Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 161-180

Nonparametric Rectangular Prediction Regions for Setting Reference Regions in Laboratory Medicine

Michael Daniel Lucagbo[1,2] and Thomas Mathew[1]
[1] *Department of Mathematics & Statistics, University of Maryland Baltimore County, Baltimore, Maryland, USA*
[2] *School of Statistics, University of the Philippines Diliman, Quezon City, Philippines*

Received: 21 May 2024; Revised: 20 August 2024; Accepted: 22 August 2024

Abstract

The interpretation of clinical laboratory results of patients depends crucially on an established reference interval for each biochemical analyte. Often, the health status of a patient is decided based on the values of multiple analytes, and this calls for the use of a multivariate reference region, so that the possible cross-correlations among the analytes can be taken into consideration. If multivariate normality can be assumed, one of the recommendations in the laboratory medicine literature is to use an ellipsoidal prediction region as a reference region. However, an ellipsoidal region cannot detect if a particular analyte is within the normal range; a rectangular prediction region is necessary for this purpose. Under multivariate normality, rectangular prediction regions are available in the literature, and these can be used as reference regions for assessing the outlyingness of individual analytes. The present work is motivated by the need to construct such regions without making the multivariate normality assumption. Two approaches are pursued in our work: based on Box-Cox transformation of each marginal variable, and based on estimating each marginal density using a kernel density estimator. A non-parametric bootstrap is then employed for estimating the required prediction factors. Through simulations, it is noted that the resulting rectangular prediction regions meet the coverage probability requirements satisfactorily. The methodology can also be adopted for computing one sided prediction limits, or a combination of one-sided prediction limits for some variables, and two-sided prediction intervals for the rest. Algorithms are provided to compute the regions, and illustrative examples are also given.

Key words: Bootstrap; Box-Cox transformation; Kernel density estimator; One-sided prediction limits; Two-sided prediction intervals.

1. Introduction

Reference intervals are used in numerous medical applications such as the interpretation of blood tests, clinical urine tests, vital signs, and so forth. Due to the extensive

Corresponding Author: Thomas Mathew
Email: mathew@umbc.edu

applications of reference intervals in the field of laboratory medicine, Horn and Pesce (2005) have called it "the most widely used medical decision-making tool." A reference interval is defined as the interval that contains 95% of the "central measurements" for a reference population. Thus, the endpoints of a two-sided reference intervals are the 2.5th and 97.5th percentiles of the reference population. If only a one-sided reference limit is of interest, then the required reference limit is either the 95th percentile (for an upper reference limit), or the 5th percentile (for the lower reference limit).

Since the population percentiles are unknown in actual practice, reference ranges are typically constructed based on data from a random sample of individuals (reference subjects). The selection of reference subjects is obviously critically important, and the prevailing view is that the reference subjects should consist of healthy subjects. For example, Wellek (2011) mentions that a population suitable for establishing reference values should consist of people free of the disease condition one aims to detect. A naive way to construct reference intervals using the available data is to use estimated sample percentiles. Under such a scheme, the percentage of the population covered by the resulting reference interval will be different from 95%. As an alternative to this, a common approach is to compute a 95% prediction interval and use it as a reference interval. This approach has been recommended in practice; see the document by the National Committee for Clinical Laboratory Standards (2010) and the User's Guide by Horn and Pesce (2005). Another option, advocated by authors such as Liu et al. (2021) and Lucagbo and Mathew (2023) is to compute a 95% tolerance interval, which can be used to assess the uncertainty in the estimated reference intervals. In this study, we adopt the prediction interval criterion.

For complex diagnoses, such as for kidney function or liver function, several analytes are needed to properly assess the health status of a patient. For such scenarios, the use of separate univariate reference intervals is an inefficient way to proceed since such an approach disregards the cross-correlations among the analytes. Moreover, it increases the risk of false-positive diagnoses (Harris, 1981; Winkel et al., 1972). When multiple analytes are needed to assess the health status of individuals, a multivariate reference region (MRR), which accounts for the cross-correlations among analytes, is needed. Nonetheless, MRRs are not without shortcomings. The conventional approach to compute MRRs, especially under the assumption of multivariate normality, is to construct ellipsoidal regions. Unfortunately, ellipsoidal reference regions are difficult to interpret. Moreover, ellipsoidal reference regions tend to produce false negative results in the presence of only one or two extreme components (Albert and Harris, 1987; Strike, 1991). Finally, ellipsoidal regions are unable to detect component-wise outliers. In other words, whenever patients are diagnosed as non-healthy based on an MRR, no conclusion can be drawn on which specific analyte/s have caused the positive result. For this reason, Wellek (2011) notes that MRRs "have only a marginal role in the practice of clinical chemistry and laboratory medicine."

To address the above difficulties associated with ellipsoidal reference regions, this paper aims to derive rectangular reference regions, which are easily interpretable regions that can detect the outlyingness of specific analytes. In view of the fact that laboratory test results are typically skewed (or at least not normally distributed), we shall derive such regions under a nonparametric framework. Previous work on nonparametric reference regions includes that of Wellek (2011) and Young and Mathew (2020). The work of Wellek (2011) includes both parametric and nonparametric estimation of rectangular reference regions.

Rectangular prediction regions are derived in Young and Mathew (2020); however, fairly large sample sizes are required for meeting the coverage probability requirement.

In the present investigation, we aim to develop rectangular nonparametric prediction regions to be used as reference regions. The methodologies described in this study are based on either transforming the marginal data using a Box-Cox transformation, or estimating the marginal densities through kernel density estimation. The accuracy of these approaches will be assessed by reporting the relevant coverage probabilities. We investigate the performance of the proposed methodologies using sample sizes starting from $n = 50$.

1.1. Rectangular prediction regions

We now define the criterion to be used in obtaining the rectangular nonparametric reference region. Our goal is to find a rectangular reference region of the form (1)

$$[c_1, d_1] \times [c_2, d_2] \times \cdots \times [c_p, d_p]\,, \tag{1}$$

subject to the prediction region criterion in (2)

$$P\left(\bigcap_{i=1}^{p} \{X_i \in [c_i, d_i]\}\right) = 1 - \alpha. \tag{2}$$

It should be clear that the set of intervals $[c_i, d_i]$, $i = 1, 2, \cdots, p$, satisfying the above requirement is not unique. Nevertheless, it is to be expected that each marginal interval $[c_i, d_i]$, $i = 1, 2, \cdots, p$, can be appropriately specified if we know the marginal distributions. In the absence of any information on the marginal distributions, we shall explore two options. The first option is to apply separate Box-Cox transformations to each set of marginal data, so that each marginal distribution is approximately normal. We can now specify a common prediction factor on the transformed scale, which can be estimated via a nonparametric bootstrap subject to the requirement in (2). Details of this appear in Section 2. The second approach, described in Section 3, uses the kernel density estimate (KDE) of the marginal densities, which also leads to a common prediction factor. In the same section, we also extend the KDE idea in order to construct mixed-sided prediction regions. These are regions where some variables have an upper prediction limits and the rest have two-sided prediction intervals. Section 4 gives numerical results on estimated coverage probabilities in order to assess the accuracy of the proposed methodologies and illustrates the methodologies through a real-life example. Section 5 gives some brief concluding remarks.

