國 立 交 通 大 學

統計學研究所

碩 士 論 文

多維度參考區間方法

An Approach for Multivariate Reference Region

研究生：楊子賢

指導教授：陳鄰安 博士

中華民國一百零一年六月

多維度參考區間方法

An Approach for Multivariate Reference Region

研 究 生：楊子賢　　Student: Tzu-Hsien Yang

指導教授：陳鄰安 博士　　Advisors: Dr. Lin-An Chen

國立交通大學

統計學研究所

碩士論文

A Thesis

Submitted to Institute of Statistics College of Science

National Chiao Tung University

In partial Fullfillment of the Requirement

For the Degree of Master

In

Statistics

June 2012

Hsinchu, Taiwan, Republic of China

中華民國一百零一年六月

多維度參考區間方法

研究生：楊子賢　　指導教授：陳鄰安 博士

國立交通大學統計學研究所

摘要

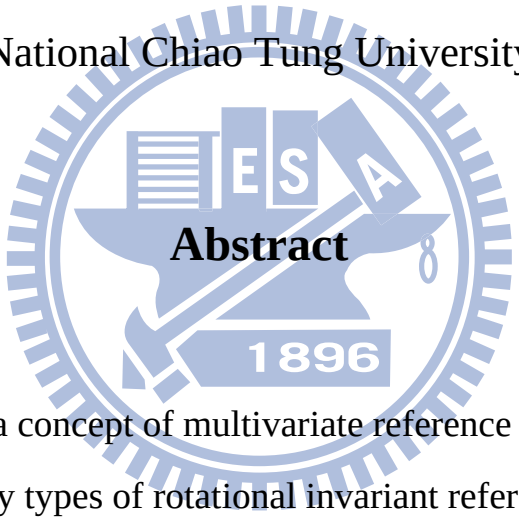
我們介紹了一個多維度參考區間的觀念。它准許我們去建造多種旋轉不變性參考區間的型態。呈現對未知多維參考區間估計之效率的模擬和偵測常態或非常態之檢定力的問題。模擬結果顯示了這篇文章中介紹的方法具有良好的效果。

An Approach for Multivariate Reference Region

Student: Tzu-Hsien Yang Advisors: Dr. Lin-An Chen

Institute of Statistics

National Chiao Tung University



Abstract

We introduce a concept of multivariate reference region. This allows us to construct many types of rotational invariant reference regions. Simulation studies of efficiency in estimation of unknown multivariate reference region and power in detection of normal or non-normal subject are performed. The simulation results show that the techniques introduced in this paper are desirable.

誌 謝

又到了鳳凰花開的六月，轉眼間研究所兩年的日子過去了，彷彿昨天才剛踏進研究所展開新的生活、新的學習，而現在即將劃下求學生涯的句點。

在研究所這兩年，首要感謝的是指導教授陳鄰安老師細心的指導和教誨，在做研究的過程中，老師會引導我們如何發現問題，然後進一步去思考，每當我遇到問題時，老師都會耐心的為我解惑。在做研究外，老師也常常關心我們的生活。除此之外，也從老師身上學到很多人生哲理，使我受益良多。

再者感謝所有的同學們，帶給我多采多姿的兩年生活，研究室裡總是充滿歡笑、課業上有問題時總能夠和我討論及給予我協助。最後謝謝家人支持與鼓勵，讓我能夠無憂無慮地念書。謝謝你們。

楊子賢 謹誌于
國立交通大學統計學研究所
中華民國一百零一年六月

Contents

List of Tables

1. Introduction

The determination of intervals for reference limits is fundamentally important in clinical chemistry. The reference interval in laboratory chemistry refers to population-based reference values obtained from a well-defined group of reference individuals. It is an interval with two confidence limits which covers the measurement values in the population in some probabilistic sense. The reference interval tells the physician if the patient's measured value is expected in a healthy or ill person or if further testing is warranted. Review of reference intervals can be found in Horn and Pesce (2003) and its statistical theory can be found in Huang, Chen and Welsh (2010).

Most medical decisions require consideration of several co-existing pieces of information, and because such pieces such as blood constituents are often correlated, the multivariate reference region is more useful than conventional univariate reference intervals for interpreting clinical laboratory results. It is an uncomfortable fact that there is a high probability that at least one result will lie outside its reference interval when many clinical tests are run on a blood sample from a healthy person. This indicates that a multidimensional point of correlated observations is likely to lie within the individual's multivariate reference region, even when one or more of the observations lie outside their separate reference intervals for the individual (see Schoen and Brooks (1970) and Harris, Yasaka et al. (1982)).

Although multivariate reference regions in the practice of clinical chemistry and laboratory medicine is very important, it has received limited attention in literature and applications while the ellipsoid method is the most popularly used multivariate one. The lack of a natural ordering for multivariate data makes the existing proposals of multivariate reference regions more or less ad hoc, hence, most com-

mon ones do not have parametrized versions and their applications are extremely limited (Chen and Welsh (2002)). One exceptional example sharing a feature of parametric version is the rectangular type multivariate reference regions using the multivariate distribution function (Wellek (2011)). This attempt is valuable in contribution toward multidimentional generalization of reference region. Unfortynately this approach lacks some desirable geometric properties, e.g., it is not rotationally equivalent.

