Screening on Correlated Variables: A Bayesian Approach

Audrey Wong and Jeff B. Meeker
Research Statistics
Pharmaceuticals Division
CI BA-G EIGY Corporation
Summit, NJ 07901

Murray R. Selwyn
Analysis & Inference, Inc.
Boston, MA 02109

Let X and Y be two correlated variables, where the measurement of X is easier or less expensive to make than the one on Y. Suppose that the Y measurement is required to meet a certain specification limit. Given that the joint distribution of (X, Y) is bivariate normal and given the sufficient statistics from a random sample from this distribution, we propose a Bayesian method for determining a cutoff limit on X so that, with a guaranteed probability, Y meets its specification limit. Monte Carlo simulations are used to evaluate the Bayesian procedure. It is seen to be less conservative than the method of Owen, Li, and Chou (1981), and it is fairly robust under nonnormality. An example using data from one time point to predict results at another time point illustrates the application of the method.

KEY WORDS: Specification limit; Bivariate normal model; Bayesian methods; Monte Carlo methods.

1. INTRODUCTION

Consider two correlated measurements X and Y, where the measurement of X is easier or less expensive to make than the one on Y. Suppose that the Y measurement is required to meet a certain specification—for example, \( Y \geq y^* \). We present a statistical procedure, based on the assumption of bivariate normality with parameters unknown, for determining a specification limit on X so that with a guaranteed probability \( \delta \), Y meets its specification. With such a procedure, acceptance of future items can be based on the observed X values.

In a pharmaceutical testing environment, useful applications of this procedure would include the prediction of results for a long experiment based on a short experiment (see the example in Section 4) or the correlation of screening experiments in which a “secondary” screen is more expensive or time-consuming to conduct than a “primary” screen.

For instance, suppose that an investigator has a series of new compounds that need to be screened for biological activity in an in vivo (whole animal) experiment. Such activity, however, is found to be correlated to a particular measurement from an in vitro (e.g., test tube) experiment, which is easy and less expensive to perform. The investigator would like to be able to select compounds for further testing in the secondary screen that have a high probability of producing the required biological activity based on the results of the simpler test. This allows the investigator to reduce the number of in vivo experiments to those of the compounds that are more likely to be “active.”

Such a procedure could be applied in a manufacturing situation with respect to quality assessment of items, where one quality (e.g., nondestructive) might be predictive of another (e.g., destructive).

For X and Y that are positively correlated, the solution to such a screening problem is equivalent to solving for the specification limit on X (call it \( x^* \)) such that

\[
P(Y \geq y^* | X = x) \geq \delta \quad \text{for all } x \geq x^*,
\]

\[
P(Y \geq y^* | X = x) < \delta \quad \text{for all } x < x^*. \tag{1.1}
\]

Given that X and Y have a bivariate normal distribution and given a random sample \((x_1, y_1), \ldots, (x_n, y_n)\) from this distribution, we propose a Bayesian method of determining \( x^* \), which satisfies (1.1). Based on this value, we then accept new items if the observed X values are greater than \( x^* \). For simplicity we are assuming that X and Y are positively correlated, although the procedure applies to any nonzero correlation coefficient. A different approach to the screening problem given by Owen, Li, and Chou (OLC) (1981) provided the solution \( x^* \) such that

\[
P(Y \geq y^* | X \geq x^*) \geq \delta \tag{1.2}
\]

under the same normality assumption. Because their
procedure entails the use of Bonferroni's inequality, the computed value \( x^* \) is not exact for a specified \( \delta \). Moreover, since the probability distribution in (1.2) is conditioned on all values greater than \( x^* \), the actual probability of an accepted item (by using this cutoff) in meeting the \( Y \) specification could be less than \( \delta \), since one has “averaged” over \( x \)-values. Therefore, considering Equations (1.1) and (1.2), both of which could be employed to formulate the problem, comparison of the associated methods is also of interest.

The derivation of the Bayesian method and conditions necessary for solving (1.1) are presented in Sections 2 and 3. An example (in Section 4) using dissolution data from one timepoint to predict results at another timepoint illustrates the application of the method. Section 5 presents the results of Monte Carlo simulations that evaluate the Bayesian procedure and compare it to the OLC method.

