Abstract—A statistical quality control (QC) process for raw materials and/or the final product of traditional Chinese medicine (TCM) is proposed based on a tolerance region. We construct a tolerance region for a multivariate random effects model to assess the quality control of TCM products from different regions and different batches. The product is then declared to pass the QC only if the constructed region is within the permitted range. An example concerning the development of a TCM for treating patients with rheumatoid arthritis is used to illustrate the proposed approach based on the tolerance region for QC.

Keywords—Tolerance Region; Random effects model.

I. INTRODUCTION

Westernization of traditional Chinese medicine (TCM) has attracted much attention. TCM is usually a complex compound, whereas Western medicines typically contain a single active ingredient. TCM usually contains a combination of individually processed botanical substances. The balance and interaction of all the ingredients is considered more important than the effect of each individual ingredient. Appropriate formula prescriptions based on accurate differential diagnoses can be made only by qualified TCM physicians. In practice, the raw materials of TCM are often from different locations, and final products may be manufactured at different sites. As a result, product variability from different sources is inevitable. Thus, how to ensure consistency in raw materials, the processing of materials and the final product is a quality control concern for TCM.

In some quality control applications, pharmaceutical products are subject to content-uniformity testing. For a single component, Tse et al. [12] proposed testing consistency across two different study sites for uniformity. A new parameter, the consistency index, was defined in [12] and the large sample theory of the estimator is derived for constructing a 95% confidence interval used to assess the quality control of the materials. The TCM is regarded as passing the QC if the constructed 95% confidence lower limit is greater than a pre-specified QC limit; otherwise it should be rejected. Lu et al. [9] extended the results in [12] to the case of two correlative components. The consistency index of the two components case was defined for each pair of two sites as the minimum of individual consistency indices of active components. The performance of this measure when the number of active components is relatively large may lead to a rather conservative measure.

In this paper, a multivariate random effects model for multiple active components is adopted to describe the random effects from sources, such as locations and batches, and a tolerance region of the multiple components is given for assessment of quality control. In the current study, the approach in [5] for constructing a tolerance region of a normally distributed random vector is applied to establish a tolerance region of a random effects model with batch and regional effects for the QC of TCM.

Instead of assessing the consistency of the materials from two sites as in [9], our approach manages the quality of the materials from different locations simultaneously. If the materials do not pass our quality control via the tolerance region of the random effects model, we perform pairwise comparisons to extract the failed one in order to reduce the cost of materials.

In the next section, a multivariate random effects model is introduced and a tolerance region for the active components is constructed. In Section III, a numerical example is employed to illustrate the construction of the tolerance region A short summary is given is Section IV.

II. TOLERANCE REGION OF A MULTIVARIATE RANDOM EFFECTS MODEL

Tolerance regions for a multivariate normal population have been well investigated. Wald [13] proposed a general parametric method, applicable to any given density function, to construct tolerance regions for large samples. However, the adequacy of the method for small samples is unknown. John [5] provided theoretical formulation for the problem of constructing tolerance region and gave a few approximate methods. Since then, many authors proposed approximate methods to construct tolerance regions for multivariate normal
distributions, see for instance, [1], [2]-[4], [6], [8], [11], among others. The result in [5] is quite significant because he not only developed the theoretical framework for the problem but also provided a simple and easy-to-use approximation for computing the tolerance factors. In this section, we first introduce the multivariate random effects model for discussing TCM with multiple active components. The approach in [5] is then adopted to derive the tolerance region of the multivariate random effects model.

Denote by p-vector \( \mathbf{Y}_{ijk} \) the measurement of the \( p \) active components from the \( k \)th sample of the \( j \)th batch from the \( i \)th region, respectively. Consider the following multivariate random effects model:

\[
\mathbf{Y}_{ijk} = \mu + \mathbf{A}_i + \mathbf{B}_j + \mathbf{e}_{ijk}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K,
\]

where the constant vector \( \mu \) is the true mean of \( \mathbf{Y}_{ijk} \) and the random vectors \( \mathbf{A}_i, \mathbf{B}_j, \) and \( \mathbf{e}_{ijk} \) are completely independent and multivariate normally distributed with mean zero and covariance matrices \( \Sigma_{A_i}, \Sigma_{B_j}, \) and \( \Sigma_{e} \), respectively. That is

\[
\mathbf{A}_i \sim N_p(0, \Sigma_{A_i}), \quad \mathbf{B}_j \sim N_p(0, \Sigma_{B_j}), \quad \mathbf{e}_{ijk} \sim N_p(0, \Sigma_{e}).
\]

According to model (1), we then obtain that \( \mathbf{Y}_{ijk} \) follows \( N_p(0, \Sigma_{A_i} + \Sigma_{B_j} + \Sigma_{e}) \) distribution.

