



# Minimum variance quadratic unbiased estimators as a tool to identify compound normal distributions

Jean-Daniel Rolle\*

*HEC-Management Studies, University of Geneva, CH-1211 Geneva, Switzerland, HEG-Haute Ecole de Gestion, CH-1700, Fribourg, Switzerland*

Received 31 August 1997; received in revised form 7 September 1998

---

## Abstract

We derive the minimum variance quadratic unbiased estimator (MIVQUE) of the variance of the components of a random vector having a compound normal distribution (CND). We show that the MIVQUE converges in probability to a random variable whose distribution is essentially the mixing distribution characterising the CND. This fact is very important, because the MIVQUE allows us to make out the signature of a particular CND, and notably allows us to check if an hypothesis of normality for multivariate observations  $y_1, \dots, y_M$  is plausible. © 1999 Elsevier Science B.V. All rights reserved.

**Keywords:** Normal linear regression; Compound normal distributions; Quadratic estimation; Error components model

---

## 1. Introduction

Suppose that our data are individual  $N$ -variate observations  $y_m$ ,  $m = 1, \dots, M$ , that are (multivariate) measures of a phenomenon. The measurements were performed under changing conditions. A reasonable model for these observations (or measures) is the linear system  $y_m = \mu + U_m$ , where  $\mu$  is an  $N \times 1$  location vector, and, conditional on a random scale parameter  $\tau$ ,  $U_m$  is Gaussian:  $U_m \sim N(0, \tau \Sigma)$ . This means that the  $y_m$  have a compound normal distribution. We examine here how we can learn something about the distribution of  $\tau$ , and enlarge the problem by assuming that we have known covariate information on the location parameter, so that  $\mu = \mu(X) = X\beta$ . Zellner [13] describes cases where the scale mixtures of normal prove useful in practice. He notes that one may look at the multivariate realisations  $y_m$  of  $Y$ ,  $m = 1, \dots, M$ , as being generated by a measuring instrument. The variability of the instrument, represented by a scale factor for the covariance matrix, has an unknown

---

\* E-mail: rolle@uni2a.unige.ch.

value within a run and is known to vary over the  $M$  runs. Rolle [10] used this model to study aggregated multivariate measures performed under changing circumstances, when data are produced in a network setting.

It is well known that compound normal distributions with high kurtosis will produce outliers. We propose here a procedure to detect departures from the multivariate normal distribution which is not directly related to the question of detection of multivariate outliers. The latter has known important developments during the last decade. Rousseeuw and Van Zomeren [12] note that usual techniques mask outlier detection, and propose to replace in the Mahalanobis distance the arithmetic mean of the data set and the sample covariance matrix by estimators with high breakdown point. More precisely, they use the minimum volume ellipsoid estimator introduced by Rousseeuw [11]. Their technique immediately applies to identification of leverage points in regression. In this context, the authors propose a plot of standardized least median of squares residuals versus robust distances; this plot proves very useful to classify observations. Cook and Hawkins [3] showed that a method based on minimum volume ellipsoid estimators may indicate too many outliers, and that the approximate algorithm used for their computation may be instable. Rocke and Woodruff [7] give insights into why the problem of outlier detection is so difficult, specially in high dimensionalities, and a method incorporating an algorithm proposed by Atkinson [1]. Atkinson's [2] forward search is based on robust estimators: the least median of squares estimators for regression and the minimum volume ellipsoid for multivariate outliers. Our aim here is not to unmask outliers, but to find what kind of random mechanism produced the outliers in a well-defined parametric setting.