2. DERIVATION

We assume that \((X, Y)\) have a bivariate normal distribution with mean \( \mu = (\mu_x, \mu_y)' \) and covariance matrix

\[
\Sigma = \begin{pmatrix} \sigma_x^2 & \rho \sigma_x \sigma_y \\ \rho \sigma_x \sigma_y & \sigma_y^2 \end{pmatrix}.
\]

Given a random sample of \((x, y)\) of size \( n \) from this distribution, we need to derive the predictive distribution of \( Y_{n+1} \), given \( X_{n+1} \), for a new observation. This can be obtained by combining the posterior density of \((\mu, \Sigma)\) with the conditional density of \( Y_{n+1} \), given \( X_{n+1} \), and integrating out the parameters; that is,

\[
p(Y_{n+1} | X_{n+1} = x_{n+1}, x, y) = \int p(\mu, \Sigma | x, y) \times p(Y_{n+1} | X_{n+1} = x_{n+1}, \mu, \Sigma) d\mu d\Sigma.
\]

The prior on \((\mu, \Sigma)\) is noninformative (Box and Tiao 1973, pp. 20-60) and obtained via Jeffreys’ rule as

\[
p(\mu, \Sigma) \propto |\Sigma|^{-3/2}.
\]

Combining this prior with the multivariate normal likelihood yields the posterior density of \((\mu, \Sigma)\). Since \( p(Y_{n+1} | X_{n+1} = x_{n+1}, \mu, \Sigma) \) is

\[
N(\mu_y + \rho \sigma_y (x_{n+1} - \mu_x) / \sigma_x, \sigma_y^2 (1 - \rho^2)),
\]

the predictive density defined in (2.1) reduces to

\[
\frac{y_{n+1} - \bar{y}_n - r \sigma_y (x_{n+1} - \bar{x}_n) / \sigma_x}{(\sigma_y^2 (1 - \rho^2) [1 + 1/n + (x_{n+1} - \bar{x}_n)^2 / ((n-1) \bar{y}_n)]^{-1/2}} = t_{n-1,1-\delta}.
\]

where \( \bar{x}_n, \bar{y}_n, \sigma_x^2, \sigma_y^2, \) and \( r \) are the usual means, variances, and correlation coefficients (i.e., the sufficient statistics) calculated from the sample of size \( n \) [see the Appendix for the derivation of (2.2)]. The frequency interpretation of (2.2) is the sampling density of the difference of \( y_{n+1} \) and its least squares estimate divided by its sampling variance conditional on \( x \) and \( x_{n+1} \). In the Bayesian approach, however, \( x \) and \( x_{n+1} \) are considered random.

3. CONDITIONS FOR SOLUTION

With the distribution of a future \( Y_{n+1} \), given \( X_{n+1} \), defined in (2.2) and assuming \( r > 0 \), the solution of (1.1) is the value of \( x^* \) such that

\[
u(x) = \frac{y^* - y_n - rs_y (x - \bar{x}_n) / \sigma_y}{c(x)} < t_{n-1,1-\delta} \quad \text{for } x \geq x^*,
\]

\[
u(x) > t_{n-1,1-\delta} \quad \text{for } x < x^*,
\]

where \( c(x) = \left[ s_y^2 (1 - r^2) [1 + 1/n + (x - \bar{x}_n)^2 / ((n-1) \bar{y}_n)]^{-1/2} \right] 
\] and \( t_{n-1,1-\delta} \) is the deviate corresponding to tail area \( 1 - \delta \) in the lower tail of the Student-t distribution with \( n - 1 \) degrees of freedom. Since

\[
u(x) = -rs_y (x_{n+1} - \bar{x}_n) / \sigma_y
\]

\[
\frac{s_x c(x)}{\{s_x^2 (1 - r^2) [1 + 1/n + (x - \bar{x}_n)^2 / ((n-1) \bar{y}_n)]^{-1/2} \}^{1/2}} = M,
\]

where \( M = r \sigma_y [s_y^2 (1 - r^2) [1 + 1/n + (x - \bar{x}_n)^2 / ((n-1) \bar{y}_n)]^{-1/2}] = -M \),

\[
\lim_{x \to \infty} \frac{rs_y (x - \bar{x}_n) / \sigma_y}{\{s_x^2 (1 - r^2) [1 + 1/n + (x - \bar{x}_n)^2 / ((n-1) \bar{y}_n)]^{-1/2} \}^{1/2}} = -M,
\]

where \( M = r \sigma_y [s_y^2 (1 - r^2) [1 + 1/n + (x - \bar{x}_n)^2 / ((n-1) \bar{y}_n)]^{-1/2}] = -M \). Conditions (i) and (ii) imply that \( u(x) < -M \) for any \( x \) in \((x_0, \infty)\); that is, the increasing portion of the function stays below \(-M\). Thus, above \(-M\), \( u \) is a decreasing function, and as \( x \to \infty \), \( u(x) \to -M \), \( u(x) \to M \), \( u(x) \to M \), \( u(x) \to M \).