2. Nonparametric prediction regions using the Box-Cox transformation

Our first strategy to deal with the problem of computing nonparametric rectangular prediction regions is to transform the marginal data so that it has a normal distribution, approximately. Ichihara and Boyd (2010) note that “Since almost all distributions of laboratory test results are non-Gaussian, it is essential to convert these to a Gaussian distribution.” They investigate transformation to normality using the Box-Cox transformation (Box and Cox, 1964) and also a modified Box-Cox formula introduced by Ichihara and Kawai (1997). The International Federation of Clinical Chemisty (IFCC) Expert Committee on Reference Intervals has actually recommended the Box-Cox transformation to normality for the purpose of computing reference intervals (Solberg, 1987). In other words, the idea of using the

Box-Cox transformation in the context of computing reference intervals is already mentioned in the literature, but only in the univariate context.

We shall employ the Box-Cox transformation to develop rectangular reference regions in a nonparametric setup. The prediction factor is computed based on the transformed data, under the assumption that the data are approximately normal. Here we want to point out that we shall apply the Box-Cox transformation to the sample from each univariate marginal distribution. In other words, the transformation is univariate, not multivariate. As will be seen shortly, normality will not be fully utilized when we derive the prediction factor, since we will be employing a nonparametric bootstrap procedure. However, normality is perhaps necessary to justify the use of a common prediction factor.

Suppose that the data $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ consist of a random sample coming from an unknown multivariate distribution with nonnegative support, where

$$\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$$

is a p-variate vector, $i = 1, 2, \ldots, n$. For each fixed $j = 1, 2, \ldots, p$, $X_{1j}, X_{2j} \ldots, X_{nj}$, form a random sample from the univariate marginal distribution of the jth component. We assume that these random variables can be transformed as

$$Y_{ij} = g_j\left(X_{ij}\right) \tag{3}$$

so that their distribution is approximately normal. The transformation could be different for the different components. Once such a transformation has been identified, we can then construct prediction regions for the transformed data Y_{ij}, $i = 1, 2, \ldots, n; j = 1, 2, \ldots, p$. Since the transformed data are assumed to be approximately normal, we restrict the two-sided prediction region to be of the symmetric form

$$\bar{Y}_j \pm \kappa\sqrt{S_{y,jj}} \tag{4}$$

for $j = 1, 2, \ldots, p$, and our goal is to estimate κ. Here $\bar{\boldsymbol{Y}} = \left(\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_p\right)'$ is the sample mean vector, and $S_{y,jj}$ is the jth diagonal element of the sample covariance matrix among the transformed sample values $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})'$, $i = 1, 2, \ldots, n$.

Introduced by Box and Cox (1964), the Box-Cox transformation, as it has come to be called in the statistical literature, is a well-known method to transform skewed data to normality. For a random variable X that assumes positive values, the Box-Cox transformed quantity, say Y, takes the form

$$Y = \begin{cases} \frac{X^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log X, & \lambda = 0 \end{cases} \tag{5}$$

where $\lambda > 1$ for negatively skewed data and $\lambda < 1$ for positively skewed data. The value of λ is to be estimated using the data on X. In this study, λ is estimated through maximum likelihood.

Once a λ_j has been estimated based on $X_{1j}, X_{2j}, \ldots, X_{nj}$, for each fixed $j = 1, 2, \ldots, p$, we shall choose the form of $g_j\left(\cdot\right)$ in (3) to be

$$Y_{ij} = g_j\left(X_{ij}\right) = \begin{cases} X_{ij}^{\lambda_j}, & \lambda_j \neq 0 \\ \log X_{ij}, & \lambda_j = 0, \end{cases} \tag{6}$$

instead of (5). Since X_{ij} assumes positive values, if $\lambda_j \neq 0$ it follows that Y_{ij} in (6) is always positive. Therefore, the back-transformed value Y_{ij}^{1/λ_j} is always defined. This is not always the case with (5), where the back-transformed value is $(\lambda Y + 1)^{1/\lambda}$, which can be undefined if $\lambda Y + 1 < 0$. For this reason, once the value of λ_j is identified, the power transformation in (6) will be adopted. The next goal is to estimate κ to form prediction regions of the form (4).

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ be the future observation to be predicted, and let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_p)' = \left(X_1^{\lambda_1}, X_2^{\lambda_2}, \ldots, X_p^{\lambda_p}\right)'$ be its transformed version. We shall find the value of κ that satisfies

$$P\left(Y_j \in \bar{Y_j} \pm \kappa\sqrt{S_{y,jj}} \quad \forall\, j = 1, 2, \ldots, p\right) = 1 - \alpha.$$

That is,

$$P\left(\left|\frac{Y_j - \bar{Y_j}}{\sqrt{S_{y,jj}}}\right| \leq \kappa \quad \forall\, j = 1, 2, \ldots, p\right) = 1 - \alpha.$$

Equivalently,

$$P\left(\max_{1 \leq j \leq p}\left|\frac{Y_j - \bar{Y_j}}{\sqrt{S_{y,jj}}}\right| \leq \kappa\right) = 1 - \alpha. \tag{7}$$

The choice of a common κ is justified in view of the approximate normality of the marginal components of the $\boldsymbol{Y}$. The statement (7) facilitates the estimation of κ via a nonparametric bootstrap, by sampling with replacement from the collection $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$. Since the λs are unknown parameters, they are also estimated in each bootstrap sample. Algorithm 1 gives the procedure to estimate κ; we shall denote the estimate by k.

2.1. Remarks on back-transforming the data

Caution must be taken in Step 10 of Algorithm 1, where the prediction region is transformed back to the original scale. First of all, since for $\lambda_j \neq 0$, $Y_{ij} = X_{ij}^{\lambda_j}$ where X_{ij}s are positive, the prediction regions in the Y_{ij} scale should contain only nonnegative limits. In some instances, however, the lower limit of an interval $\bar{Y_j} - k\sqrt{S_{jj,y}}$ could be negative. In such a case, we recommend that the lower limit just be changed to 0.

It is possible for $\hat{\lambda}_j$ to be negative. Whenever this happens concurrently with a negative lower limit, then the quantity

$$\left(\bar{Y_j} - k\sqrt{S_{jj,y}}\right)^{1/\hat{\lambda}_j} \tag{10}$$

is undefined, even when the lower limit is changed to 0. The case where both lower limit and λ_j are negative occurs rarely in the simulations, but it occurs more often when the sample size is small than when large, presumably because λ_j cannot be estimated accurately from a small sample. Such a phenomenon never occurred in the simulations included in this paper, but the authors have seen it occur when the sample size is small (such as when $n = 30$). Nonetheless, such an occurrence is still highly unlikely when p is small (*e.g.*, 2 or 3). Since

Algorithm 1 Nonparametric prediction regions using the Box-Cox transformation

1. Let $\boldsymbol{X}_1, \boldsymbol{X}_2 \ldots, \boldsymbol{X}_n$ be the data and write $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $i = 1, 2, \ldots, n$.

2. For each $j = 1, 2, \ldots, p$, estimate the Box-Cox transformation parameter λ_j for the observations $X_{1j}, X_{2j}, \ldots, X_{nj}$. Let $\hat{\lambda}_j$ be the estimated value of λ_j.

3. For each $\hat{\lambda}_j$ in Step 2, compute
$$Y_{ij} = \begin{cases} X_{ij}^{\hat{\lambda}_j}, & \hat{\lambda}_j \neq 0 \\ \log X_{ij}, & \hat{\lambda}_j = 0 \end{cases}, \quad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, p.$$
Define $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})'$, $i = 1, 2, \ldots, n$.

4. Take B random samples with replacement of size $n+1$ from the collection $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$, call these $\boldsymbol{X}_{1b}^*, \boldsymbol{X}_{2b}^*, \ldots \boldsymbol{X}_{nb}^*, \boldsymbol{X}_b^*$, where $b = 1, 2, \ldots, B$. Write $\boldsymbol{X}_b^* = \left(X_{1b}^*, X_{2b}^*, \ldots, X_{pb}^*\right)'$ and $\boldsymbol{X}_{ib}^* = \left(X_{i1b}^*, X_{i2b}^*, \ldots, X_{ipb}^*\right)'$, $i = 1, 2, \ldots, n$, $b = 1, 2, \ldots, B$.