Chen and Welsh (2002) and Shiau and Chen (2003) considered a normalization of the original measurement vector to construct multivariate quantiles and use them for building methods of statistical inferences for distributional parameters. We propose multivariate reference region of the original measurement sample space as the back transformation of a γ confidence set constructed from the normalized sample space. This allows us to construct specific rotational invariant multivariate reference regions of cube, ellipsoid, parallelogram, trapezoid or others. This generalization offers not only the choice of design for any specific distribution but also methods of efficiency.

2. Multivariate Reference Region

Let X be a bivariate random vector with joint probability density function (pdf) $f(x_1, x_2)$. Of interest is how to develop γ reference region C_x for a distribution of X. Elements of X are generally correlated that makes it difficult in constructing a rotationally invariant γ reference region. The interest in this paper is to develop rectangular or parallel multivariate reference region as an alternative choice of multivariate reference region.

Let the population mean of X be $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and the population covariance matrix of X be $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$. We consider the transformed random vector $Y = \Sigma^{-1/2'} \begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix}$ and define a γ confidence set C_y in y-space with

$$C_y = \{ \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} : F_{Y_1}^{-1}(\alpha) \leq y_1 \leq F_{Y_1}^{-1}(1-\alpha), F_{Y_2}^{-1}(\alpha) \leq y_2 \leq F_{Y_2}^{-1}(1-\alpha) \}\} \quad (2.1)$$

where α satisfies $\gamma = P\{Y \in C_y\}$ and $F_{Y_1}^{-1}$ and $F_{Y_2}^{-1}$ are, respectively, quantile functions of Y_1 and Y_2. A multivariate reference region is thus obtained by transforming this region back to the x-space as

$$C_x = \{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \Sigma^{1/2} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in C_y \} \quad (2.2)$$

A fundamental problem in multivariate statistics is the development of affine equivariant inference procedure. The following theorem states that with the reference region being the back transformation of the γ confidence set to the scale of X, our proposal allows the user to design various affine equivariant multivariate reference regions.

Theorem 2.1. We re-denote the covariance matrix, mean vector and γ reference region obtained from vector X by $\Sigma(X), \mu(X)$ and $C_x(X)$, respectively. Suppose that $\Sigma^{1/2}(AX+b) = A\Sigma^{1/2}(X)$ and $\mu(AX+b) = A\mu(X)+b$. Then the γ reference region is affine equivariant with

$$C_x(AX+b) = AC_x(X) + b.$$

Proof. We re-denote the parameters Y, Σ, μ, C_y and C_x respectively by $Y(X), \Sigma(X)$, $\mu(X)$, $C_y(X)$ and $C_x(X)$. Notice that

$$\Sigma^{-1/2'}(AX+b)(AX+b-\mu(AX+b)) = \Sigma^{-1/2'}(X)(X-\mu(X))$$

and $Y(AX+b) = Y(X)$ leading to $C_y(AX+b) = C_y(X)$ so

$$\begin{aligned} C_x(AX+b) &= \{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \Sigma^{1/2}(AX+b)\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \mu(AX+b), \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in C_y(AX+b)\} \\ &= \{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = A\Sigma^{1/2}(X)\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + A\mu(X) + b, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in C_y(X)\} \\ &= AC_x(X) + b. \quad \square \end{aligned}$$

There are many choices in setting of γ confidence set C_y. We here give several choices, some constructed by population quantile functions $F_{Y_1}^{-1}$ and $F_{Y_2}^{-1}$, of interesting γ confidence set C_y:

$$\text{Upper region } C_U : \{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} : y_1 \geq F_{Y_1}^{-1}(\alpha), y_2 \geq F_{Y_2}^{-1}(\alpha)\},$$

$$\text{Lower region } C_L : \{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} : y_1 \leq F_{Y_1}^{-1}(\alpha), y_2 \leq F_{Y_2}^{-1}(\alpha)\},$$

$$\text{Parallelogram } C_P : \{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} : F_{Y_1}^{-1}(\alpha) \leq y_1 \leq F_{Y_1}^{-1}(1-\alpha), F_{Y_2}^{-1}(\alpha) \leq y_2 \leq F_{Y_2}^{-1}(1-\alpha)\},$$

$$\text{Ellipsoid } C_{ellip} : \{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} : y_1^2 + y_2^2 \leq \chi_\alpha^2\},$$

where χ_α^2 is the α quantile point of the chi-square distribution $\chi^2(p)$ and the parameters α and γ are chosen such that (2.1) holds. The most popular in application of multivariate reference region is the γ-ellipsoid which is then a special case in our design. This allows the user to design various rotational invariant multivariate reference regions.

The rectangular type γ reference region by Wellek (2011) in the form

$$C_{rect} = \{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : F_{X_1}^{-1}(q_\gamma) \leq x_1 \leq F_{X_1}^{-1}(1-q_\gamma), F_{X_2}^{-1}(q_\gamma) \leq x_2 \leq F_{X_2}^{-1}(1-q_\gamma)\}$$

is not a case of our transformed multivariate reference region. This rectangular type multivariate reference region does not satisfy any interesting rotational invariant property. In an attempt of power comparison, we hereafter consider only the parallel, rectangular reference regions and ellipsoid.

We consider that X has a bivariate normal distribution $N_2(\mu, \Sigma)$ as an example for interpretation of three reference regions. By letting $\mu = (0,0)'$ and $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$, we present pictures C_x, C_{rect} and C_{ellip} for several σ_{12} and two $\sigma_1 = \sigma_2$ in Figures 1 and 2.