Conversely, if \( y^* > \bar{y}_n \),

\[
u(x) > 0 \quad \text{for } x > x_0.
\]

and

\[
u(x) < 0 \quad \text{for } x < x_0.
\]

For illustration purpose, consider the case \( y^* < \bar{y}_n \). Then (i) \( u(x) \) is decreasing for \( x \) in \((\infty, x_0)\) and increasing for \( x \) in \((x_0, \infty)\). As \( x \to \infty \), however, \( u(x) \to M \).

\[
u(x) = 0 \quad \text{for } x = \bar{x}_n - r \sigma_y / c(x),
\]

and

\[
u(x) > 0 \quad \text{for } x < x_0.
\]

and

\[
u(x) < 0 \quad \text{for } x > x_0.
\]

The condition for the existence of \( x^* \) also holds for \( y^* > \bar{y}_n \), where, below \( M \), \( u \) is a decreasing function. Therefore, given that \( r^2 (1 - \delta) = M \), the solution \( x^* \) is such that

\[
y^* - \bar{y}_n - rs_y (x^* - \bar{x}_n) / \sigma_y < t_{n-1,1-\delta}.
\]
This solution can be obtained by solving the quadratic equation $A(x^* - \bar{x}_n)^2 + B(x^* - \bar{y}_n) + C = 0$, obtained by squaring (3.2), where $A = b^2 - t^2s^2/(n - 1)s^2$; $B = -2b(y^* - \bar{y}_n)$; and $C = (y^* - \bar{y}_n)^2 - t^2s^2(n + 1)/n$ with $b = r\bar{y}_n/s\bar{x}_n$, $s^2 = s^2_1(1 - r^2)$, and $t = t_{n - 1, \delta}$. It can be shown that the condition $t^2 < M^2$, which is equivalent to $A > 0$, implies that

1. the discriminant (i.e., $B^2 - 4AC$) is positive (therefore the roots are real)
2. only one of the roots can satisfy (3.2).

Thus the solution is the one of the two roots

$$x^* = \bar{x}_n + (-B \pm \sqrt{B^2 - 4AC})/(2A)$$

that corresponds to $r_{n - 1, 1 - \delta}$ in (3.2). "Erratic solutions," however, may occur when $A$ is close to 0.

Graphically the solution $x^*$ can be obtained by plotting the prediction intervals for $Y$ over the range of $X$ values using the formula

$$y = \bar{y}_n + rs_n(x - \bar{x}_n)/s_x \pm tc(x).$$

The cutoff $x^*$ corresponds to the $x$ value at which $y^*$ intersects the lower or upper prediction curve depending on whether $\delta > .5$ or $\delta < .5$. If $t^2 > M^2$ (i.e., the correlation is small) we can have the situation in which $y^*$ never intersects either the lower or the upper prediction curve or that in which $y^*$ intersects only the lower or the upper prediction curve. The first case [Figure 2(a)] corresponds to the discriminant of the quadratic form ($B^2 - 4AC$) being negative, and the latter occurs when the discriminant is positive and $A < 0$ [see Figure 2(b)]. Therefore, if the condition $t^2 < M^2$ is satisfied, one can always obtain a solution $x^*$ either graphically or algebraically as described before.

4. AN EXAMPLE

In the dissolution testing to assure the quality of batches of a pharmaceutical product, the active ingredient in samples was measured at two points in time (Times 1 and 2). In order to accept new batches, we want to be able to guarantee that with 99% probability, the cumulative release at Time 2 exceeds 1,500. Using the correlation between the dissolution at Time 1 ($X$) and Time 2 ($Y$), acceptance of new batches can, therefore, be based on the amount released at Time 1 without having to prolong the test to Time 2.

