6) the identification of the socioeconomic behavioral changes and the increasing importance of social protection and public expenditure, and 7) a policy oriented problem solving approach through sectoral disaggregation of structural models on the one hand, and central and satellite modelling, exploration modelling, and valid reduced forms for policy analysis on the other (Artus, Güvenen, 1986).

### 3.3. Information Systems

Information systems are necessary at almost all levels in society. All organizations need a sufficient flow of information in order for decision makers to make the decisions that face them each day. Indeed, the entire decisional structure depends on the free flow of information. The information system, in order to be useful to the decision makers, must be in a form so as to promote the dissemination of information in a useable form and in a timely fashion. In the context of the present information age, an information system will be considered successful based on its ability to condense the most significant information quickly to its users. Additionally, since decision-making is usually an interdisciplinary process, the information used must suit the nature of this process, and be therefore, of use in this regard.

When examining information system in the context of understanding social and economic phenomena, it is necessary to point out the uniqueness of social and economic phenomena in a particular point in history, geography, and society. This makes it difficult to undertake controlled experiments as in the natural sciences. Although recent literature has propounded the "birth of a new science", where order and pattern replace what was formerly considered random (Gleick, 1987), there still exist many gaps in our knowledge base. The issue is further complicated because of the close relationship between human behavior and the existing social and economic phenomena.

As was mentioned earlier, individuals, or decision markers, are affected by the existing social and economic phenomena when making decisions. Once decisions are made and acted upon, the social and economic phenomena that form the basis of future decisions are changed. As was also mentioned, alleviating these problems requires the analysis and quantification of any social and economic phenomena within the context of understanding the global structure. Understanding the global structure necessitates the use of the quantification methodology of the social sciences, which are intimately tied to behavior patterns.

Because the social sciences are primarily linked to behavior patterns, the strong bond between behavior patterns and information flows will be a main determinant in the evolution of the social sciences and the decision making structure. As mentioned above, the volume of the disaggregated information available necessitates quantification for simplification purposes; and the underlying tool for simplification purposes; and the underlying tool for this within the context of behavioral patterns in human society is econometrics. Thus, in the 1990s and the decades that follow, we will be witnessing a phase where an interdisciplinary and quantitative approach, both theoretically and empirically, will become a necessary precondition. With this, the endogenization of information flows, and thus information systems, will also become a necessary precondition. These trends are already having a strong impact on the decisional systems, and this has shown the need for a reconsideration of the existing theories and research being undertaken from this perspective.

Other extremely important factors in creating optimal decisional structures by extracting the maximum value possible from information flows are the freedom of opinion and the freedom of expression. These allow the creation of information and its transmission in the broadest possible terms, which positively impact the decisional structure in that decision makers have wider scope from which to view their alternatives. This is the basis on which policy makers rely in making economic, social and political decisions. Clearly then, a clear link and interdependency can bee seen between the scientific process of collecting, and disseminating information on the one hand, and decision making on the other. In this regard, the independence of science is most essential to the decision making process, and the means for transforming the information flows generated by scientists into digestible form, namely econometrics, is the bridge between scientists and decision makers.

The role of having flexible, information system cannot be underestimated. So, the first stage in this process is the analysis and understanding of the global structure within an interdisciplinary approach. Via this information, the elaboration of any specification at the structural level may then occur.

Information flows are the core ingredients in the functioning of open market economies in democratic societies. The markets would be unable to function properly without transparency and unrestricted access to the most current information. While information system are at the center of the world's economic structure, it is also widely appreciated that the transparency that result from free information flows, and appropriate statistical infrastructures to present this information in a useable format, is vital to achieving the highest possible level of democracy in any given society.

However, it must also be acknowledged that certain specification problems arise from the lack of data or from mismeasurement problems because the specification methodology is determined independently of the data collection process. In other words, social scientists must rely on database which basically answer to the needs of governments and business, thus causing an alteration in scope of the specification methodologies (Güvenen, 1991).

### 3.4. Statistical Inferastructures

The main purpose of a statistical infrastructure is to provide information services. Thus, it may be considered a subset of a given information infrastructure, or of the information system. The statistical infrastructure uses an information system to collect, store, retrieve, transform, process and communicate information through the latest available technology.

Today, as mandated by the United Nation's "Fundamental Principles of Official Statistics in the Region of the Economic Commission for Europe," the primary aim of official statistics, and therefore of the statistical infrastructures, is to "provide an indispensable element in the information system of a democratic society, serving the government, the economy, and the public with data about the economic, demographic, social and environmental situation... to be compiled and made available on an impartial basis by official statistical agencies to honor citizens entitlement to public information." This adoption of the 47th session of the United Nations Economic Commission for Europe goes on to include the resolution "to facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics." It adds that "the coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system. The use of statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels. Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries." These points clearly show the emphasis being placed today on the scientific approach to information systems and the standardization of statistical infrastructures to the benefit of the global system of collecting and utilizing the voluminous amount of disaggregated information flows.