Firgures 1 and 2 are here

If we consider smallest volume reference region, then we can expect different shapes for different distributions. For example, ellipsoid is better than the others when the distribution is bivariate normal while the result is opposite when the distribution is bivariate exponential or chi-square. For comparison, we compute true areas of these three γ-reference regions under some specified values of covariate σ_{12}. The results are listed in Table 1.

Table 1. True areas for three bivariate reference regions $(1 - 2\alpha = 0.9, \gamma = 0.81)$

Covariance	Ellipsoid	Rectangle	Parallelogram
$\sigma_1^2 = 0.3$			
$\sigma_{12} = 0.05$	3.09	3.23	3.20
$\sigma_{12} = 0.12$	2.87	3.15	2.98
$\sigma_{12} = 0.15$	2.71	3.09	2.81
$\sigma_{12} = 0.25$	1.73	2.73	1.79
$\sigma_1^2 = 1$			
$\sigma_{12} = 0.3$	9.95	10.64	10.32
$\sigma_{12} = 0.5$	9.037	10.29	9.37
$\sigma_{12} = 0.7$	7.45	9.70	7.73
$\sigma_{12} = 0.9$	4.55	8.63	4.72

In this setting of normal distribution, ellipsoid is uniformly better than the parallel multivariate reference region. The rectangular one is the poorest.

3. Estimations of Multivariate Reference Region

The multivariate reference region can be estimated either parametrically or non-parametrically. Suppose that we have a random sample $\begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix}, i = 1, ..., n$ from

$f(x_1, x_2)$. The parametric method assumes that the underlying distribution $f(x_1, x_2)$ is the form $f_\theta(x_1, x_2)$ with known function f but unknown parameters θ so that the population mean, covariance matrix and population quantiles are functions of parameters θ such as μ_θ, Σ_θ and $F^{-1}_{Y_1,\theta}(\gamma_1)$ and $F^{-1}_{Y_2,\theta}(\gamma_2)$ respectively. Formulation of the parallel reference region is

$$C_x = \{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \Sigma_\theta^{1/2} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \mu_\theta, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in C_y\} \tag{3.1}$$

with

$$C_y = \{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} : F^{-1}_{Y_1,\theta}(\alpha) \le y_1 \le F^{-1}_{Y_1,\theta}(1-\alpha), F^{-1}_{Y_2,\theta}(\alpha) \le y_2 \le F^{-1}_{Y_2,\theta}(1-\alpha)\}\}$$

where α satisfies $\gamma = P\{\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \in C_y\}$. The parametric estimator of parallel γ reference region is

$$\hat{C}_{p,x} = \{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \Sigma_{\hat{\theta}}^{1/2} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \mu_{\hat{\theta}}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in \hat{C}_{p,y}\} \tag{3.2}$$

with

$$\hat{C}_{p,y} = \{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} : F^{-1}_{Y_1,\hat{\theta}}(\alpha) \le y_1 \le F^{-1}_{Y_1,\hat{\theta}}(1-\alpha), F^{-1}_{Y_2,\hat{\theta}}(\alpha) \le y_2 \le F^{-1}_{Y_2,\hat{\theta}}(1-\alpha)\}\}$$

when $\hat{\theta}$ is an available estimator of θ. We say that $\hat{C}_{p,x}$ is the maximum likelihood estimator (mle) of C_x when $\hat{\theta}$ is mle of θ. For some distributions such as normal distribution, the γ confidence set C_y is free of parameters and then we let $\hat{C}_{p,y} = C_y$.

We consider estimations of population mean and population covariance matrix by sample mean and sample covariance matrix as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \text{ and } \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} X_{1i} - \hat{\mu}_1 \\ X_{2i} - \hat{\mu}_2 \end{pmatrix} \begin{pmatrix} X_{1i} - \hat{\mu}_1 \\ X_{2i} - \hat{\mu}_2 \end{pmatrix}'.$$

With the transformed sample $\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} = \hat{\Sigma}^{-1/2'} \begin{pmatrix} X_{1i} - \hat{\mu}_1 \\ X_{2i} - \hat{\mu}_2 \end{pmatrix}, i = 1, ..., n$, the empirical distribution functions for Y observations are $\hat{F}_{Y_1}(y_1) = \frac{1}{n}\sum_{i=1}^{n} I(Y_{1i} \leq y_1)$ and $\hat{F}_{Y_2}(y_2) = \frac{1}{n}\sum_{i=1}^{n} I(Y_{2i} \leq y_2)$ that allows us to estimate the population quantiles $F_{Y_1}^{-1}(\gamma_1)$ and $F_{Y_2}^{-1}(\gamma_2)$ of (2.1) by empirical quantiles $\hat{F}_{Y_1}^{-1}(\gamma_1)$ and $\hat{F}_{Y_2}^{-1}(\gamma_2)$. Then estimator of nonparametric multivariate γ reference region is

$$\hat{C}_{np,x} = \{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \hat{\Sigma}^{1/2} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \hat{\mu}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in \hat{C}_{np,y}\}. \tag{3.3}$$

with $\hat{C}_{np,y}$ the estimator of $C_{np,y}$ as

$$\hat{C}_{np,y} = \{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} : \hat{F}_{Y_1}^{-1}(\alpha) \leq y_1 \leq \hat{F}_{Y_1}^{-1}(1-\alpha), \hat{F}_{Y_2}^{-1}(\alpha) \leq y_2 \leq \hat{F}_{Y_2}^{-1}(1-\alpha)\} \tag{3.4}$$

where α is chosen satisfying $\frac{1}{n}\sum_{i=1}^{n} I(\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} \in \hat{C}_{np,y}) \approx \gamma$. Parametric estimator of (3.2) and nonparametric estimator of (3.3) both have unknown quantity of (3.1) as target for estimation.