## 2. The necessary tools

To provide a procedure able to detect departures from the multivariate normal distribution, we analyze the behavior of quadratic estimators. First of all, let us recall a few definitions needed in the sequel. Let  $\mu$  be a  $N \times 1$  vector and  $\Sigma$  a  $N \times N$  symmetric matrix. A  $N \times 1$  random vector  $Z$  is said to have a CND with parameters  $\mu$  and  $\Sigma$ , and is denoted  $Z \sim \text{CN}(\mu, \Sigma, \phi_H)$ , if its density function has the form

$$f(z) = \int_0^\infty N(z; \mu, \tau \Sigma) dH(\tau), \quad (1)$$

where  $N(z; \mu, \tau \Sigma) = (2\pi\tau)^{-N/2} (\det \Sigma)^{-1/2} \exp(-(z - \mu)' \Sigma^{-1} (z - \mu)/2\tau)$  is the density function of the Gaussian distribution, and  $H$  is a distribution function for the nonnegative random scale parameter  $\tau$ . From this definition, if  $Z \sim \text{CN}(\mu, \Sigma, \phi_H)$ , then equivalently one has that, conditional on  $\tau$ ,  $Z \sim N(\mu, \tau \Sigma)$ . The distribution of  $\tau$  is a mixing distribution, and the class of CND is obtained by varying this distribution [6]. If  $H$  is degenerated on  $\tau = 1$ , that is if  $\tau$  takes the value 1 with probability 1, then  $f(z) = N(z; \mu, \Sigma)$ , that is  $Z$  is Gaussian. If  $Z$  has a CND, then its characteristic function is of the form  $c_Z(t) = e^{it'\mu} \phi_H(t'\Sigma t)$  for some function  $\phi_H$  (depending on  $H$ , as the notation suggests). Provided that relevant moments exist, one has  $E(Z) = \mu$  and  $\text{Cov}(Z) = -2\phi_H'(0)\Sigma = E_H(\tau)\Sigma = V$ , say. For example, if  $Z$  is Gaussian with  $f(z) = N(z; \mu, \Sigma)$ , then  $\phi_H(s) = \exp(-s/2)$ , and  $V = \Sigma$ . Moreover, if  $Z = (Z_1, \dots, Z_N)'$  has finite fourth moments, then the marginal distributions  $Z_j$  all have zero skewness and the same kurtosis  $3(\phi_H''(0) - \phi_H'^2(0))/\phi_H'^2(0) = 3\kappa_H$ , say [6, p. 41]. Using the characteristic function  $c_Z(t)$  to generate moments, one can show that  $\kappa_H = \text{var}_H(\tau/E_H(\tau))$ .

### 3. The model and the main results

We consider here the linear regression model

$$y = X\beta + U, \quad (2)$$

where  $U$  is a  $N \times 1$  vector of disturbances,  $X$  is a  $N \times K$  known matrix, not necessarily of full rank, and  $\beta$  is a  $K \times 1$  vector of regression coefficients. We assume that  $U$  in (2) is such that  $U \sim \text{CN}(0, \gamma^2 I, \phi_H)$ . Let the covariance matrix of the error vector be  $V = \text{Cov}(U) = \sigma^2 I$ . A special case of this model was considered by Zellner [13], where  $U$  was assumed to follow a multivariate Student- $t$  distribution. More general forms can be assumed for  $V$  without changing the qualitative results. Notably, the conclusions are the same if  $V$  has the structure of a covariance matrix from the error components model, a model well known by econometricians working on  $p$ -way classified data. Here we will

1. Find the minimum variance unbiased estimator  $\hat{\sigma}^2$  of  $\sigma^2$  among the quadratics  $y' Ay$  with  $A \geq 0$ , i.e. we will seek the nonnegative MIVQUE. We show that this estimator is

$$\hat{\sigma}^2 = \frac{1}{N - \text{r}X} y'(I - XX^+)y, \quad (3)$$

where  $X^+$  is the Moore–Penrose inverse of  $X$ , and  $\text{r}X$  denotes the rank of  $X$ . Note that this estimator does not depend on any particular (mixing) distribution of  $\tau$  and hence on any particular CND (in that sense, it is a uniform MIVQUE).

2. Show that  $\hat{\sigma}^2$  converges in probability (and hence in distribution) to  $\sigma^2 \tau / E_H(\tau)$ , where  $E_H$  denotes mathematical expectation.