5. For each bootstrap sample in Step 4, estimate the transformation parameter for the observations in the jth column of the data matrix, and denote this estimate by $\hat{\lambda}_{jb}^*$.

6. For each $\hat{\lambda}_{jb}^*$ in Step 5, compute
$$Y_{ijb}^* = \begin{cases} \left(X_{ijb}^*\right)^{\hat{\lambda}_{jb}^*}, & \hat{\lambda}_{jb}^* \neq 0 \\ \log X_{ijb}^*, & \hat{\lambda}_{jb}^* = 0 \end{cases}, \quad \text{and} \quad Y_{jb}^* = \begin{cases} \left(X_{jb}^*\right)^{\hat{\lambda}_{jb}^*}, & \hat{\lambda}_{jb}^* \neq 0 \\ \log X_{jb}^*, & \hat{\lambda}_{jb}^* = 0 \end{cases},$$
where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$, $b = 1, 2, \ldots, B$. Write $\boldsymbol{Y}_{ib}^* = \left(Y_{i1b}^*, Y_{i2b}^*, \ldots, Y_{ipb}^*\right)'$ and $\boldsymbol{Y}_b^* = \left(Y_{1b}^*, Y_{2b}^*, \ldots, Y_{pb}^*\right)'$.

7. Compute $k_b^* = \max\limits_{1 \le j \le p} \left| \frac{Y_{jb}^* - \bar{Y}_{jb}^*}{\sqrt{S_{b,jj}^*}} \right|$, $b = 1, 2, \ldots, B$, where $\bar{Y}_{jb}^*$ is the jth element in the sample mean of $\boldsymbol{Y}_{1b}^*, \boldsymbol{Y}_{2b}^*, \ldots, \boldsymbol{Y}_{nb}^*$; Y_{jb}^* is the jth component of $\boldsymbol{Y}_b^*$; and $S_{b,jj}^*$ is the jth diagonal element in the sample covariance matrix of $\boldsymbol{Y}_{1b}^*, \boldsymbol{Y}_{2b}^*, \ldots \boldsymbol{Y}_{nb}^*$.

8. Compute k as the $(1-\alpha)$-quantile of $k_1^*, k_2^*, \ldots, k_B^*$.

9. The prediction region for the transformed data is given by
$$\left[\bar{Y}_1 \pm k\sqrt{S_{y,11}}\right] \times \left[\bar{Y}_2 \pm k\sqrt{S_{y,22}}\right] \times \cdots \times \left[\bar{Y}_p \pm k\sqrt{S_{y,pp}}\right], \tag{8}$$
where $\bar{Y}_j$ and $S_{y,jj}$ are respectively the jth element and jth diagonal element in the sample mean vector and sample covariance matrix of $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$.

10. Finally, the prediction region for the original data is given by
$$\left[\bar{Y}_1 \pm k\sqrt{S_{y,11}}\right]^{1/\hat{\lambda}_1} \times \left[\bar{Y}_2 \pm k\sqrt{S_{y,22}}\right]^{1/\hat{\lambda}_2} \times \cdots \times \left[\bar{Y}_p \pm k\sqrt{S_{y,pp}}\right]^{1/\hat{\lambda}_p}, \tag{9}$$
where we define $[a, b]^{1/q} = \left[a^{1/q}, b^{1/q}\right]$ if $q > 0$, $\left[b^{1/q}, a^{1/q}\right]$ if $q < 0$, and $\left[e^a, e^b\right]$ if $q = 0$.

(10) is undefined when we run into this situation, we can address this problem by redefining the interval corresponding to the particular component of $\boldsymbol{X}$, say the jth. For example, it is reasonable to redefine it as

$$\left[\left(\bar{Y}_j + k\sqrt{S_{jj,y}}\right)^{1/\hat{\lambda}_j}, \infty\right).$$

We therefore end up with a prediction region that is a mix of one and two-sided intervals (called a mixed-sided reference region in Section 5). It is not clear how we can compare expected volumes in such a scenario.

The reference region given in (9) can also exhibit erratic behaviors, for example, when at least one of $\bar{Y}_j - k\sqrt{S_{jj,y}}$ and $\bar{Y}_j + k\sqrt{S_{jj,y}}$ is very close to 0 and $\hat{\lambda}_j < 0$. Ideally, having a large sample size is the best remedy to avoid the above undesirable behaviors of reference regions. When this is not possible (for example, due to cost considerations), we would like to recommend the following: instead of back-transforming to the original scale, do a "forward-transform" of the future observation to see if it falls inside the reference range in the transformed scale. That is, whenever $\hat{\lambda}_j \neq 0$, consider the transformation

$$(X_1, X_2, \ldots, X_p)' \longmapsto \left(X_1^{\hat{\lambda}_1}, X_2^{\hat{\lambda}_2}, \ldots, X_p^{\hat{\lambda}_p}\right)'$$

and when $\hat{\lambda}_j = 0$, use $X_j \longmapsto \log X_j$, and then use the region in (8) as the reference region, instead of (9). We emphasize that the limits in (8) are always defined.

2.2. One-sided prediction regions

Modifications of Algorithm 1 that are necessary to compute one-sided regions are straightforward whenever we do not run into problems involving the sign of λ_j or the lower limit, or when we choose to adopt the reference region in the transformed scale. To compute a one-sided upper prediction region, we first estimate the prediction factor κ that satisfies

$$\begin{aligned} &P\left(Y_j \leq \bar{Y}_j + \kappa\sqrt{S_{y,jj}} \quad \forall\, j = 1, 2, \ldots, p\right) = 1 - \alpha \\ &\Longleftrightarrow P\left(\max_{1\leq j\leq p} \frac{Y_j - \bar{Y}_j}{\sqrt{S_{y,jj}}} \leq \kappa\right) = 1 - \alpha. \end{aligned} \tag{11}$$

Condition (11) implies that the modification is to be done in Step 7 of Algorithm 1, in which the quantity k_b will be redefined as

$$k_b = \max_{1\leq j\leq p} \frac{Y_{jb}^* - \bar{Y}_{jb}^*}{\sqrt{S_{jj,b}^*}}.$$

For a $(1-\alpha)$-one-sided upper prediction region, we can take k to be the $(1-\alpha)$-quantile of $k_1, k_2, \ldots, k_B$, and the prediction region in the $\boldsymbol{Y}$-scale is given by

$$\left(-\infty, \bar{Y}_1 + k\sqrt{S_{11,y}}\right] \times \cdots \times \left(-\infty, \bar{Y}_p + k\sqrt{S_{pp,y}}\right].$$

If $\hat{\lambda}_j > 0$ for all $j = 1, \ldots, p$, the prediction region in the $\boldsymbol{X}$-scale can then be defined as

$$\left(-\infty, \left(\bar{Y}_1 + k\sqrt{S_{11,y}}\right)^{1/\hat{\lambda}_1}\right] \times \cdots \times \left(-\infty, \left(\bar{Y}_p + k\sqrt{S_{pp,y}}\right)^{1/\hat{\lambda}_p}\right].$$