Dissolution data at Time 1 and Time 2 were collected from 10 batches for a particular compound, and the following sufficient statistics were computed: $\bar{x}_n = 1,256$, $\bar{y}_n = 1,969$, $s_x = 133$, $s_y = 177$, and $r = .975$. Normal plots of $x$ and $Y$ and a half-normal plot of the residuals from regressing $Y$ on $x$ are displayed in Figure 3, and they can be used to check the normality assumption. Since

$$M^2 = r^2(n - 1)/(1 - r^2) = 173.3 > t^2_{99, 10} = 7.95,$$

a solution can be obtained for $\delta = .99$. By applying the Bayesian method as described in Section 3, we obtain an $x^*$ value of 1,000, which corresponds to the upper 1st percentile of the predictive distribution. Hence, using 1,000 as the specification limit for the cumulative release at Time 1, the probability that new accepted batches will meet the specification of 1,500 at Time 2 is .99. For this example the specification limit $x^*$ obtained from the OLC method is 1,340, considerably higher than that obtained from the Bayesian method and even close to the 1,500 cumulative release specification limit for Time 2.

5. MONTE CARLO SIMULATIONS

Monte Carlo studies were performed to evaluate the Bayesian procedure. The behavior of $x^*$ was studied under various values of $n$, $y^*$, and positive $\rho$. In each case we generated 100 independent samples using the normal pseudo-random number generator in SAS® (SAS Institute, Inc. 1982). Each sample consisted of $n$ independent observations from a bivariate normal distribution with $\mu_x = \mu_y = 0$, $\sigma^2_x = \sigma^2_y = 1$, and a particular $\rho$. The solution $x^*$ was calculated for each sample as described in Section 3, except for cases in which $t^2 > M^2$ (see Section 3) or the sample correlation coefficient was negative. In particular, results on the distribution of $x^*$ will be used to evaluate the performance of this procedure. Alternatively we could study values of $y$ likely to be accepted in practice or probabilities that certain $y$'s would be acceptable. Since these results can easily be computed once $x^*$ is determined, information on the distribution of $x^*$ will thus be presented.

To give an idea of both the central tendency of $x^*$ estimates and the variation from sample to sample, the 5th, 25th, 50th, 75th, and 95th percentiles of the distribution of $x^*$ are presented in Table 1 for $n = 6$, 11, 31; $y^* = 0$, .675, 1.28; $\delta = .75, .90, .99$; and $\rho = .75, .90, .99$. For comparison purposes we also
display the "correct" value of \( x^* \)—that is, the value of \( x^* \) that could be calculated if all parameters were known, given by

\[
x^* = \left[ y^* - (1 - \rho^2)^{1/2} \Phi^{-1}(1 - \delta) \right] / \rho,
\]

where \( \Phi^{-1}(\cdot) \) is the inverse of the standard normal cdf. Sample sizes \( n \) of 6, 11, and 31 were chosen to allow comparison with the OLC method using available tables.

Results obtained with \( n = 31 \) show that the distribution of \( x^* \) is nearly symmetric and that the median of \( x^* \) is close to the correct value; the spread of \( x^* \) values seen by comparing the 5th and 95th percentiles is fairly small. Similar results are also obtained with \( n = 11 \), except for the combinations of small \( \rho(0.75) \) and high \( \delta(0.99) \), where the percentiles at the upper end are large. This is due to cases in which \( A \), the coefficient of the quadratic term, is near 0. The median of \( x^* \), however, is still close to the correct value. Even for \( n = 6 \), satisfactory results are also obtained for \( \rho = 0.9 \) and 0.99.

Monte Carlo studies were also performed to compare the Bayesian procedure with the OLC method. Again 100 independent samples were generated for each case in the same manner as before. In each sample the solution \( x^* \) using the OLC procedure was computed for accepting a single future observation. There were cases in which no solution was found for the OLC method because the lower confidence limit for \( y = P(Y \geq y^*) \) or \( \rho \) was too low and was not covered by the available tables.

Because of the limitations of the available tables needed for applying the OLC procedure, the choices of \( y^* \), \( \delta \), and \( \rho \) are quite restricted. Results on the distribution of \( x^* \) are presented in Table 2 for \( n = 31 \); \( y^* = -0.675 \) (\( \delta = 0.90, 0.95 \)), \( -1.28 \) (\( \delta = 0.95, 0.98 \)); and \( \rho = 0.75, 0.90, 0.99 \). The correct value of \( x^* \) is obtained from the tables of Odeh and Owen (1980).
and Odeh (1982), satisfying $P(Y \geq y^* | X \geq x^*) = \delta$ under the case of all parameters known.