Suppose that X_1 and X_2 have distribution functions F_{X_1,θ_1} and F_{X_2,θ_2} respectively. The rectangular type γ reference region of Wellek (2011) may be formed as

$$C_{rect} = \{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : F_{X_1,\theta_1}^{-1}(q_\gamma) \leq x_1 \leq F_{X_1,\theta_1}^{-1}(1-q_\gamma), F_{X_2,\theta_2}^{-1}(q_\gamma) \leq x_2 \leq F_{X_2,\theta_2}^{-1}(1-q_\gamma)\} \tag{3.5}$$

where constant q_γ is chosen to fulfill $\gamma = P\{\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \in C_{rect}\}$. Suppose that estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ are available for unknown parameters θ_1 and θ_2. The parametric rectangular type γ reference region is

$$\hat{C}_{p,rect} = \{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : F_{X_1,\hat{\theta}_1}^{-1}(q_\gamma) \leq x_1 \leq F_{X_1,\hat{\theta}_1}^{-1}(1-q_\gamma), F_{X_2,\hat{\theta}_2}^{-1}(q_\gamma) \leq x_2 \leq F_{X_2,\hat{\theta}_2}^{-1}(1-q_\gamma)\}. \tag{3.6}$$

If $\hat{F}_{X_1}^{-1}$ and $\hat{F}_{X_2}^{-1}$ represent, respectively, the empirical quantile functions for variables X_1 and X_2. The nonparametric rectangular type γ reference region is

$$\hat{C}_{np,rect} = \{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \hat{F}_{X_1}^{-1}(q_\gamma) \le x_1 \le \hat{F}_{X_1}^{-1}(1-q_\gamma), \hat{F}_{X_2}^{-1}(q_\gamma) \le x_2 \le \hat{F}_{X_2}^{-1}(1-q_\gamma)\} \tag{3.7}$$

where q_γ satisfies $\gamma \approx \frac{1}{n}\sum_{i=1}^{n} I(\hat{F}_{X_1}^{-1}(q_\gamma) \le x_{1i} \le \hat{F}_{X_1}^{-1}(1-q_\gamma), \hat{F}_{X_2}^{-1}(q_\gamma) \le x_{2i} \le \hat{F}_{X_2}^{-1}(1-q_\gamma)\}$.

In practical application, parametric estimation of multivariate γ reference region needs first derive the explicit forms of $C_{p,y}$ and $C_{p,x}$ for implementation of imposing estimate $\hat{\theta}$ into the equations for regions. However, the nonparametric estimation of multivariate γ reference region is given in straight way requiring only quantiles $\hat{F}_{Y_1}^{-1}$, $\hat{F}_{Y_2}^{-1}$ and $(\hat{\mu}, \hat{\Sigma})$.

4. Efficiency Evaluation for Estimators of Multivariate Reference Regions

Now, suppose that random sample $\begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix}, i = 1, ..., n$ is drawn from the bivariate normal distribution $N_2(\mu, \Sigma)$. The parametric type unknown multivariate reference region is C_x of (2.2) with

$$C_y = \{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} : -\Phi^{-1}(\frac{1+\delta}{2}) \le y_j \le \Phi^{-1}(\frac{1+\delta}{2}), j = 1, 2\}$$

where Φ^{-1} represents the quantile function of the standard normal distribution. Hence, C_y is a known region without need of estimation and the mle of C_x is

$$\hat{C}_{p,x} = \{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \hat{\Sigma}^{1/2}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \hat{\mu}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in C_{p,y}\} \tag{4.1}$$

since estimator of $\theta = \{\mu, \Sigma\}$ is $\hat{\theta} = \{\hat{\mu}, , \hat{\Sigma}\}$ of (3.2). It is interesting to compare parametric and nonparametric estimators $\hat{C}_{p,x}$ of (4.1) and $\hat{C}_{np,x}$ of (3.3).

We further denote $\mu = (\mu_1, \mu_2)'$ and let σ_1^2 and σ_2^2 be variances of X_1 and X_2 respectively. The rectangular reference region is

$$C_{p,rect} = \{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \mu_1 + \Phi^{-1}(q_\gamma)\sigma_1 \le x_1 \le \mu_1 + \Phi^{-1}(1-q_\gamma)\sigma_1,$$
$$\mu_2 + \Phi^{-1}(q_\gamma)\sigma_2 \le x_2 \le \mu_2 + \Phi^{-1}(1-q_\gamma)\sigma_2 \}$$

where q_γ satisfies $\gamma = P\{ \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \in C_{p,rect} \}$ with estimator

$$\hat{C}_{p,rect} = \{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \hat{\mu}_1 + \Phi^{-1}(q_\gamma)\hat{\sigma}_1 \le x_1 \le \hat{\mu}_1 + \Phi^{-1}(1-q_\gamma)\hat{\sigma}_1,$$
$$\hat{\mu}_2 + \Phi^{-1}(q_\gamma)\hat{\sigma}_2 \le x_2 \le \hat{\mu}_2 + \Phi^{-1}(1-q_\gamma)\hat{\sigma}_2 \} \qquad (4.2)$$