When $\hat{\lambda}_j < 0$, the corresponding univariate reference limit simply becomes a lower limit instead of an upper limit. Finally, if $\hat{\lambda}_j = 0$, then the corresponding interval becomes

$$\left(-\infty, \exp\left(\bar{Y}_j + k\sqrt{S_{jj,y}}\right)\right].$$

Similarly, for a $(1-\alpha)$-one-sided lower prediction region, we note that the prediction factor should satisfy

$$P\left(\bar{Y}_j + \kappa\sqrt{S_{y,jj}} \leq Y_j \quad \forall\, j = 1, 2, \ldots, p\right) = 1 - \alpha$$
$$\iff P\left(\min_{1\leq j\leq p} \frac{Y_j - \bar{Y}_j}{\sqrt{S_{y,jj}}} < \kappa\right) = \alpha.$$

Thus we change the definition of k_b in Step 7 of Algorithm 1 to be

$$k_b = \min_{1\leq j\leq p} \frac{Y_{jb}^* - \bar{Y}_{jb}^*}{\sqrt{S_{jj,b}^*}},$$

and then we take the estimated prediction factor k to be the α-quantile of $k_1, k_2 \ldots, k_B$. The prediction region in the $\boldsymbol{Y}$-scale is given by

$$\left[\bar{Y}_1 + k\sqrt{S_{11,y}}, \infty\right) \times \cdots \times \left[\bar{Y}_p + k\sqrt{S_{pp,y}}, \infty\right).$$

If $\hat{\lambda}_j > 0$ for all $j = 1, 2, \ldots, p$, then in the $\boldsymbol{X}$-scale the $(1-\alpha)$-one-sided lower prediction region is given by

$$\left[\left(\bar{Y}_1 + k\sqrt{S_{11,y}}\right)^{1/\hat{\lambda}_1}, \infty\right) \times \cdots \times \left[\left(\bar{Y}_p + k\sqrt{S_{pp,y}}\right)^{1/\hat{\lambda}_p}, \infty\right),$$

and we deal with a zero or negative $\hat{\lambda}_j$ analogously.

3. Nonparametric prediction regions using kernel density estimation

We shall now explore an alternative approach to construct nonparametric rectangular prediction regions. The approach consists of obtaining a kernel density estimate of the unknown probability density function, and then use the probability integral transform based on the estimated density function to derive a rectangular prediction region. Such an approach provides us with a justification for using common prediction limits for each marginal component in the transformed scale, quite analogous to the use of a common prediction factor κ in the transformed scale in Section 2. An inverse transformation can then be used to obtain the required prediction limits in the original scale. We shall now present the details.

3.1. One-sided upper and lower prediction regions

We shall first present our approach for computing upper prediction limits; the case of lower prediction limits can be handled similarly. The case of two-sided prediction regions will be explained later. Suppose we want to compute upper one-sided prediction limits for the components of the random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$, using the sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ from the distribution of $\boldsymbol{X}$. Thus we have to use the data to estimate $\gamma_1, \gamma_2, \ldots, \gamma_p$ that satisfy

$$P\left(X_1 \leq \gamma_1, X_2 \leq \gamma_2, \ldots, X_p \leq \gamma_p\right) = 1 - \alpha. \tag{12}$$

If we can find one-to-one transformations $Y_j = g_j\left(X_j\right), j = 1, 2, \ldots, p,$ so that $Y_1, Y_2, \ldots, Y_p$ are identically distributed random variables, then it makes sense to have a common upper limit ζ that satisfies

$$P\left(Y_1 \leq \zeta, Y_2 \leq \zeta, \ldots, Y_p \leq \zeta\right) = P\left(\max_{1 \leq j \leq p} Y_j \leq \zeta\right) = 1 - \alpha. \tag{13}$$

If the distribution functions $F_j\left(x\right)$ of all the X_js, $j = 1, 2, \ldots, p$, were completely known, then an obvious transformation that one can use is $Y_j = g_j\left(X_j\right) = F_j\left(X_j\right)$, $j = 1, 2, \ldots, p$. Clearly, the transformed random variables $F_j\left(X_j\right), j = 1, 2, \ldots, p$, are identically distributed as $U\left(0, 1\right)$ random variables. However, since the $F_j\left(X_j\right), j = 1, 2, \ldots, p$ are unknown, the idea is to estimate them marginally using *kernel density estimation* (KDE). Call the estimated distribution functions $\hat{F}_j, j = 1, 2, \ldots, p$, and let z be an estimate of ζ satisfying (13), where $Y_j = \hat{F}_j$, $j = 1, 2, \ldots, p$. We can now obtain estimates of the upper limits γ_j, $j = 1, 2, \ldots, p$, satisfying (12) as $\hat{\gamma}_j = c_j = \hat{F}_j^{-1}\left(z\right), j = 1, 2, \ldots, p$. Obviously, since the upper limits so computed are estimates obtained from the data, we do not expect (12) to hold exactly. We shall explore this shortly based on numerical results.

Kernel density estimation is a nonparametric statistical method used to estimate an unknown PDF or CDF. Whenever we have a random sample $X_1, X_2, \ldots, X_n$ from a continuous univariate distribution with an unknown density function $f\left(x\right)$, the kernel density estimate of $f\left(x\right)$, say $\hat{f}\left(x\right)$, is given by:

$$\hat{f}\left(x\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

where h denotes the bandwidth, which can be thought of as a smoothing parameter. The *kernel function* $K\left(\cdot\right)$ satisfies $K\left(\cdot\right) \geq 0$ and

$$\int\limits_{-\infty}^{\infty} K\left(t\right) dt = 1.$$

In this study, we shall use the Gaussian kernel $K\left(t\right) = \phi\left(t\right)$, where $\phi\left(\cdot\right)$ is the standard normal density function. Our interest is in estimating the CDF, say $F\left(\cdot\right)$. The corresponding estimate, say $\hat{F}(t)$, is given by:

$$\begin{aligned}\hat{F}(t) &= \int_{-\infty}^{t} \hat{f}(u)\,du = \int_{-\infty}^{t} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{u-X_i}{h}\right)du\\ &= \int_{-\infty}^{t} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\phi\left(\frac{u-X_i}{h}\right)du = \frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{t-X_i}{h}\right), \end{aligned} \tag{14}$$

where $\Phi(\cdot)$ denotes the standard normal CDF. Moreover, our choice for the bandwidth h will be *Silverman's Rule of Thumb*, given by

$$h = 0.9\min(S, IQR/1.34)\,n^{-1/5}. \tag{15}$$

This bandwidth appears to be the preferred choice whenever $K(\cdot)$ is chosen to be the Gaussian kernel (Silverman, 1986).

We point out that since $\hat{F}_j$ is a one-to-one function, $\hat{F}_j^{-1}(z)$ always exists and hence the resulting region based on KDE always has defined limits, unlike the case for Box-Cox transformation. Algorithm 2 gives the steps necessary to compute nonparametric one-sided upper and lower prediction limits using KDE. In order to understand Step 5 in the algorithm, we recall that when we have a random sample of univariate observations $X_1, X_2, \ldots, X_n$, and we want to construct a $100(1-\alpha)\%$ nonparametric upper prediction limit for a future observation, the upper prediction limit is given by the rth order statistic $X_{(r)}$, where $r = \lceil (1-\alpha)(n+1) \rceil$. Similarly, to construct a $100(1-\alpha)\%$ lower prediction limit for a future observation, the required limit is given by $X_{(r)}$, where $r = \lfloor \alpha(n+1) \rfloor$ (Meeker et al., 2017).