We can draw two conclusions from 2. First,  $\hat{\sigma}^2$  converges in probability to a random variable, and hence is inconsistent, unless  $U$  is Gaussian — in that case,  $\tau$  is degenerated on 1. Thus, the MIVQUE cannot be used per se. This shows that the minimum variance criterion can be a rather poor criterion for choosing estimators. Second, and more promisingly, we have that for reasonably large  $N$ , the distribution function of  $\hat{\sigma}^2$  is essentially  $H$ . Hence, if we have a set of i.i.d.  $N$ -variate observations  $y_i$ ,  $i = 1, \dots, M$ , of mean  $X\beta$ , that we suspect to come from a heavy-tailed distribution, we can model  $y_i \sim \text{CN}(X\beta, \gamma^2 I, \phi_H)$ . An analysis of the (empirical) distribution of the independent  $\hat{\sigma}_i^2 = y_i' M_X y_i / (N - \text{r}X)$ , where  $M_X = I - XX^+$ , will provide much information about how the  $y_i$  were generated. In particular, concentration of the  $\hat{\sigma}_i^2$  around a single point indicates normality. As  $\hat{\sigma}^2$  has approximatively (for large  $N$ ) the distribution of  $\sigma^2 \tau / E_H(\tau)$ , we propose the following process: (i) We assume that we have i.i.d. realizations  $y_1, \dots, y_M$  of  $y$ . (ii) We compute  $\hat{\sigma}_1^2, \dots, \hat{\sigma}_M^2$  and their mean  $a = (\sum \hat{\sigma}_i^2) / M$ . (iii) We compute  $t_1 = \hat{\sigma}_1^2 / a, \dots, t_M = \hat{\sigma}_M^2 / a$ . The  $t_1, \dots, t_M$  approximate realizations of  $\tau / E_H(\tau)$ . (iv) We analyze the distribution of the  $t_i$  (thanks to usual tools such as boxplots, histograms, etc.) to detect potential non-Gaussian features. Next, as  $\kappa_H = \text{var}_H(\tau / E_H(\tau))$ , the sample variance  $s_t^2$  of the  $t_i$  will provide an estimate of  $\kappa_H$ . Note that  $\kappa_H = 0$  in the Gaussian case, and  $\kappa_H > 0$  otherwise. The higher the  $\kappa_H$ , the larger the departure from normality. Various simulations showed that this technique works very well. Although more formal procedures can be developed to exploit this finding, we shall here content ourselves with a brief description. Let  $\mathbf{1}$  be a  $N \times 1$  vector of ones. We simulated  $M = 2000$  i.i.d.  $N$ -variate (with  $N = 100$ ) realizations  $y_i = \mathbf{1} + U_i$  in three cases: (a)  $U_i \sim \text{N}(0, 4I)$ , (b)  $U_i \sim$  contaminated normal  $0.8 \text{N}(0, I) + 0.2 \text{N}(0, 16I)$ , and (c)  $U_i \sim$  another heavy-tailed CND where  $H$  was chosen to be the chi-square distribution with 5