With nonparametric estimates (3.3) and (3.7) and parametric estimates (4.1) and (4.2), we set $1 - 2\alpha = 0.9$ and replications $m = 10,000$ to generate random sample of size n from normal distribution $N_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix})$ and we denote the areas of jth parametric and nonparametric estimate of parallel reference regions by $A_{p,x}^j$ and $A_{np,x}^j$ and those of rectangular reference regions by $A_{p,rect}^j$ and $A_{np,rect}^j$. We then have averaged areas as

$$A_{p,x} = \frac{1}{m}\sum_{j=1}^{m} A_{p,x}^j, \; A_{np,x} = \frac{1}{m}\sum_{j=1}^{m} A_{np,x}^j, \; A_{p,rect} = \frac{1}{m}\sum_{j=1}^{m} A_{p,rect}^j$$
$$\text{and } A_{np,rect} = \frac{1}{m}\sum_{j=1}^{m} A_{np,rect}^j.$$

The simulated results of these averaged areas are dispalyed in Table 2.

Table 2. Simulated averaged areas of four multivariate reference regions ($\sigma_1 = 1$)

σ_{12}	$n = 30$	$n = 50$	$n = 100$
$\sigma_{12} = 0.3$			
$A_{p,x}(10.32)$	9.647	9.915	10.11
$A_{np,x}$	9.795	10.01	10.19
$A_{p,rect}(10.64)$	9.450	9.809	10.39
$A_{np,rect}$	13.12	12.22	11.21
$\sigma_{12} = 0.5$			
$A_{p,x}(9.372)$	8.737	8.973	9.183
$A_{np,x}$	8.873	9.068	9.260
$A_{p,rect}(10.29)$	9.126	9.450	10.06
$A_{np,rect}$	11.97	11.39	10.92
$\sigma_{12} = 0.7$			
$A_{p,x}(7.728)$	7.220	7.416	7.579
$A_{np,x}$	7.346	7.491	7.638
$A_{p,rect}(9.696)$	8.633	8.925	9.484
$A_{np,rect}$	11.03	10.53	10.29
$\sigma_{12} = 0.9$			
$A_{p,x}(4.717)$	4.396	4.537	4.618
$A_{np,x}$	4.463	4.582	4.653
$A_{p,rect}(8.631)$	7.597	7.928	8.430
$A_{np,rect}$	10.20	9.277	9.148

We have conclusions for the results in Table 2:

(a) The parametric reference region estimator, parallel or rectangular, have areas mostly smaller than nonparametric reference region estimator. Hence, we propose to apply the parametric reference regions when the underlying distribution is known.

(b) When $\sigma_{12} = 0.5$ and 0.9 the parallel reference region estimators, parametric or nonparametric, has area smaller than it of the rectangular reference region estimator. This show that we should apply the parallel reference region when the covariance σ_{12} is large.

Table 3. Simulated averaging areas of four multivariate reference regions ($\sigma_1^2 = 0.3$)

σ_{12}	$n = 30$	$n = 50$	$n = 100$
$\sigma_{12} = 0.05$			
$A_{p,x}(3.201)$	2.986	3.075	3.134
$A_{np,x}$	3.034	3.116	3.158
$A_{p,rect}(3.232)$	2.865	2.983	3.158
$A_{np,rect}$	4.084	3.776	3.399
$\sigma_{12} = 0.12$			
$A_{p,x}(2.975)$	2.776	2.859	2.919
$A_{np,x}$	2.816	2.884	2.942
$A_{p,rect}(3.147)$	2.792	2.903	3.083
$A_{np,rect}$	3.767	3.557	3.353
$\sigma_{12} = 0.15$			
$A_{p,x}(2.811)$	2.610	2.799	2.757
$A_{np,x}$	2.653	2.803	2.774
$A_{p,rect}(3.090)$	2.730	3.080	3.019
$A_{np,rect}$	3.586	3.112	3.286
$\sigma_{12} = 0.25$			
$A_{p,x}(1.794)$	2.633	2.703	2.756
$A_{np,x}$	2.671	2.731	2.775
$A_{p,rect}(2.729)$	2.750	2.843	3.017
$A_{np,rect}$	3.596	3.434	3.284

Again, nonparametric parallel reference region estimator has area smaller uniformly than the nonparametric rectangular reference region estimator. For larger σ_{12} (0.12, 0.15, 0.25), the parametric parallel reference region estimator has uniformly smaller area than the parametric rectangular reference region.

Next, we evaluate the efficiencies of these four estimators by computing simulated mean squares errors as

$$\text{MSE}_p = \frac{1}{m}\sum_{j=1}^{m}(A^j_{p,x} - A^0_{p,x})^2,\ \text{MSE}_{np} = \frac{1}{m}\sum_{j=1}^{m}(A^j_{np,x} - A^0_{p,x})^2$$

$$\text{MSE}_{p,rect} = \frac{1}{m}\sum_{j=1}^{m}(A^j_{p,rect} - A^0_{rect})^2,\ \text{MSE}_{np,rect} = \frac{1}{m}\sum_{j=1}^{m}(A^j_{np,rect} - A^0_{rect})^2$$

where $A^0_{p,x}$ and A^0_{rect} are areas of true regions $A_{p,x}$ of (4.1) and C_{rect} of (3.5).