3.2. Two-sided prediction regions

We shall now develop the methodology to compute nonparametric two-sided prediction regions using kernel density estimation. In constrast to (12), we need to estimate $\gamma_{11}, \gamma_{21}, \ldots, \gamma_{p1}$ and $\gamma_{12}, \gamma_{22}, \ldots, \gamma_{p2}$ that satisfy the condition:

$$P(\gamma_{11} \le X_1 \le \gamma_{12}, \gamma_{21} \le X_2 \le \gamma_{22}, \ldots, \gamma_{p1} \le X_p \le \gamma_{p2}) = 1-\alpha. \tag{16}$$

Similar to the development of one-sided prediction regions, if the distribution functions $F_j(x)$, $j = 1, 2, \ldots, p$, were completely known, then we can use the transformation $Y_j = F_j(X_j)$ and find the common upper and lower prediction limits ζ_1 and ζ_2 satisfying

$$P(\zeta_1 \le Y_j \le \zeta_2, j = 1, 2, \ldots, p) = 1-\alpha. \tag{17}$$

Since Y_j, $j = 1, 2, \ldots, p$ are all $U(0,1)$ random variables, and since the $U(0,1)$ distribution is symmetric, it makes sense to set $\zeta_1 = 1-\zeta_2$. Thus, write (17) as

$$\begin{aligned}P(1-\zeta_2 \le Y_j \le \zeta_2, j = 1, 2, \ldots, p) &= P(\max\{Y_j, 1-Y_j\} \le \zeta_2, j = 1, 2, \ldots, p) \quad (18)\\ &= P\left(\max_{1\le j\le p}\max\{Y_j, 1-Y_j\} \le \zeta_2\right) = 1-\alpha. \quad (19)\end{aligned}$$

As in the one-sided case, since $F_j(x), j = 1, 2, \ldots, p$, are unknown, we estimate them via KDE. Let $\hat{F}_j$, $j = 1, 2, \ldots, p$, be the estimated CDFs and let z be the estimate of ζ_2 that

Algorithm 2 Nonparametric one-sided upper (lower) prediction regions based on KDE

1. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be the random sample, where each $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $i = 1, 2, \ldots, n$ is a $(p \times 1)$ column vector of measurements from the ith subject.
2. For each $j = 1, 2, \ldots, p$, estimate the distribution function of the jth component using KDE (see (14)).The data used to estimate F_j are $X_{1j}, X_{2j}, \ldots, X_{nj}$. Call the estimated CDF $\hat{F}_j$.
3. Compute $Y_{ij} = \hat{F}_j\left(X_{ij}\right)$ for each X_{ij}, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, p$.
4. Compute $z_i = \max\limits_{1 \leq j \leq p} Y_{ij} \left(z_i = \min\limits_{1 \leq j \leq p} Y_{ij}\right)$, for each $i = 1, 2, \ldots, n$.
5. Compute the nonparametric $(1-\alpha)$ upper (lower) prediction limit of the $z_1, z_2, \ldots, z_n$. Denote this upper (lower) limit by z; thus $z = z_{(r)}$, where $r = \lceil (1-\alpha)(n+1) \rceil$ $(r = \lfloor \alpha (n+1) \rfloor)$.
6. Now compute for $c_j = \hat{F}_j^{-1}(z)$, $j = 1, 2, \ldots, p$.
7. The $(1-\alpha)$-nonparametric upper (lower) prediction region is given by

$$(-\infty, c_1] \times (-\infty, c_2] \times \cdots \times (-\infty, c_p] \qquad ([c_1, \infty) \times [c_2, \infty) \times \cdots \times [c_p, \infty)).$$

satisfies (19). We can estimate the prediction limits in (16) as $\hat{\gamma}_{j1} = \hat{F}_j^{-1}(1-z)$ and $\hat{\gamma}_{j2} = \hat{F}_j^{-1}(z)$, $j = 1, 2, \ldots, p$. Algorithm 3 gives the procedure to compute the nonparametric two-sided prediction region using KDE. Step 6 of Algorithm 3 is motivated by the fact that ζ_2 has been expressed in (19) as the $(1-\alpha)$-quantile of the random variable $\max\limits_{1 \leq j \leq p} \max\{Y_j, 1 - Y_j\}$.

Here we would like to make an important remark concerning Step 6 in Algorithm 2. The computation of the order statistic-based nonparametric upper prediction limit in Step 6 requires the independence of $z_1, z_2, \ldots, z_n$. However, these quantities are not independent since the $\hat{F}_j$s are not independent. In formulating the algorithm, we have simply ignored this. The estimated coverage probabilities that we shall shortly report will indicate the effect of ignoring the lack of independence among $z_1, z_2, \ldots, z_n$.

3.3. Mixed-sided nonparametric prediction regions using kernel density estimation

In many applications, we are interested in prediction regions that are a combination of one-sided and two-sided intervals, since some variables may require two-sided reference limits while others are appropriately bounded by one-sided reference limits. We shall refer to such regions as *mixed-sided prediction regions*. For example, we may be interested in finding the region $[c_1, d_1] \times (-\infty, d_2]$ such that

$$P\left(c_1 \leq X_1 \leq d_1, X_2 \leq d_2\right) = 1 - \alpha.$$

Algorithm 3 Nonparametric two-sided prediction regions based on KDE

1. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be the random sample, where each $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $i = 1, 2, \ldots, n$, is a $(p \times 1)$ column vector of measurements from the ith subject.
2. For each $j = 1, 2, \ldots, p$, estimate the distribution function of the jth component using KDE (see (14)).The data used to estimate F_j are $X_{1j}, X_{2j}, \ldots, X_{nj}$. Call the estimated CDF $\hat{F}_j$.
3. Compute $Y_{ij} = \hat{F}_j\left(X_{ij}\right)$ for each X_{ij}, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, p$.
4. Compute $U_{ij} = \max\left\{Y_{ij}, 1 - Y_{ij}\right\}$ for each Y_{ij}, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, p$.
5. Compute $z_i = \max\limits_{1 \leq j \leq p} U_{ij}$, for each $i = 1, 2, \ldots, n$.
6. Compute the nonparametric $(1 - \alpha)$ upper prediction limit of the $z_1, z_2, \ldots, z_n$. Denote this upper limit by z; thus $z = z_{(r)}$, where $r = \lceil (1 - \alpha)(n + 1) \rceil$.
7. Compute $c_j = \hat{F}_j^{-1}(1 - z)$ and $d_j = \hat{F}_j^{-1}(z)$, $j = 1, 2, \ldots, p$.
8. The $(1 - \alpha)$-nonparametric two-sided prediction region is given by

$$[c_1, d_1] \times [c_2, d_2] \times \cdots \times [c_p, d_p].$$

We now take up the problem of computing mixed-sided nonparametric prediction regions. Suppose our data consists of the random sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, where each $\boldsymbol{X}_i$ is p-variate. Moreover, let $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ be the observation that we wish to predict and assume that it has the same distribution as the $\boldsymbol{X}_i$s and is independent of them. Without loss of generality, we develop a procedure to compute two-sided prediction limits for the first p_1 components of $\boldsymbol{X}$ and upper prediction limits for the remaining $p - p_1$ components. In doing so, we use a KDE-based approach since this approach generally shows superior performance over the Box-Cox transformation-based approach, as we have seen in Section 4.