For comparison purposes, $x^*$ values obtained from separate independent samples using the Bayesian method are also given. The solutions obtained from the OLC method in all cases are much higher than the correct value and are more conservative than the ones from the Bayesian method, even though the OLC method is less stringent in theory. In fact, there is no overlap in the 5th percentile of the OLC method and the 95th percentile of the Bayesian method.

Further Monte Carlo studies were performed to investigate the "robustness" of the Bayesian procedure. Primarily, we are concerned with the sensitivity of this procedure to deviations from the normality assumption. Three types of distributions (symmetric short-tailed, symmetric long-tailed, asymmetric) were considered:

1. a bivariate uniform distribution in which both $X$ (with mean 0, variance 1) and $Y$, given $X = x$ (with mean $\rho x$, variance $1 - \rho^2$), were uniformly distributed.

2. a bivariate scale contaminated normal (SCN): $X$ and $Y$ distributed as $N((0, \sigma), \delta (\phi, \phi))$ with probability .9 and $N((0, 5), (\phi, \phi))$ with probability .1, the result is a symmetric long-tailed distribution, where the marginal distributions of $X$ and $Y$ have means 0 and variances 1, and the correlation coefficient between $X$ and $Y$ is $\rho$.

3. a bivariate location contaminated normal (LCN): $X$ and $Y$ are distributed as

$$N\left(\left(\frac{-3}{\sqrt{1.81}}, \frac{1}{1.81}\right), \left(\rho, \rho\right)\right)$$

with probability .9.

and

$$N\left(\left(\frac{2.7}{\sqrt{1.81}}, \frac{1}{1.81}\right), \left(\rho, \rho\right)\right)$$

with probability .1;

the result is a bimodal skewed distribution, where $X$ and $Y$ each has mean 0 and variance 1, and the correlation coefficient between $X$ and $Y$ is $(\rho + .31)/1.81$.

One hundred independent samples were generated for each combination of $n = 11, 31$; $y^* = .675, 1.28$; $\delta = .75, .90, .99$; and $\rho = .75, .95$, with respect to each type of distribution. Higher $\rho$ values were not presented, since for high values of $\rho$, $y$ values are always tightly clustered around the regression line, irrespective of distribution. In each sample the solution $x^*$ was calculated, and the results on the distribution of $x^*$ are presented in Table 3, with correct $x^*$ values given by:

1. for the uniform,

$$x^*_T = \frac{y^* + 2(3(1 - \rho^2))^{1/2}(\delta - .5)}{\rho};$$

2. for the SCN, $x^*_C$ is the solution of

$$0.9\Phi\left(\frac{y^* - \rho x^*_C}{\sqrt{5(1 - \rho^2)}}\right) + 0.1\Phi\left(\frac{y^* - \rho x^*_C}{\sqrt{5(1 - \rho^2)}}\right) = 1 - \delta;$$

3. for the LCN, $x^*_L$ is the solution of

$$0.9\Phi\left(\frac{y^* - \mu_1 - \rho(x^*_L - \mu_2)}{\sigma(1 - \rho^2)^{1/2}}\right) + 0.1\Phi\left(\frac{y^* - \mu_1 - \rho(x^*_L - \mu_2)}{\sigma(1 - \rho^2)^{1/2}}\right) = 1 - \delta,$$

where $\mu_1 = -3/\sqrt{1.81}$, $\mu_2 = 2.7/\sqrt{1.81}$, and $\sigma = 1/\sqrt{1.81}$. 

TECHNOMETRICS, NOVEMBER 1985, VOL. 27, NO. 4
Table 1. Monte Carlo Results for the Normal