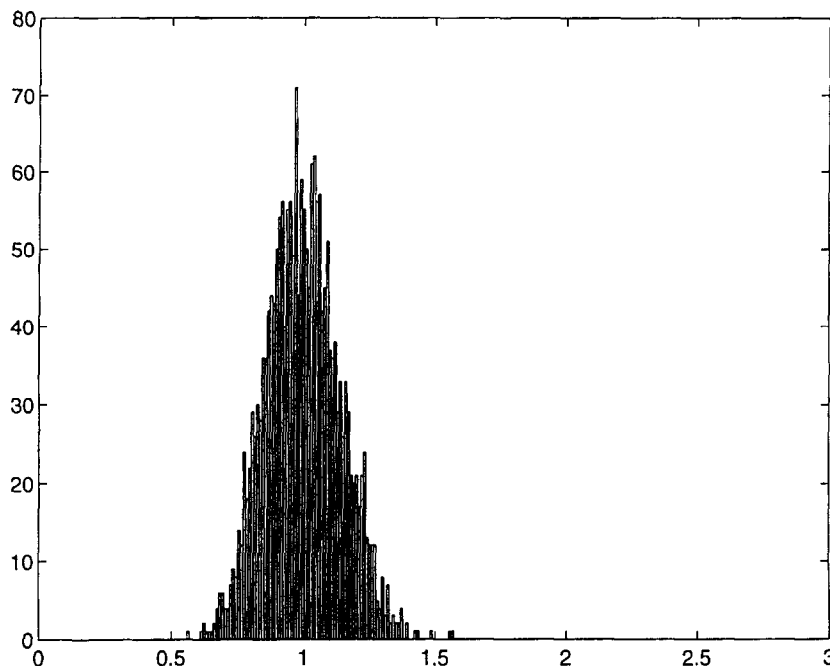


Fig. 1. We simulated  $M=2000$   $N$ -variate ( $N=100$ ) Gaussian realizations  $y_i=1+U_i$ , where  $1$  is a  $N \times 1$  vector of ones, and  $U_i \sim N(0, 4I)$ ,  $i=1, \dots, M$ . In such a case, we have simply  $X=1$ . Then we computed the quadratics  $\hat{\sigma}_i^2 = y_i' M_X y_i / (N-1)$ . The  $M$  values  $t_i$  were obtained by dividing the  $\hat{\sigma}_i^2$  by their mean. Fig. 1 displays the resulting histogram of the  $t_i$  and suggest concentration around 1.

d.f., that is  $H = \chi_5^2$ . The histograms of the  $t_i$  for the three cases appear in Figs. 1–3. As the theoretical findings predict, they indicate concentration around one point for (a), 80% concentration around  $\frac{1}{4}$ , and 20% around 4 for (b), and the shape of the  $\chi_5^2$  divided by its mean, 5, for (c). We took  $N=100$  for this simulation, but even for  $N$  with magnitude 20 or 30, non-Gaussian features can be detected. Of course, the smaller the  $N$ , the more diffuse this information. The corresponding approximations  $s_t^2$  of  $\kappa_H$  may be read in Table 1. More formal analysis such as approximate normality tests, or density estimation, may be applied to this data generating mechanism, and are the object of current research.

#### 4. Proof of the statements

We first prove 1. above, i.e. we show that (3) is the MIVQUE. Using Li Gang's result from Fang and Anderson [4], one can show that if  $Z \sim \text{CN}(\mu, \Sigma, \phi_H)$ , and  $Z'AZ$  is such that  $A\mu=0$ , then

$$\text{Var}(Z'AZ) = \kappa_H(\text{tr} AV)^2 + 2(\kappa_H + 1)\text{tr} AVAV. \quad (4)$$

Next, since  $U \sim \text{CN}(0, \gamma^2 I, \phi_H)$ , then  $y = X\beta + U \sim \text{CN}(X\beta, \gamma^2 I, \phi_H)$ . Let  $y'Ay$  be a potential estimator for  $\sigma^2$ . Nonnegativity and unbiasedness ( $E(y'Ay) = \sigma^2$ ) imply  $AX=0$ , as one can show easily. Hence  $A\mu = AX\beta = 0$  and, since  $E(Z'AZ) = \text{tr} AV$ , we have that  $E(y'Ay) = \sigma^2 \text{tr} A$ . From (4),