Let $F_j(\cdot)$ be the CDF of X_j, $j = 1, 2, \ldots, p$. If we can find scalar quantities u, u', and v, all three being functions of $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, that satisfy

$$\begin{aligned} P_{\boldsymbol{X}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n} \Big(& F_j^{-1}(u') \leq X_j \leq F_j^{-1}(u), \forall j = 1, \ldots, p_1 \\ & \text{and } X_j \leq F_j^{-1}(v), \forall j = p_1 + 1, \ldots, p \Big) = 1 - \alpha, \end{aligned} \tag{20}$$

then the region

$$\left[F_1^{-1}(u'), F_1^{-1}(u)\right] \times \cdots \times \left[F_{p_1}^{-1}(u'), F_{p_1}^{-1}(u)\right] \times \left(-\infty, F_{p_1+1}^{-1}(v)\right] \times \cdots \times \left(-\infty, F_p^{-1}(v)\right] \tag{21}$$

is a $(1 - \alpha)$-mixed-sided nonparametric prediction region for $\boldsymbol{X}$. The condition in (20) is equivalent to

$$\begin{aligned} P_{\boldsymbol{X}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n} (& u' \leq F_j(X_j) \leq u, \forall j = 1, \ldots, p_1 \\ & \text{and } F_j(X_j) \leq v, \forall j = p_1 + 1, \ldots, p) = 1 - \alpha. \end{aligned} \tag{22}$$

Since each $F_j(X_j)$ follows a $U(0,1)$ distribution, we can choose $u' = 1 - u$. Furthermore, since infinitely many possible values of u and v can satisfy (22), we shall impose a constraint on u and v so as to arrive at a unique solution. The constraint to be imposed is that the marginal probabilities in (22) should be equal. This amounts to choosing u and v such that $v = 2u - 1$. We can see this by observing that if U and V are $U(0,1)$ random variables, then imposing the condition $P(1 - u \leq U \leq u) = P(V \leq v)$ implies $v = 2u - 1$. Substituting these expressions for u' and v, (22) becomes

$$\begin{aligned} P_{\boldsymbol{X},\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n} \big(1 - u &\leq F_j(X_j) \leq u, \forall j = 1, 2, \ldots, p_1 \\ &\text{and } F_j(X_j) \leq 2u - 1, \forall j = p_1 + 1, \ldots, p\big) = 1 - \alpha. \end{aligned} \tag{23}$$

Since

$$\begin{aligned} &1 - u \leq F_j(X_j) \leq u, \forall j = 1, 2, \ldots, p_1 \\ &\iff \max\left\{\max_{1\leq j\leq p_1} F_j(X_j), \max_{1\leq j\leq p_1} (1 - F_j(X_j))\right\} \leq u, \end{aligned} \tag{24}$$

$$F_j(X_j) \leq 2u - 1, \forall j = p_1 + 1, \ldots, p \iff \max_{p_1+1\leq j\leq p} \left\{\frac{1 + F_j(X_j)}{2}\right\} \leq u \tag{25}$$

then we can write (23) as

$$P_{\boldsymbol{X},\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n} \left(\max\left\{h_1(\boldsymbol{X}), h_2(\boldsymbol{X}), h_3(\boldsymbol{X})\right\} \leq u\right) = 1 - \alpha, \tag{26}$$

where

$$\begin{aligned} h_1(\boldsymbol{X}) &= \max_{1\leq j\leq p_1} F_j(X_j) \\ h_2(\boldsymbol{X}) &= \max_{1\leq j\leq p_1} (1 - F_j(X_j)) \\ h_3(\boldsymbol{X}) &= \max_{p_1+1\leq j\leq p} \left\{\frac{1 + F_j(X_j)}{2}\right\}. \end{aligned}$$

From (26) we can conclude that u is a $(1 - \alpha)$-upper prediction limit of

$$\max\left\{h_1(\boldsymbol{X}), h_2(\boldsymbol{X}), h_3(\boldsymbol{X})\right\}.$$

Since the distribution functions $F_j(\cdot)$ are unknown, we estimate them using KDE. Algorithm 4 gives the steps to compute the mixed-sided nonparametric prediction region using KDE.

3.4. KDE with logarithmic transformation

Studies such as Geenens and Wang (2016) and Jones et al. (2018) suggest that whenever the density is supported on the set of positive real numbers, we should first apply a logarithmic transformation on the observations before estimating the density function. Geenens and Wang (2016) argue that the KDE approach to estimate the density of a positive random variable is inadequate due to the boundary bias problem and the fact that such a density might have a long right tail. Charpentier and Flachaire (2014) also mention that

Algorithm 4 Mixed-sided nonparametric prediction regions based on KDE

1. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be the random sample, where each $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $i = 1, 2, \ldots, n$ is a $(p \times 1)$ column vector of measurements from the ith subject.
2. For each $j = 1, 2, \ldots, p$, estimate $\hat{F}_j$ via KDE using $X_{1j}, X_{2j}, \ldots, X_{nj}$.
3. Compute $Y_{ij} = \hat{F}_j\left(X_{ij}\right)$ for all $j = 1, 2, \ldots, p$, and $i = 1, 2, \ldots, n$.
4. Compute $u_i = \max\left\{\max\limits_{1\leq j\leq p_1} Y_{ij}, \max\limits_{1\leq j\leq p_1}\left(1 - Y_{ij}\right), \max\limits_{p_1+1\leq j\leq p}\left(\frac{1+Y_{ij}}{2}\right)\right\}$, for each $i = 1, 2, \ldots, n$.
5. Compute the nonparametric $(1-\alpha)$-upper prediction limit of the $u_1, u_2, \ldots, u_n$. Denote this upper limit by u; thus $u = u_{(r)}$, where $r = \lceil (1-\alpha)(n+1) \rceil$.
6. Compute $c_j = \hat{F}_j^{-1}(1-u)$ and $d_j = \hat{F}_j^{-1}(u)$, $j = 1, 2, \ldots, p_1$; and $d_j = \hat{F}_j^{-1}(2u-1)$, $j = p_1 + 1, \ldots, p$.
7. The $(1-\alpha)$-mixed-sided nonparametric prediction region is given by $[c_1, d_1] \times \cdots \times [c_{p_1}, d_{p_1}] \times (-\infty, d_{p_1+1}] \times \cdots \times (-\infty, d_p]$.

doing a preliminary logarithmic transformation before applying KDE can provide a better fit for heavy-tailed densities.

To apply this idea to the proposed KDE-based procedure, we can modify Algorithm 2 (or Algorithm 3 for the two-sided case) by taking the logarithm component-wise of each $\boldsymbol{X}_i$, $i = 1, 2, \ldots, n$ in Step 1 before proceeding to the other steps, and then exponentiating each limit in Step 6 of Algorithm 2 (or Step 7 Algorithm 3) to get the reference limits in the original scale. We shall refer to this procedure as the *KDE with log transform* procedure.

4. Numerical results and an example

In order to evaluate the performance of the proposed procedures to construct prediction regions in the nonparametric case, simulations will be carried out to estimate coverage probabilities and expected volumes for data generated from a multivariate lognormal distribution with mean vector in the logarithmic scale of $\mathbf{0}$, and covariance matrix in the logarithmic scale $\boldsymbol{\Sigma} = (1-\rho)\boldsymbol{I}_p + \rho\mathbf{1}_p\mathbf{1}_p'$ where $\rho = 0.5$, and $\mathbf{1}_p$ is the $(p \times 1)$ column vector of 1s and $\boldsymbol{I}_p$ is the $(p \times p)$ identity matrix. We use the R package **compositions** of van den Boogaart and Tolosana-Delgado (2008) to generate samples from the multivariate lognormal distribution. We examine the performance for sample sizes $n =$ 50, 100, and 200 and refer to these as small, moderate and large sample sizes. We also use dimensions $p = 2$ and 3 since most applications of MRRs involve only at most three analytes. For the Box-Cox transformation-based procedure, we use $B = 500$ bootstrap samples to estimate the prediction factor. For the KDE-based procedure, the inverse function in Step 6 of Algorithm 2 and Step 7 in Algorithm 3 and all other occurrences of the inverse function in this study are computed using the R package **GoFKernel** of Pavia (2015). The coverage probabilities are based on 5000 simulated samples, and the results are given in Table 1.