Table 4. Mean square error for (parallel) rectangle reference regions

σ_{12}	$n = 30$	$n = 50$	$n = 100$
$\sigma_1^2 = 0.3, \sigma_{12} = 0.05$			
MSE_p	0.362	0.212	0.103
MSE_{np}	0.496	0.300	0.158
$\text{MSE}_{p,rect}$	0.590	0.351	0.170
$\text{MSE}_{np,rect}$	2.522	0.948	0.216
$\sigma_{12} = 0.12$			
MSE_p	0.323	0.185	0.088
MSE_{np}	0.436	0.260	0.133
$\text{MSE}_{p,rect}$	0.597	0.353	0.168
$\text{MSE}_{np,rect}$	1.863	0.772	0.242
$\sigma_{12} = 0.15$			
MSE_p	0.281	0.015	0.081
MSE_{np}	0.380	0.023	0.122
$\text{MSE}_{p,rect}$	0.586	0.033	0.174
$\text{MSE}_{np,rect}$	1.439	0.034	0.252
$\sigma_{12} = 0.25$			
MSE_p	0.275	0.168	0.079
MSE_{np}	0.377	0.240	0.120
$\text{MSE}_{p,rect}$	0.455	0.320	0.248
$\text{MSE}_{np,rect}$	1.947	1.060	0.515
$\sigma_1 = 1, \sigma_{12} = 0.3$			
MSE_p	3.733	2.204	1.070
MSE_{np}	5.105	3.180	1.608
$\text{MSE}_{p,rect}$	6.498	3.977	1.882
$\text{MSE}_{np,rect}$	24.91	9.775	2.455
$\sigma_{12} = 0.5$			
MSE_p	3.140	1.823	0.921
MSE_{np}	4.330	2.546	1.369
$\text{MSE}_{p,rect}$	6.411	3.942	1.951
$\text{MSE}_{np,rect}$	16.15	7.158	2.627
$\sigma_{12} = 0.7$			
MSE_p	2.137	1.267	0.613
MSE_{np}	2.929	1.811	0.914
$\text{MSE}_{p,rect}$	6.238	3.939	1.909
$\text{MSE}_{np,rect}$	11.20	5.623	2.495
$\sigma_{12} = 0.9$			
MSE_p	0.804	0.457	0.220
MSE_{np}	1.095	0.663	0.330
$\text{MSE}_{p,rect}$	5.820	3.580	1.719
$\text{MSE}_{np,rect}$	11.39	4.474	2.245

We have two conclusions for the results in Table 4:

(a) Parametric parallel reference region has MSE's uniformly smaller than other

three versions. This supports to apply the parametric parallel reference region when the distribution is known normal.

(b) The nonparametric parallel reference region has MSE's uniformly smaller than the nonparametric rectangular reference region. This then supports to apply the nonparametric parallel reference region when the distribution is unknown.

5. Evaluation of Error Probabilities for Estimators of Multivariate Reference Regions

Laboratory test results are commonly compared to a reference region (interval) before caregivers make physiological assessments, medical diagnoses, or management decisions. An individual who is being screened for a disorder based on a measurement is suspected to be abnormal if his/her measurement value lies outside the reference region. The reference region plays exactly the role of acceptance region in hypothesis testing. Hence, the quality of a reference region relies on its accuracy in detection of abnormality that can be studied with the probabilities of two types of error.

With estimates $\hat{C}_{p,x}$ and $\hat{C}_{np,x}$, the type 1 error probabilities with parametric estimation and nonparametric estimation from repeated samples $\begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix}, i = 1, ..., k$ are defined as

$$p^0_{p,err1} = \frac{1}{k}\sum_{i=1}^{k} I(\begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \in C_{p,x}, \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \notin \hat{C}_{p,x})$$

$$p^0_{n,err1} = \frac{1}{k}\sum_{i=1}^{k} I(\begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \in C_{p,x}, \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \notin \hat{C}_{np,x})$$

with $p^0_{p,err1}$ the parametric estimation and $p^0_{n,err1}$ the non-parametric estimation. This measures the probability that the individual is expected to be identified normal since the measurement is truely lies in true reference region but actually this

measurement falls out estimate $\hat{C}_{p,x}$ and $\hat{C}_{np,x}$. Now, if we have m replications in generating estimates $\hat{C}_{p,x}$ and $\hat{C}_{np,x}$, we have errpr probabilities estimates $p^j_{p,err1}$ and $p^j_{n,err1}, j = 1, ..., m$. We then have averaged estimates of error probabilities as

$$p_{p,err1} = \frac{1}{m}\sum_{j=1}^{m} p^j_{p,err1}, \quad \text{and } p_{n,err1} = \frac{1}{m}\sum_{j=1}^{m} p^j_{n,err1}.$$

Considering the parallel reference region only, results of the two error probabilities from a simulation study under the normal distribution are displayed in Table 5.

Table 5. Type 1 error probabilities for parallel reference regions (k=m=10,000)

σ_{12}	$n = 30$	$n = 50$	$n = 100$
$\sigma_1^2 = 0.3, \sigma_{12} = 0.05$			
$p_{p,err1}$	0.073	0.050	0.032
$p_{n,err1}$	0.091	0.064	0.042
$\sigma_{12} = 0.12$			
$p_{p,err1}$	0.073	0.050	0.031
$p_{n,err1}$	0.091	0.065	0.042
$\sigma_{12} = 0.15$			
$p_{p,err1}$	0.074	0.051	0.031
$p_{n,err1}$	0.092	0.065	0.042
$\sigma_{12} = 0.25$			
$p_{p,err1}$	0.072	0.050	0.031
$p_{n,err1}$	0.091	0.064	0.042
$\sigma_1^2 = 1, \sigma_{12} = 0.3$			
$p_{p,err1}$	0.073	0.050	0.031
$p_{n,err1}$	0.091	0.065	0.042
$\sigma_{12} = 0.5$			
$p_{p,err1}$	0.073	0.051	0.032
$p_{n,err1}$	0.091	0.065	0.042
$\sigma_{12} = 0.7$			
$p_{p,err1}$	0.073	0.051	0.031
$p_{n,err1}$	0.091	0.065	0.042

Two conclusions are available for the results in Table 5:

(a) Two error probabilities achieve smallest when the sample size is 100, the largest one.