Before we derive the tolerance region (TR) for model (1), we recall the TR proposed in [5] for a general multivariate normal distribution first. Let \( \mathbf{Y} \) be a \( N_p(\mu, \Sigma_Y) \) distributed random vector with mean vector \( \mu \) and covariance matrix \( \Sigma_Y \). Based on a random sample \( \mathbf{Y}_1, \ldots, \mathbf{Y}_N \) from \( N_p(\mu, \Sigma_Y) \), a \( \beta \)-content, \( \gamma \)-confidence tolerance region \( T(\beta, \gamma) \) of \( \mathbf{Y} \) is defined to be a subset of \( \mathbb{R}^p \) such that

\[
P_{Y_1, \ldots, Y_N} \{ \mathbf{P}_Y( \mathbf{Y} \in T(\beta, \gamma) | Y_1, \ldots, Y_N) \geq \beta \} = \gamma.
\]

Denote the sample mean and the sample covariance matrix as \( \bar{Y} \) and \( \Sigma_Y \), respectively. Moreover, the sample covariance matrix \( \hat{\Sigma}_Y \) follows a Wishart distribution with certain scale matrix \( \Lambda \) and degree of freedom \( \nu \), which will be discussed later in this section.

Based on a random sample \( \mathbf{Y}_1, \ldots, \mathbf{Y}_N \) from \( N_p(\mu, \Sigma_Y) \), [5] showed that the tolerance ellipsoid within which, with 100\(\gamma\) % confidence level, includes at least 100\(\beta\)% of the population satisfies the following inequality

\[
(Y - \bar{Y})\hat{\Sigma}_Y^{-1/2} (Y - \bar{Y}) \leq \chi^2_{\nu, \beta, p}/p,
\]

where \( \chi^2_{\nu, \beta, p}/p(2N) \) denotes the upper 100(1-\(\beta\))% point of the non-central Chi-squared distribution with degree of freedom \( p \) and non-centrality parameter \( \nu p/(2N) \), and \( \chi^2_{\nu, \beta, p}/p(2N) \) denotes the upper 100\(\gamma\)% point of Chi-squared distribution with degree of freedom \( \nu p \). We may represent the \( (\beta, \gamma) \) tolerance region \( T(\beta, \gamma) \) as follows

\[
T(\beta, \gamma) = \{ y : (y - \bar{Y})\hat{\Sigma}_Y^{-1/2} (y - \bar{Y}) \leq \chi^2_{\nu, \beta, p}/p(2N) \}.
\]

The upper bound in (3) is called the tolerance factor of the tolerance region \( T(\beta, \gamma) \).

Using model (1), to define the tolerance region \( T(\beta, \gamma) \) as (3) of the response \( \mathbf{Y} \), we have to calculate the degree of freedom of \( \hat{\Sigma}_Y \). Based on the random samples, \( \mathbf{Y}_{ijk}, i=1, \ldots, I, j=1, \ldots, J, \) and \( k=1, \ldots, K \), we calculate the following averages and sum of squares:

\[
\bar{Y} = \frac{1}{JK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \mathbf{Y}_{ijk},
\]

\[
\hat{Y}_j = \frac{1}{K} \sum_{k=1}^{K} \mathbf{Y}_{ijk}, \quad \text{for } i=1, \ldots, I,
\]

\[
\tilde{Y}_i = \frac{1}{I} \sum_{i=1}^{I} \mathbf{Y}_{ijk}, \quad \text{for } j=1, \ldots, J, \text{ and } k=1, \ldots, K,
\]

\[
SSA = JK \sum_{i=1}^{I} (\mathbf{Y}_{i\cdot}-\bar{Y})(\mathbf{Y}_{i\cdot}-\bar{Y})',
\]

\[
SSB = K \sum_{j=1}^{J} (\mathbf{Y}_{\cdot j}-\bar{Y})(\mathbf{Y}_{\cdot j}-\bar{Y})',
\]

\[
SSe = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (\mathbf{Y}_{ijk}-\tilde{Y}_i)(\mathbf{Y}_{ijk}-\tilde{Y}_i)'.
\]