From the numerical results in Table 1, we can see that the coverage probabilities of the proposed methodologies are generally close to the nominal level of 0.95, even for a sample of size $n = 50$. Furthermore, it seems that the KDE-based procedure is slightly more accurate than the Box-Cox transformation-based procedure. It is worth comparing our sample sizes with those of Young and Mathew (2020), who also propose nonparametric reference regions. Young and Mathew (2020) examine the performance of their procedure only for sample sizes 300 and 1000. We note that for these dimensions, the coverage probabilities for $n = 100$ in Table 1 are already comparable to Young and Mathew's results for $n = 300$.

Table 2 gives the expected volumes obtained from the proposed methodologies. Table 2 shows that for both the Box-Cox transformation-based procedure and the KDE-based procedure, the expected volume decreases with the sample size. We can see that the KDE-based prediction regions have smaller expected volumes than the Box-Cox transformation-based prediction regions. This implies that the KDE-based procedure results in better precision in estimating the prediction region. On the basis of the results in Tables 1 and 2, the KDE-based procedure has better overall performance than the Box-Cox transformation-based procedure. We note that in computing the expected volume for the Box-Cox transformation-based procedure, we replaced any negative lower limit with zero except when $\hat{\lambda}_j = 0$, in which case a negative lower limit is kept negative. In Table 3 we present the results of the proposed KDE-based one-sided lower and upper prediction regions. The results show accurate coverage, even for small sample sizes.

Table 1: Estimated coverage probabilities of the nonparametric rectangular prediction regions based on Box-Cox transformation and KDE for nominal level = 0.95

	Box-Cox		KDE	
	$p = 2$	$p = 3$	$p = 2$	$p = 3$
$n = 50$	0.9344	0.9396	0.9582	0.9414
$n = 100$	0.9408	0.9396	0.9472	0.9428
$n = 200$	0.9480	0.9398	0.9428	0.9460

Table 2: Expected volumes of the nonparametric two-sided prediction regions based on Box-Cox transformation and KDE for nominal level = 0.95

	Box-Cox		KDE	
	$p = 2$	$p = 3$	$p = 2$	$p = 3$
$n = 50$	116.06	2,476.42	95.41	1061.11
$n = 100$	96.67	1,582.25	60.11	823.90
$n = 200$	89.76	1,342.86	53.02	660.93

Table 3: Estimated coverage probabilities of the nonparametric one-sided lower and upper prediction regions based on KDE for nominal level = 0.95

	Lower		Upper	
	$p = 2$	$p = 3$	$p = 2$	$p = 3$
$n = 50$	0.9588	0.9594	0.9518	0.9512
$n = 100$	0.9430	0.9436	0.9526	0.9526
$n = 200$	0.9476	0.9452	0.9464	0.9488

4.1. Comparison of nonparametric procedures when sampling from a highly skewed distribution

We now compare the performances of the KDE-based procedures (both with and without a preliminary log transformation) and the Box-Cox transformation approach to compute prediction regions when we sample from a highly skewed distribution. In the simulations, we generate the data from a gamma distribution with density function given in (27)

$$f(x) = \frac{1}{\lambda^{\eta}\Gamma(\eta)} x^{\eta-1} e^{-x/\lambda}, \quad x \geq 0, \tag{27}$$

with shape parameter $\eta = 0.04$ and scale parameter $\lambda = 1$. This distribution has skewness $2/\sqrt{\eta} = 10$. Table 4 shows the estimated coverage probabilities. We can see that Box-Cox transformation-based procedure results in estimated coverage probabilities very close to 0.95. On the other hand, the usual KDE procedure on the original data is too conservative. While in the previous results, we have seen that the KDE-based procedure outperforms the Box-Cox transformation-based procedure, Table 4 suggests that the Box-Cox-based procedure is more robust to highly skewed distributions, and the KDE-based procedure breaks down under such extreme skewness. Nonetheless, the KDE with log transform procedure rectifies the coverage.

Table 4: Estimated coverage probabilities of the Box-Cox transformation-based and the KDE-based two-sided prediction regions under highly skewed distributions

	Box-Cox transformation	KDE	KDE with log transform
$n = 50$	0.9424	0.9934	0.9668
$n = 100$	0.9500	0.9898	0.9514
$n = 200$	0.9490	0.9936	0.9564

4.2. An example: assessment of liver function

To apply the proposed procedure to compute nonparametric rectangular prediction regions, we use the liver function data from Appendix 4.2 of Harris and Boyd (1995). The measurements are from single blood specimens taken from 596 male medical students during the years 1987-1991 at the University of Virginia. Among the measurements taken from each subject are two liver enzymes: alanine transaminase (ALT) in U/L and aspartate transaminase (AST) in U/L. After the removal of three outliers, the summary statistics are given in Table 5.

Table 5: Summary statistics for measurements on ALT and AST taken from the liver function data of Harris and Boyd (1995)

Analyte	Mean	Median	S.D.	Skewness
ALT	26.97	23.00	17.83	3.63
AST	23.66	22.00	9.51	1.80

Figure 1 shows the density plots for these two analytes. Clearly, both analytes are skewed to the right. Table 5 above also shows that the sample coefficient of skewness is positive. Thus, we use our proposed procedures to compute nonparametric prediction regions. The resulting MRRs using both the Box-Cox transformation and KDE-based approaches are given in Table 6. According to Mayo Clinic (2020), the normal levels for ALT and AST are, respectively, 7-55 and 8-48. Therefore, while the lower limits of the MRR for our proposed procedures agree closely with the lower limits of the reference intervals used in practice, the upper limits are quite different. We hasten to say that these enzymes can be erratically large, in some conditions they can be in the 1000s range (eMedicine Health, 2020). Figure 1 also shows that there are several outlying measurements for ALT, and this could be a factor leading to the unexpectedly high upper reference limit for ALT.

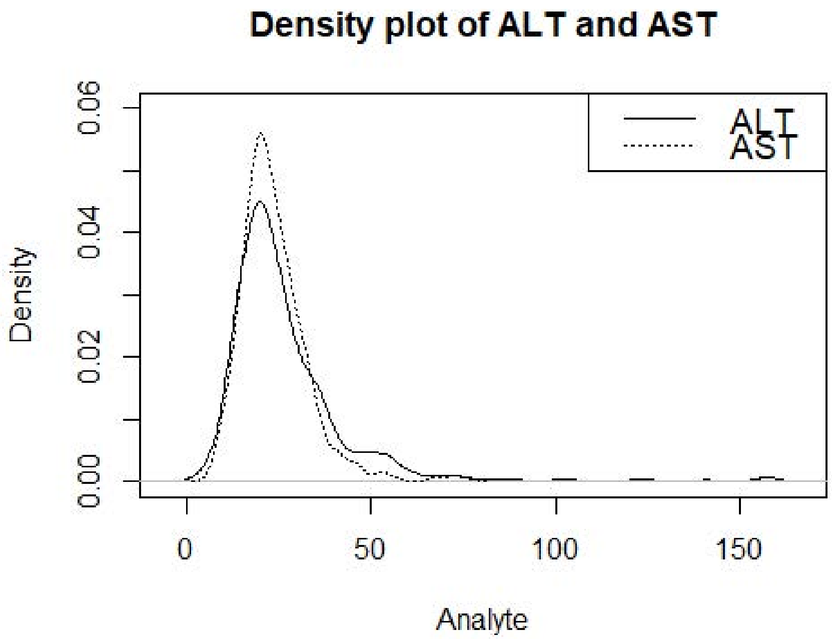

Figure 1: Density plot of ALT and AST

Table 6: MRR for liver function data computed as a two-sided prediction region using Box-Cox transformation and KDE

Analyte	Box-Cox	KDE
ALT	7.2-84.8	7.1-79.1
AST	9.0-54.8	8.8-52.0

4.3. Numerical results on mixed-sided nonparametric prediction regions

We shall now evaluate the performance of our proposed procedure to compute mixed-sided nonparametric prediction regions using KDE, described in Section 3. We generate data from the same distribution used in previous subsections. That is, we estimate coverage probabilities for data generated from a multivariate lognormal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma} = (1-\rho)\,\boldsymbol{I}_p + \rho\mathbf{1}_p\mathbf{1}_p'$, where $\rho = 0.5$ on the logarithmic scale. The 95% prediction regions will be computed based on 5000 simulated samples. Moreover, we use sample sizes n = 50, 100, 200, and $(p, p_1) = (2, 1)$ and $(3, 2)$. Table 7 shows the results. It appears that a sample of size $n = 50$ is sufficient for the proposed methodology to yield accurate results.