<table>
<thead>
<tr>
<th>$\gamma^*$</th>
<th>$\delta$</th>
<th>$p$</th>
<th>$n$</th>
<th>$N$</th>
<th>$5%$</th>
<th>$25%$</th>
<th>$50%$</th>
<th>$75%$</th>
<th>$95%$</th>
<th>$\chi^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.75</td>
<td>.75</td>
<td>6</td>
<td>93</td>
<td>-.18</td>
<td>.12</td>
<td>.43</td>
<td>.87</td>
<td>2.42</td>
<td>.59</td>
</tr>
<tr>
<td>90</td>
<td>6</td>
<td>99</td>
<td>0</td>
<td>15</td>
<td>.08</td>
<td>.28</td>
<td>.42</td>
<td>.76</td>
<td></td>
<td>.33</td>
</tr>
<tr>
<td>90</td>
<td>6</td>
<td>100</td>
<td>0</td>
<td>.08</td>
<td>.21</td>
<td>.34</td>
<td>.51</td>
<td>.76</td>
<td></td>
<td>.99</td>
</tr>
<tr>
<td>90</td>
<td>6</td>
<td>100</td>
<td>0</td>
<td>.17</td>
<td>.27</td>
<td>.34</td>
<td>.41</td>
<td>.52</td>
<td></td>
<td>31</td>
</tr>
<tr>
<td>90</td>
<td>6</td>
<td>100</td>
<td>0</td>
<td>.06</td>
<td>.08</td>
<td>.09</td>
<td>.11</td>
<td>.14</td>
<td></td>
<td>99</td>
</tr>
<tr>
<td>.00 .75</td>
<td>0</td>
<td>6</td>
<td>95</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
</tr>
<tr>
<td>.00 .75</td>
<td>0</td>
<td>6</td>
<td>85</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
</tr>
<tr>
<td>.00 .75</td>
<td>0</td>
<td>6</td>
<td>85</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
</tr>
<tr>
<td>.00 .75</td>
<td>0</td>
<td>6</td>
<td>85</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
</tr>
</tbody>
</table>

NOTE: $N$ is the number of samples out of 100 in which solutions $x^*$ are obtained.
Table 2. Monte Carlo Comparison of Methods (n = 31)

<table>
<thead>
<tr>
<th>y*</th>
<th>δ</th>
<th>p</th>
<th>Method</th>
<th>N</th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
<th>x*</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.675</td>
<td>.90</td>
<td>.75</td>
<td>Bayesian</td>
<td>100</td>
<td>-0.02</td>
<td>.15</td>
<td>.27</td>
<td>.40</td>
<td>.63</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td>OLC</td>
<td>.78</td>
<td>1.27</td>
<td>1.68</td>
<td>2.02</td>
<td>2.64</td>
<td>.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.90</td>
<td>Bayesian</td>
<td>100</td>
<td>-0.34</td>
<td>-0.18</td>
<td>-0.10</td>
<td>-0.01</td>
<td>.07</td>
<td>-1.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OLC</td>
<td>.13</td>
<td>.33</td>
<td>.46</td>
<td>.62</td>
<td>.86</td>
<td>.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.99</td>
<td>Bayesian</td>
<td>100</td>
<td>-0.55</td>
<td>-0.52</td>
<td>-0.50</td>
<td>-0.47</td>
<td>-0.43</td>
<td>-0.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OLC</td>
<td>.31</td>
<td>-0.26</td>
<td>-0.21</td>
<td>-0.18</td>
<td>-0.14</td>
<td>-0.96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.95</td>
<td>Bayesian</td>
<td>100</td>
<td>.20</td>
<td>.46</td>
<td>.60</td>
<td>.80</td>
<td>1.04</td>
<td>.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OLC</td>
<td>1.61</td>
<td>2.67</td>
<td>3.52</td>
<td>4.48</td>
<td>6.41</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.90</td>
<td>Bayesian</td>
<td>100</td>
<td>-0.16</td>
<td>0.00</td>
<td>0.08</td>
<td>0.17</td>
<td>0.30</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OLC</td>
<td>0.87</td>
<td>1.09</td>
<td>1.27</td>
<td>1.47</td>
<td>1.95</td>
<td>.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.99</td>
<td>Bayesian</td>
<td>100</td>
<td>-0.50</td>
<td>-0.47</td>
<td>-0.44</td>
<td>-0.42</td>
<td>-0.37</td>
<td>-0.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OLC</td>
<td>0.93</td>
<td>0.07</td>
<td>0.01</td>
<td>0.05</td>
<td>0.11</td>
<td>0.18</td>
<td>-0.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1.28</td>
<td>.95</td>
<td>.75</td>
<td>Bayesian</td>
<td>100</td>
<td>-0.70</td>
<td>-0.41</td>
<td>-0.20</td>
<td>-0.07</td>
<td>0.21</td>
<td>-0.26</td>
</tr>
<tr>
<td></td>
<td>OLC</td>
<td>0.64</td>
<td>1.59</td>
<td>2.18</td>
<td>2.80</td>
<td>3.69</td>
<td>-1.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.90</td>
<td>Bayesian</td>
<td>100</td>
<td>-0.86</td>
<td>-0.72</td>
<td>-0.62</td>
<td>-0.52</td>
<td>-0.35</td>
<td>-0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OLC</td>
<td>0.92</td>
<td>0.29</td>
<td>0.46</td>
<td>0.68</td>
<td>1.10</td>
<td>-1.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.99</td>
<td>Bayesian</td>
<td>100</td>
<td>-1.14</td>
<td>-1.08</td>
<td>-1.04</td>
<td>-1.01</td>
<td>-0.98</td>
<td>-1.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OLC</td>
<td>0.94</td>
<td>-0.64</td>
<td>-0.57</td>
<td>-0.52</td>
<td>-0.48</td>
<td>-0.36</td>
<td>-1.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.98</td>
<td>.75</td>
<td>Bayesian</td>
<td>100</td>
<td>-0.36</td>
<td>-0.05</td>
<td>0.17</td>
<td>0.35</td>
<td>0.69</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OLC</td>
<td>1.35</td>
<td>2.08</td>
<td>2.26</td>
<td>3.04</td>
<td>4.00</td>
<td>-0.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.90</td>
<td>Bayesian</td>
<td>100</td>
<td>-0.65</td>
<td>-0.48</td>
<td>-0.30</td>
<td>-0.26</td>
<td>-0.00</td>
<td>-0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OLC</td>
<td>0.52</td>
<td>0.67</td>
<td>0.94</td>
<td>1.19</td>
<td>1.60</td>
<td>-0.95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.99</td>
<td>Bayesian</td>
<td>100</td>
<td>1.08</td>
<td>1.02</td>
<td>0.98</td>
<td>0.96</td>
<td>0.90</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OLC</td>
<td>0.47</td>
<td>-0.38</td>
<td>-0.33</td>
<td>-0.27</td>
<td>-0.18</td>
<td>-1.34</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: N is the number of samples out of 100 in which solutions x* are obtained. OLC—Owen, Li, and Chou (1981).