Given the enormous amount of change that technological advancements and information technology are causing in the global system, the statistical infrastructure becomes increasingly important in collecting, sorting and producing the information in a meaningful way, The flexibility of the statistical infrastructure must be proportional to the speed at which information flows are generated.

At present, information structures are insufficient and have not been maximizing the potential of the information flows in terms of their beneficial possibilities for the decision making structure. This is largely due to the fact that the existing economic theories are limited in their current ability to optimize the use of the massive amount of information for the decision making structure. What is needed is a new, interdisciplinary approach that combines the scientific quantification and statistical-economic methodologies.

By combining these elements, statistical infrastructures can achieve improved levels of accurate and timely information. This is the direction, it would seem, that future research needs to be directed towards to reach a flexible statistical infrastructure that decision makers can utilize to make their decisions among a given set of alternatives.

To conclude this section, it should be emphasized once again that the overall goal is to benefit the global system by enhancing the choose of alternatives for decision makers. This can be realized by quantifying the information flows into viable means of analyzing and absorbing information at the institutional and individual levels so as to produce standardized and globally accessible data sets. In short, this is the equivalent to the optimization of the interaction between econometrics, information systems and statistical infrastructures.

## 4. The Interdependence Between Information Structures and Decision Making Structures

In rapidly changing economic conditions, an information structure should alert decision making structures so that they may undertake timely policy measures. Even in the most developed countries, such as the United State, it is well known that monthly and quarterly figures are revised several times before the actual numbers are finalized. This implies that the preliminary numbers may suggest a downturn in the economy when actually there is an upward trend. Or, forecasts of the economy may be in the wrong direction. Especially when the outside and inside lags are taken into account, policy action may worsen the situation rather than helping to solve it. That is, the information structure of the economy and the capacity of the decision making structure to take advantage of the information structure become crucial in determining the performance of the economy.

The relationship between the information structure and the decision making structure should develop in such a way that it will lead to continuous improvements in both structures. That is, information structures should provide the decision making structure with the necessary information, whereas the decision making structure should convey its demands on the information structure in a clear manner. Through this relationship, both structures should be able to focus on and sort out the information that is crucial for decision makers out of the numerous information points seemingly relevant for policy making.

In democratic societies, information systems facilitate the transmission of information which can always be verified for validity and regenerated according to scientific criteria. The existence of such an information system is the means by which a market economy is able to function properly, for in a market economy, all agents need a credible statistical infrastructure and reliable information. This is necessarily true because decentralized decision making is inherent in any market economy. In other words, the existence of reliable and timely information are the exact reasons why market uncertainties are reduced to a viable level. This is indeed the reality of the impact that information systems have on decision making structures.

It is noteworthy to point out that the dissemination of the information coming from the information flows are, in fact, the most reliable source of data when observing economic and social trends. Clearly then, the soundness of the information infrastructure has a direct impact on how the decision making process operates. The next logical deduction is that since statistics are the foundation of information systems, and since democratic forms of governments rely heavily on the free flow of information, statistics, and their method of arriving at their quantified state, namely econometrics, play a fundamental role in the democratic process.

## 5. Conclusion

Because the world is entering an era of globalization, where local and national economies are becoming increasingly interdependent, economic, technological, and social trends are quickly transcending regional and national boundaries. This makes it imperative for each country to harmonize its interests with others to have updated information on the latest developments and advances.

Globalization is occurring very rapidly primarily because a new element in the global system has been introduced, namely disaggregated information flows at unprecedented levels. There is also an entirely new system of explanatory variables because of this. This implies that there will be rapid change in both economic theory and quantification techniques which will thereby incur a transformation in these areas. With respect to the social sciences, the impact will be possible to explain the interaction between the global structure and the "E" as well as other social sciences.

The challenge for social scientists is to restructure the formulation of the specification methodology so that it is formed in relation to the data collection process according to the needs of these scientists as opposed to only the needs of businesses and governments. In other words, the data collection process and the formulation of the data bases must also be based on the specification methodologies of social scientists. In light of the abundance of the free flow of disaggregated information, the current economic theory and quantification methods can be reformulated so as to better understand the global structure with the available tools, and perhaps with the creation of newer and more efficient tools. It is inevitable that there will be changes in the theoretical and quantification methodologies of the social sciences since the world, the global structure and therefore the basis of our quantification methods are changing.