The mean squares and expected mean squares are then as follows,

\[
S_A = \frac{SSA}{I-1}, \quad \Lambda_A = \text{ES}_A = JK\Sigma_A + \kappa \Sigma_B + \Sigma_e,
\]

\[
S_B = \frac{SSB}{I(J-1)}, \quad \Lambda_B = \text{ES}_B = K\Sigma_B + \Sigma_e,
\]

\[
S_e = \frac{SSe}{IJ(K-1)}, \quad \Lambda_e = \text{ES}_e = \Sigma_e.
\]

We list above results in Table 1. Moreover, according to definitions of sum of squares in (4), the distributions of \( SS_A \), \( SS_B \), and \( SS_e \) are \( W_p(\Lambda_A, 1-I) \), \( W_p(\Lambda_B, I(J-1)) \), and \( W_p(\Lambda_e, I(K-1)) \), respectively, where \( W_p(\Lambda, k) \) denotes the Wishart distribution with the scale matrix \( \Lambda \) and degree of freedom \( k \).

Based on model (1) and the normality assumption, the random vectors \( \mathbf{Y} \) and \( \bar{Y} \) follows \( N_p(\mu, \Sigma_Y) \) and \( N_p(\mu, \Sigma_{\bar{Y}}) \) distributions, respectively, with \( \Sigma_Y = \Sigma_A + \Sigma_B + \Sigma_e \), and \( \Sigma_{\bar{Y}} = \Sigma_A/I + \Sigma_B/(I) + \Sigma_e/(IJK) \). We may represent the covariance matrix \( \Sigma_Y \) as follows,
\[ \Sigma_y \equiv \frac{1}{JK} \Lambda_A + \frac{J-1}{JK} \Lambda_B + \frac{K-1}{K} \Lambda_c = c_A \Lambda_A + c_B \Lambda_B + c_c \Lambda_c, \]

Where the coefficients \( c_A, c_B, c_c \) are defined to be \( 1/(JK) \), \((J-1)/(JK) \), and \((K-1)/K \), respectively. An unbiased estimator of \( \Sigma_y \), namely \( \hat{\Sigma}_Y \), is then given by
\[
\hat{\Sigma}_Y = c_A \Sigma_A + c_B \Sigma_B + c_c \Sigma_c,
\]

(5)

Where

\[
c_A \Sigma_A \sim W_p \left( \frac{c_A \Lambda_A}{I(J-1)}, I \right) - 1)
\]
\[
c_B \Sigma_B \sim W_p \left( \frac{c_B \Lambda_B}{I(J-1)}, I \right) - 1)
\]
\[
c_c \Sigma_c \sim W_p \left( \frac{c_c \Lambda_c}{I(J-1)}, I \right) - 1)
\]

Applying the modified multivariate Satterthwaite approximation for the distribution of a linear combination of independent Wishart matrices by Nel and van der Merwe \[10\], the sample covariance matrix distribution with the degree of freedom \( JK \) can be constructed according to sample information.

> (J

Where

\[
\hat{\nu} = \frac{1}{m_i} \left[ \frac{tr(\hat{\Sigma}_Y^2)}{\nu} + \left[ \frac{tr(\hat{\Sigma}_Y)\hat{\Sigma}_Y}{\nu} \right]^2 \right]
\]

(6)

\[\sum m_i \]

where \( m_A = I-1, m_B = I(J-1), m_c = I(JK) \). When \( \Sigma_A, \Sigma_B, \) and \( \Sigma_c \) are unknown, the unknown degree of freedom \( \nu \) can be estimated by replacing \( \Lambda \) with \( S \), i.e.,
\[
\hat{\nu} = \frac{1}{m_i} \left[ \frac{tr(\hat{\Sigma}_Y^2)}{\nu} + \left[ \frac{tr(\hat{\Sigma}_Y)\hat{\Sigma}_Y}{\nu} \right]^2 \right]
\]

(7)

However, such an estimated degree of freedom is not invariant under nonsingular transformation. Krishnamoorthy and Yu \[7\] considered the Wishart approximation for the distribution of \( \Sigma_y^{-1/2} \Sigma_y \Sigma_y^{-1/2} \). Replacing \( \Sigma_e \) with \( \Sigma_y^{-1/2} \Sigma_y \Sigma_y^{-1/2} \), the resulting degree of freedom, say \( \tilde{\nu} \), now becomes
\[
\tilde{\nu} = \frac{p + p^2 - \nu}{\sum m_i \left[ \frac{tr(c_s \hat{\Sigma}_Y)}{\nu} + \left[ \frac{tr(c_s \hat{\Sigma}_Y)}{\nu} \right]^2 \right]}
\]

(8)

With this estimator of the degree of freedom, the tolerance region (3) can be constructed according to sample information.