Table 7: Estimated coverage probabilities of the mixed-sided nonparametric prediction regions based on KDE for nominal level = 0.95

	$p = 2, p_1 = 1$	$p = 3, p_1 = 2$
$n = 50$	0.9548	0.9448
$n = 100$	0.9510	0.9450
$n = 200$	0.9510	0.9502

5. Discussion

The problem of constructing multivariate reference regions has received proper attention in the literature only recently, except the computation of traditional ellipsoidal prediction regions under the multivariate normality assumption. There are two difficulties associated with the latter region; first, the multivariate normality assumption is not always valid and second, ellipsoidal regions are not appropriate for deciding which among several analytes are outside the normal range. The nonparametric rectangular regions that we have constructed address both of these issues satisfactorily. A different construction of nonparametric rectangular prediction regions is described in Young and Mathew (2020); however, the resulting region exhibits satisfactory coverage probabilities only under relatively large sample sizes. While our work is focused on the computation of rectangular prediction regions only, an important issue is whether a prediction region is appropriate for the purpose for which a reference region is to be used. Some of the recent literature has emphasized tolerance regions, and rectangular tolerance regions are indeed available in the parametric setup of multivariate normality, and in a nonparametric scenario; see Lucagbo and Mathew (2023) and Young and Mathew (2020). Here we do want to note that some laboratory medicine experts have pointed out the role of prediction intervals and regions; see Horn and Pesce (2005), National Committee for Clinical Laboratory Standards (2010), and Trost (2006). In particular, while discussing ellipsoidal regions, Trost (2006, p. 38) notes that "Reference intervals referred to in this document are arguably the closest to prediction intervals since we want exactly 95% of the future observations from reference individuals to fall inside the bounds". We shall not further consider the issue of what criterion is appropriate for the construction of a reference region; this clearly requires input from experts in laboratory medicine.

In our work we have employed two approaches for computing a nonparametric rectangular prediction regions: using the Box-Cox transformation and using kernel density estimation. Estimated coverage probabilities lead us to the conclusion that both approaches are satisfactory. Based on estimated coverage probabilities and expected volumes, our overall recommendation is the solution based on kernel density estimates. A problem of considerable interest in the context of reference regions is the computation of such regions that are covariate dependent, perhaps using a multivariate regression model. We hope to address this problem in the near future.

Acknowledgements

The authors are grateful to a reviewer for several constructive suggestions that resulted in clarification of some of the ideas and better organization of the results.

References

Albert, A., and Harris, E. K. (1987). *Multivariate Interpretation of Clinical Lab- oratory Data*. Marcel Dekker, Inc., New York.

Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society, Series B*, **26**, 211-252.

Charpentier, A. and Flachaire, E. (2015). Log-transform kernel density estimation of income distribution. (AMSE Working Paper 2015 No. 06). Marseille, France: Aix-Marseille School of Economics.

eMedicineHealth. (2020). Retrieved from: https://www.emedicinehealth.com/liver_blood_tests/article_em.htm#what_should_i_know_about_liver_blood_tests_why_are_they_used. Retrieved April 11, 2020.

Geenens, G. and Wang, C. (2016). Local-likelihood transformation kernel density estimation for positive. arXiv:1602.04862.

Harris, E. K. (1981). Statistical Aspects of Reference Values in Clinical Pathology. In *Progress in Clinical Pathology*, Vol. 8, Mario Stefanini and Ellis S. Benson, eds. Grune & Stratton, New York, pp. 45-66.

Harris, E. K. and Boyd, J. C. (1995). *Statistical Bases of Reference Values in Laboratory Medicine*. Marcel Dekker, Inc., New York.

Horn, P. S. and Pesce, A. J. (2005). *Reference Intervals: A Users Guide*. American Association for Clinical Chemistry Press, Washington, D. C.

Ichihara, K. and Boyd, J. C. (2010). An appraisal of statistical procedures used in derivation of reference intervals. *Clinical Chemistry and Laboratory Medicine*, **48**, 1537-1551.

Ichihara, K. and Kawai, T. (1996). Determination of reference intervals for 13 plasma proteins based on IFCC international reference preparation (CRM470) and NCCLS proposed guideline (C28- P,1992): trial to select reference individuals by results of screening tests and application of maximal likelihood method. *Journal of Clinical Laboratory Analysis*, **10**, 110-117.

Jones, A. T., Nguyen, H. D., and McLachlan, G. J. (2018). Positive data kernel density estimation via the logKDE package for R. arXiv:1804.08365v2 [stat.CO].

Liu, W., Bretz, F., and Cortina-Borja, M. (2021). Reference range: Which statistical intervals to use? *Statistical Methods in Medical Research*, **30**, 523-534.

Lucagbo, M. D., Mathew, T., and Young, D. S. (2023). Rectangular multivariate normal prediction regions for setting reference regions in laboratory medicine, *Journal of Biopharmaceutical Statistics*, **33**, 191-209.

Mayo Clinic. (2020). Retrieved from: https://www.mayoclinic.org/tests-procedures/liver-function-tests/about/pac-20394595. Retrieved April 11, 2020.

Meeker, W. Q., Hahn, G. J., and and Escobar, L. A. (2017). *Statistical Intervals: A Guide for Practitioners and Researchers*, Second Edition. John Wiley & Sons.

National Committee for Clinical Laboratory Standards. (2010). *EP28-A3C: Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline*, - Third Edition. Clinical and Laboratory Standards Institute, Wayne, PA.

Pavia, J. M. (2015). Testing goodness-of-fit with the kernel density estimator: GoFKernel. *Journal of Statistical Software*, Code Snippets, **66**, 1-27.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

Solberg, H. E. (1987). Approved recommendation on the theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits. *Journal of Clinical Chemistry and Clinical Biochemistry*, **25**, 645-656.

Strike, P. W. (1991). *Statistical Methods in Laboratory Medicine*, Butterworth- Heinemann, Oxford, UK.

Trost, D. C. (2006). Multivariate probability-based detection of drug-induced hepatic signals. *Toxicological Reviews*, **25**, 37-54.

van den Boogaart, K. G. and Tolosana-Delgado, R. (2008). compositions: A unified R package to analyze compositional data. *Computers and Geosciences*, **34**, 320-338.

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

Winkel, P., Gaede, P., and Lyngbye, J. (1976). Method for monitoring plasma progesterone concentrations in pregnancy. *Clinical Chemistry*, **22**, 422-428.

Young, D. S. and Mathew, T. (2020). Nonparametric hyperrectangular tolerance and prediction regions for setting multivariate reference regions in laboratory medicine. *Statistical Methods in Medical Research*, **29**, 3569-3585.