(b) The error probabilities under parametric estimator is uniformly smaller than

that under the nonparametric estimator.

The type 2 error probabilities with parametric estimate $\hat{C}_{p,x}$ and nonparametric estimate $\hat{C}_x$ are defined as

$$p^0_{p,err2} = \frac{1}{k}\sum_{i=1}^{k} I(\begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \in \hat{C}_{p,x}, \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \notin C_{p,x})$$

$$p^0_{n,err2} = \frac{1}{k}\sum_{i=1}^{k} I(\begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \in \hat{C}_{np,x}, \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \notin C_{p,x}).$$

The average estimates of these two error probabilities in m replications are

$$p_{p,err2} = \frac{1}{m}\sum_{j=1}^{m} p^j_{p,err2}, \quad \text{and } p_{n,err2} = \frac{1}{m}\sum_{j=1}^{m} p^j_{n,err2}.$$

where $p^j_{p,err2}$ and $p^j_{n,err2}$ represent the estimates of $p^0_{p,err2}$ and $p^0_{n,err2}$ at jth replication. The simulated results for this type 2 error probabilities are displayed in Table 6.

Table 6. Type 2 error probabilities (k=m=10,000) ($\sigma_1 = 1$)

σ_{12}	$n = 30$	$n = 50$	$n = 100$
$\sigma_{12} = 0.3$			
p_p	0.024	0.021	0.017
p_{np}	0.031	0.027	0.023
$p_{p,rect}$	0.020	0.017	0.017
$p_{np,rect}$	0.057	0.047	0.030
$\sigma_{12} = 0.5$			
p_p	0.023	0.021	0.017
p_{np}	0.030	0.027	0.023
$p_{p,rect}$	0.019	0.017	0.017
$p_{np,rect}$	0.049	0.041	0.031
$\sigma_{12} = 0.7$			
p_p	0.024	0.021	0.017
p_{np}	0.031	0.027	0.023
$p_{p,rect}$	0.019	0.016	0.016
$p_{np,rect}$	0.044	0.037	0.029
$\sigma_{12} = 0.9$			
p_p	0.023	0.021	0.017
p_{np}	0.030	0.028	0.023
$p_{p,rect}$	0.017	0.016	0.015
$p_{np,rect}$	0.045	0.033	0.026

We have several comments on the results in Table 6:

(a) Parametric rectangular reference region is generally better than the parametric parallel reference region.

(b) When we consider a nonparametric estimation, the parallel reference region estimator is more powerful than the rectangle reference region estimator.

(c) Since $\pi_{p,err2} = 1 - p_{p,err2}$ and $\pi_{n,err2} = 1 - p_{n,err2}$ are considered the powers in detection of abnormality, four reference regions are satisfactory since their powers are all larger than 0.9.

References

Chen, L.-A. and Welsh, A. H. (2002). Distribution-function-based bivariate quantiles. *Journal of Multivariate Analysis* **83**, 208-231.

Harris, E. K., Yasaka, T., Horton, M., R. and Shakarji, G. (1982). Comparing multivariate and univariate subject-specific reference regions for blood constituents in healthy persons. *Clinical Chemistry*, 28, 422-426.

Horn, P. S. and Pesce, A. J. (2003). Reference intervals: an update. *Clinica Chimica Acta*, 334, 5-23.

Huang, J.-Y., Chen, L.-A. and Welsh, A.H. (2010). A note on reference limits. *IMS Collections*, Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in Hornor of Professor Jana Jureckova. 7, 84-94.

Schoen, I. and Brooks, S. (1970) Judgement based on 95% confidence limits. *American Journal of Clinical Pathology.* 53, 190-193.

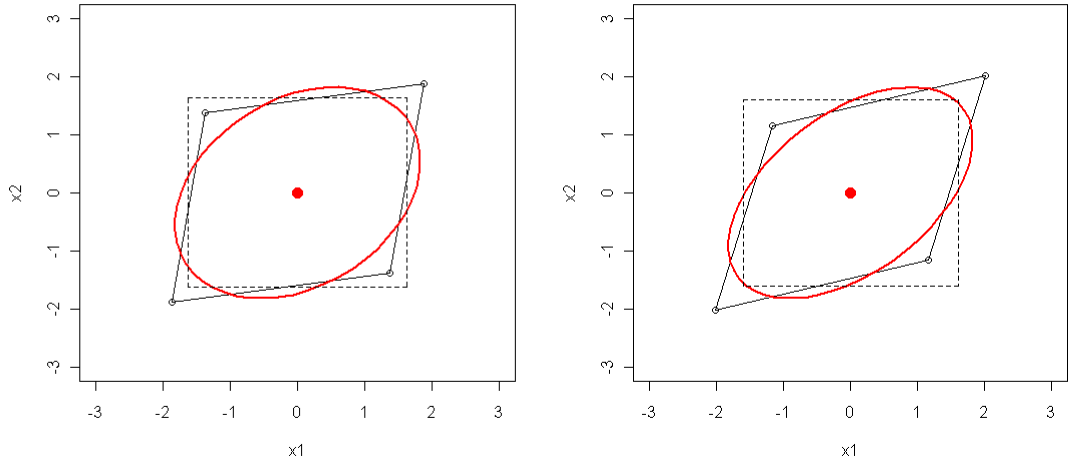
Shiau, J.-J. H. and Chen, L.-A. (2003). The multivariate parallelogram and its ap-

plications to multivariate trimmed mean. *Australian and New Zealand Journal of Statistics* **45**, 343-352.