### III. Example

To illustrate the proposed tolerance region for the statistical quality control process, we construct the proposed tolerance region for the synthetic data from \[9\]. The TCM for treating patients with rheumatoid arthritis contains two active components, namely, extract A and B. Each of these two active components has been used as herbal remedy since Ancient China. We are interested in testing the consistency of the raw materials, based on the two active components, from three different sites. Suppose that four batches from each site will be tested and nine samples from each batch will be randomly selected. Some description statistics are as follows,

\[
P = \begin{bmatrix} 4.460317 \\ 4.510282 \end{bmatrix}, \quad P_e = \begin{bmatrix} 4.598416 \\ 4.501955 \end{bmatrix}, \quad P_e = \begin{bmatrix} 4.417328 \\ 4.591696 \end{bmatrix},
\]

\[
P = \begin{bmatrix} 4.365207 \\ 4.437197 \end{bmatrix},
\]

\[
SS_A = \begin{bmatrix} 1.07875085 \\ 0.08284527 \end{bmatrix}, \quad SS_B = \begin{bmatrix} 0.01320735 \\ 0.00824539 \end{bmatrix},
\]

\[
SS_e = \begin{bmatrix} 0.33166111 \\ 0.2090694 \end{bmatrix}, \quad SS_B = \begin{bmatrix} 0.2090694 \\ 0.2091579 \end{bmatrix},
\]

\[
S_e = \frac{1}{2} SS_A + \frac{1}{9} SS_B, \quad S_c = \frac{1}{96} SS_e,
\]

and
\[
\Sigma_j = \begin{bmatrix} 0.01817597 \\ 0.00316280 \end{bmatrix}, \quad 0.00316280 \end{bmatrix}, \quad 0.00805247
\]

With equation (8), the degree of freedom of \( \hat{\nu} \) is \( \nu = 2.970581 \). The tolerance factors defined in the tolerance region (3) via the values of \( \nu \) estimated by the approaches in \[7 \] and \[10 \] are given in the Tables II with respect to different content proportion and confidence level. The ellipsoid in Fig. 1 shows the tolerance region of data with content proportion \( \gamma = 0.9 \) and the degree of freedom \( \nu = 2.970581 \).

### IV. Summary

We propose a flexible approach for assessing the consistency of raw materials and/or final products of TCM. We use the tolerance region of the multivariate random effects model to assess the quality of TCM. If the tolerance region is within the permitted range, then the product is declared to pass the QC; otherwise, it should be rejected and further comparisons should be carried out. The large sample distributions of the sum of squares have been studied.

In general, TCM contains more than two active components. Our approach provides a quality control assessment to make sure that all the materials meet the requirement of quality simultaneously. This property provides the flexibility for assessing the consistency over a high-dimensional data corresponding to multiple components from different locations.
TABLE II
THE TOLERANCE FACTORS DEFINED IN THE TOLERANCE REGION (3) WITH $\tilde{V} = 2.970581$ AND $\hat{V} = 3.127606$ CORRESPONDING TO SPECIFIC CONTENT PROPORTION $\beta$ AND CONFIDENCE LEVEL $\gamma$

<table>
<thead>
<tr>
<th>$\hat{V}$</th>
<th>$\gamma = 0.8$</th>
<th>$\gamma = 0.85$</th>
<th>$\gamma = 0.9$</th>
<th>$\gamma = 0.95$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta = 0.8$</td>
<td>2.123451</td>
<td>2.451901</td>
<td>2.966166</td>
<td>4.003442</td>
</tr>
<tr>
<td>$\beta = 0.85$</td>
<td>2.497194</td>
<td>2.883454</td>
<td>3.485881</td>
<td>4.708076</td>
</tr>
<tr>
<td>$\beta = 0.9$</td>
<td>3.026625</td>
<td>3.494776</td>
<td>4.224924</td>
<td>5.706237</td>
</tr>
<tr>
<td>$\beta = 0.95$</td>
<td>3.941354</td>
<td>4.550993</td>
<td>5.501812</td>
<td>7.460819</td>
</tr>
</tbody>
</table>

$\tilde{V} = 2.970581$ (Krishnamoorthy and Yu 2004)

$\hat{V} = 3.127606$ (Nel and van der Merwe 1986)

Fig. 1 The tolerance region of TCM data with content proportion $\beta = 0.8$, confidence level $\gamma = 0.9$ and the degree of freedom $V = 2.970581$.

REFERENCES