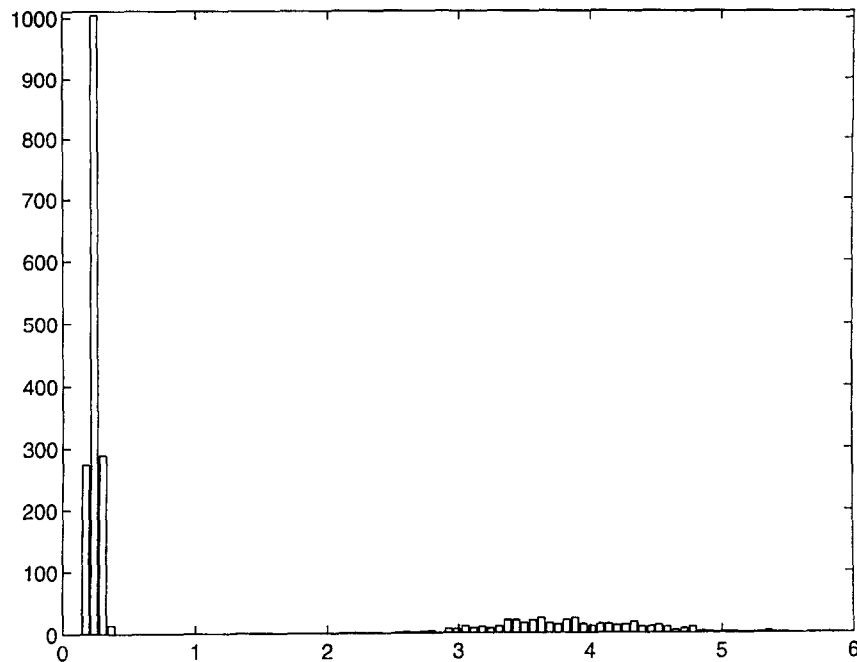


Fig. 2. Simulation of  $M = 2000$   $N$ -variate ( $N = 100$ ) contaminated realizations  $y_i = 1 + U_i$ , where  $U_i$  is  $0.8N(0, I) + 0.2N(0, 16I)$ . The  $M$  values  $t_i$  were obtained by dividing the  $\hat{\sigma}_i^2$  by their mean. The histogram of the  $t_i$  suggests contamination: nearly 80% of the  $t_i$  fell around  $\frac{1}{4}$ , and nearly 20% of them fell around 4.

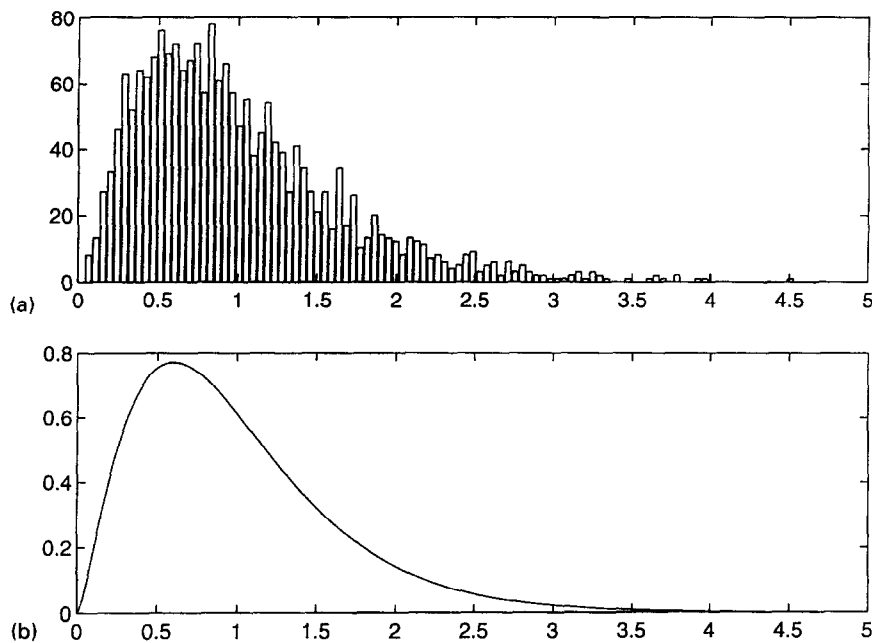


Fig. 3. Simulation of  $M = 2000$   $N$ -variate ( $N = 100$ ) realizations  $y_i = 1 + U_i$ , where  $U_i \sim N(0, \tau I)$ , conditional on  $\tau \sim \chi_5^2$ . Fig. 3 (a) displays the histogram of the  $t_i$ . For comparison, the graph of the p.d.f. of  $\tau/5$  – the p.d.f. approached by the histogram – is given in Fig. 3 (b). A look at the histogram reveals positive skewness, and hence excludes the Gaussian hypothesis for the  $y_i$ .