Ideally, such a reformulation will bring about more coherence in the system. Through the flow of information and the optimal formulation of statistical infrastructures, economic theory will be improved because the global structure will be closer to being explained in analytical terms.

The era of information will mark a higher level in the progression of mankind because an inseparable link has been formed between statistical infrastructures and democracy. Freedom of opinion and expression lie at the center of establishing scientific and transparent information flows. These freedom allow the creation of information and its subsequent transmission to occur in an optimal manner. This benefits decision making and the decisional structure at the highest level because it broadens the scope for the realm of the data in methodology specifications.

Furthermore, in market economies, where decentralized decision making is inherent to the system, the existence of reliable and timely information reduce market uncertainties to viable levels. By reducing the imperfections in the market structure through optimizing the use of information flows, the market structure functions more efficiently and results in an improved allocation of resources. Increases in the volume of information flows also mean that problem solving will occur at increasingly disaggregated levels, and by using the appropriate quantitative techniques, there will be a revision of the economic theory and of the quantification techniques that will best suit needs of the agents in the free market economy.

To summarize, the transmission of information, through its accumulation and quantification, is the definition of the statistical production process. Econometrics is the tool by which this is achieved, and is irreplaceable in quantifying the social sciences to the maximum extent possible. The availability of reliable information at all levels within a given society facilitates the decentralized decisional structure. This allows societies to successfully confront the challenges presented by the rapidly changing conditions sweeping through the global structure. Because statistical information is a tool for knowledge, analysis, decision making and evaluation, decision makers are empowered with the necessary information to advance their societies' standard of living and the overall progression of mankind. Thus, through an interdisciplinary approach, the global structure and explanatory variables used to understand it will reach higher planes.

At present, changes in the focus of modeling for achieving the above mentioned goals are vitally needed. It is hoped that subsequent discussions and further research will be stimulated according to the needs of new economic theory and quantification formulations, and of a growing and improving decisional structure.

## References

ırtus, P. and Güvenen, O. (1986). *International Macroeconomic Modelling for Policy Decisions.* Martinus Nijhoff Publishers, Dordrecht.

Barbone, L. and Poret, P. (1989). *'Structural Conditions and Macroeconomic Responses to Shocks:* A Sensitivity Analysis for Four European Countries,' OECD Economic Studies, No. 12, Spring 1989.

Dock, T. (ed.) (1972). MIS, a Managerial Perspective, Science Research Associates, Inc. Chicago.

Gleick, James. (1987). *Chaos;* Making a New Science. Penguin Books, New York.

Güvenen, O. (1991). 'Some Comments on the Specification Methodology of Social and Economic Phenomena.' *Contributed Papers Volume of the 48th Session of the International Statistical Institute,* Book 1, Cairo, Egypt, September.

Güvenen, O. (1992). 'Some Comments and Proposals on the Statistical Infrastructure and Regional Information Systems for the Black Sea Economic Cooperation Countries.' The State Institute of Statistics, Ankara, Turkey.

Johnston, J. (1991). 'Econometrics Retrospect and Prospect.' *The Economic Journal,* 101, January, pp. 51-56.

Leamer, E. (1978). *Specification Searches.* New York: Wiley.

Murray, T. (1985). Computer Based Information Systems, Richard D. Irwin, Inc. Homewood.

Neumann, A. (1982). Principles of Information Systems for Management W.C. Brown Company Publishers, Dubuque.

Peitgen, H. and Dietmar, S. (eds.), (1988). *The Science of Fractal Images,* Springer- Verlag, New York.

Sims, C. (1980). 'Macroeconometrics and Reality.' *Econometrica,* Vol. 48, pp. 1-49.

United Nations, (1992). 'Fundamental Principles of Official Statistics in the Region of the Economic Commission for Europe.' 47th Session of the U.N. Commission for Europe, Geneva, April.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics,* New York: Wiley

## Özet

Ekonometri, bilgi sistemleri ve istatistik alt yapılar bir birine bağımlı olup beraberce bir toplumda karar sistemlerini optimal olarak etkileyen araçlardır. Mevcut ekonomi ve sosyal bilimler kavramlarının yoğun bilgi akışı ve niceliklendirme tekniklerindeki ilerlemelerden önemli ölçüde etkilendiği görülmektedir. İstatistik alt yapıların hızlı ve güvenilir bilgi

sunumları ve etkili veri işleme araçları hızla artan bilgi akışı sağlamakla sosyal bilimlerin deneysel ve kuramsal yönlerinde güçlü bağların oluşumuna yol açmaktadır. Bu ise, sosyal bilimlerde varolan teorilerin ve çözümleme yöntemlerinin reforma uğratılması anlamına gelmektedir.