Wellek, S. (2011). On easily interpretable multivariate reference regions of rectangular shape. *Biometrical Journal* **53**, 491-511.

Figure1. Pictures of three bivariate reference regions $\sigma_1^2 = \sigma_2^2 = 1$

(a) $\sigma_{12} = 0.3$ (b) $\sigma_{12} = 0.5$



(c) $\sigma_{12} = 0.7$ (d) $\sigma_{12} = 0.9$

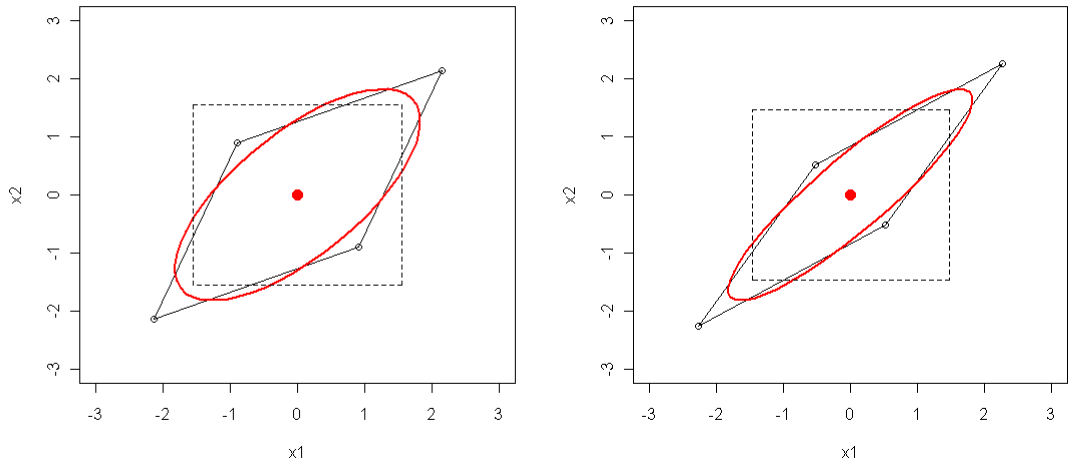
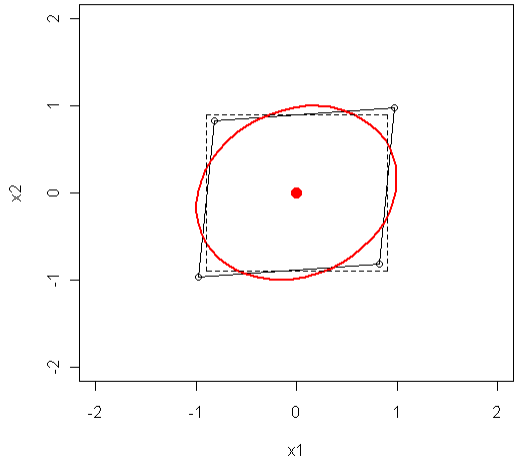
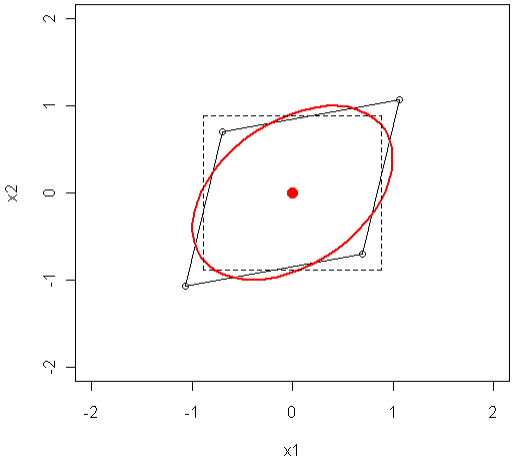


Figure2. Pictures of three bivariate reference regions $\sigma_1^2 = \sigma_2^2 = 0.3$

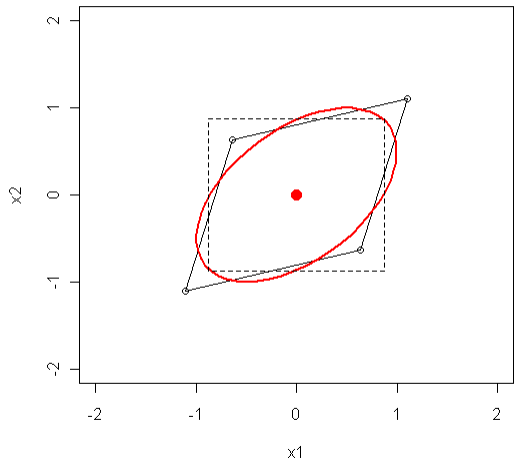
(a) $\sigma_{12} = 0.05$



(b) $\sigma_{12} = 0.12$



(c) $\sigma_{12} = 0.15$



(d) $\sigma_{12} = 0.25$