Table 1  
Kurtosis parameter

	Simulated	True
Normal	$s_t^2 = 0.0197$	$\kappa_H = 0$
Contaminated normal	$s_t^2 = 2.2502$	$\kappa_H = 2.25$
Mixing distribution $\chi_5^2$	$s_t^2 = 0.4087$	$\kappa_H = 0.4$

$\text{Var}(y'Ay) = \kappa_H \sigma^4 (\text{tr } A)^2 + 2\sigma^4 (\kappa_H + 1) \text{tr } A^2$ . Define the set of matrices  $\mathcal{A} = \{A \in \mathbb{R}^{N \times N}; A \geq 0, AX = 0, \text{tr } A = 1\}$ . The quadratics  $y'Ay$  with  $A \in \mathcal{A}$  are unbiased for  $\sigma^2$ . We are seeking an optimal estimator  $y'A^*y$  of  $\sigma^2$ , with  $A^* \in \mathcal{A}$ , such that  $\text{Var}(y'A^*y) \leq \text{Var}(y'Ay)$  for all  $A \in \mathcal{A}$ . To derive the optimal  $A^*$ , it is very useful to characterize the matrices  $A \in \mathcal{A}$ . To do so, let us define the open set of matrices  $\mathcal{P} = \{P \in \mathbb{R}^{N \times N}; PM_X \neq 0\}$ , where  $M_X = I - XX^+$ . Let us prove the following lemma:

**Lemma 1.** *The function  $h_1 : \mathcal{P} \rightarrow \mathcal{A}$  given by  $h_1(P) = \|PM_X\|^{-2} M_X P' PM_X$  is surjective (i.e.  $h_1(\mathcal{P}) = \mathcal{A}$ ).*

**Proof.** Let  $A \geq 0$ , with  $AX = 0$ .  $A \geq 0$  implies the existence of a  $N \times N$  matrix  $B$  such that  $A = B'B$ . Next,  $0 = AX = B'BX$  implies  $X'B'BX = 0$ . Therefore  $BX = 0$ . The general solution of  $BX = 0$  is  $B = PM_X$  (see [5, ex. 4 p. 38]), where  $P$  is an arbitrary matrix of appropriate order. Hence  $A = B'B = M_X P' PM_X$ . The unbiasedness condition imposes that  $\text{tr } A = 1$ , that is  $\text{tr}(M_X P' P) = \|PM_X\|^2 = 1$ . Hence  $A = \|PM_X\|^{-2} M_X P' PM_X$ .  $\square$

To continue the proof of 1., define the function  $h_2 : \mathcal{A} \rightarrow \mathbb{R}$  by  $h_2(A) = \text{Var}(y'Ay)$  and  $h : \mathcal{P} \rightarrow \mathbb{R}$  by  $h = h_2 \circ h_1$ . We will find a  $P^* \in \mathcal{P}$  such that  $h(P^*) \leq h(P)$ . The function  $h_1$  being surjective, this means that  $h_2$  takes a minimum at  $A^* = h_1(P^*)$ , i.e.  $h_2(A^*) \leq h_2(A)$  for all  $A \in \mathcal{A}$ . Using the representation  $A = \|PM_X\|^{-2} M_X P' PM_X$  and (4), one has  $h(P) = \text{Var}(y'Ay) = \kappa_H \sigma^4 + 2(\kappa_H + 1) \sigma^4 [\text{tr}(M_X P' P)^2] [\text{tr}(M_X P' P)]^{-2}$ . Therefore we can equivalently minimize on the open set  $\mathcal{P}$  the differentiable function  $\mu(P) = \text{tr}(M_X P' P)^2 / (\text{tr } M_X P' P)^2$ . Using Theorem 2, p. 265 in Rolle [9], a global minimizer  $P^*$  of  $\mu(P)$  is given by  $(N - rX)^{-1/2} J'$ , where  $J$  is such that  $JJ' = M_X$ . Hence the optimal matrix  $A^* \in \mathcal{A}$  at which  $h_2$  takes a minimum is given by  $A^* = h_1(P^*) = \|J'M_X\|^{-2} M_X JJ' M_X = M_X / (N - rX)$ , noting that idempotence of  $M_X$  implies  $rM_X = \text{tr } M_X = N - rX$ . The optimal estimator is then  $\hat{\sigma}^2 = y'M_X y / (N - rX)$ .

We have still to prove 2., i.e. we must show that  $\hat{\sigma}^2$  converges in probability to  $\gamma^2 \tau$ . Noting that  $M_X X = 0$ , we have  $\hat{\sigma}^2 = y'M_X y / (N - rX) = U'M_X U / (N - rX)$ . Next, by assumption  $U \sim \text{CN}(0, \gamma^2 I, \phi_H)$ , for some  $H$ . It follows that  $U = \tau^{1/2} Z$ , where  $Z$  is  $N(0, \gamma^2 I)$ , and  $\tau, Z$  are independent. Hence  $\hat{\sigma}^2 = 1 / (N - rX) U'M_X U = 1 / (N - rX) \tau Z' M_X Z = \tau \gamma^2 / (N - rX) Z'(M_X / \gamma^2) Z = \tau \gamma^2 / (N - rX) \chi_{N-rX}^2$ . But  $\chi_{N-rX}^2 / (N - rX)$  converges in probability to 1, and thus  $\hat{\sigma}^2$  converges in probability to  $\gamma^2 \tau$ . Moreover,  $\sigma^2 = -2\phi'(0)\gamma^2$ , and  $-2\phi'(0) = E_H(\tau)$ , by a moment generating property of the characteristic function. That is,  $\hat{\sigma}^2$  converges in probability to  $\sigma^2 \tau / E_H(\tau)$ . Another way to see this is to note that

the components of  $U$  are exchangeable (but not independent). The law of large numbers for such sequences allows the limit to be a random variable. This is what happens here.

## References

- [1] A.C. Atkinson, Stalagmite plots and robust estimation for the detection of multivariate outliers, in: S. Morgenthaler, E. Ronchetti, W. Stahel, *Data Analysis and Robustness* (Eds.), Birkhauser, Basel, 1993.
- [2] A.C. Atkinson, Fast very robust methods for detection of multiple outliers, *J. Amer. Statist. Assoc.* 428 (1994) 1329–1339.
- [3] R.D. Cook, D.M. Hawkins, Comments on Unmasking multivariate outliers and leverage points, by P.J. Rousseeuw and B.C. Van Zomeren, *J. Amer. Statist. Assoc.* 85 (1990) 640–644.
- [4] K.T. Fang, T.W. Anderson, *Statistical Inference in Elliptically Contoured and Related Distributions*, Allerton Press, New York, 1990.
- [5] J.R. Magnus, H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, New York, 1988.
- [6] R.J. Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley, New York, 1982.
- [7] D.M. Roche, D.L. Woodruff, Identification of outliers in multivariate data, *J. Amer. Statist. Assoc.* 428 (1996) 1329–1339.
- [8] J.D. Rolle, Best nonnegative invariant partially orthogonal quadratic estimation in normal regression, *J. Amer. Statist. Assoc.* 428 (1994) 1378–1385.
- [9] J.D. Rolle, Optimization of functions of matrices with application in statistics and econometrics, *Linear Algebra Appl.* 234 (1996) 261–275.
- [10] J.D. Rolle, Aggregated multivariate measures performed under changing circumstances, *The Indian J. Statist.: Sankhya, Ser. A*, 60 (2) (1998) 232–248.
- [11] P.J. Rousseeuw, Multivariate estimation with high breakdown point, in: W. Grossman et al. (Eds.), *Mathematical Statistics and Applications*, vol. B, Reidel, Dordrecht, 1985.
- [12] P.J. Rousseeuw, B.C. Van Zomeren, Unmasking multivariate outliers and leverage points, *J. Amer. Statist. Assoc.* 441 (1990) 633–639.
- [13] A. Zellner, Bayesian and non-Bayesian analysis of the regression model with multivariate student  $t$  error terms, *J. Amer. Statist. Assoc.* 71 (1976) 400–405.