Results from the Monte Carlo simulations show that for δ = .75 and .90, the median of x* is fairly close to the correct value even under distributions other than normal. The distribution of x* is nearly symmetric for n = 31 and is skewed to the right for n = 11. For high δ(.99) the percentiles at the upper end are large, particularly for small n, and the median is somewhat higher than the correct value, as seen in the normal case. Thus the procedure seems to behave reasonably, even under the alternative distributions studied.

6. DISCUSSION

A Bayesian method to determine the specification x* to assure, with probability δ, that a specification y* is met has been presented. Monte Carlo simulations show that the Bayesian procedure gives satisfactory results for the cases studied, though for small samples and low correlation the solutions can be somewhat conservative. Hence a moderate sample size is recommended for this procedure unless X and Y are very highly correlated. The method of Owen, Li, and Chou (1981) is seen to be much more conservative than the Bayesian method, and values obtained are not nearly as close to the correct values.

The Bayesian method is easy to apply and is not restricted because of the necessity of having special tables.

In practice, the application of a screening method such as the one proposed should be implemented carefully along with other appropriate techniques. Since x* may be used as a cutoff value for screening many future items, a periodic resampling of (x, y) should be used to validate the procedure as well as to provide additional data. Strategies for updating the bivariate sample, however, are beyond the scope of this article.

Since the Bayesian procedure depends on the bivariate normal model, diagnostic techniques (such as normal plots of raw data or residuals) should be used as an aid in checking the normality assumption. If the normality assumption is noticeably violated, one may try to normalize the data using transformations. Nevertheless, the procedure is seen to be fairly insensitive to departures from normality.
## Table 3. Monte Carlo Simulations on Robustness (\( p = .75 \))

<table>
<thead>
<tr>
<th>( n )</th>
<th>( y^* )</th>
<th>( \delta )</th>
<th>Distribution</th>
<th>( N )</th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
<th>( x^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>.675</td>
<td>.75</td>
<td>Normal</td>
<td>100</td>
<td>.74</td>
<td>1.22</td>
<td>1.54</td>
<td>2.00</td>
<td>4.12</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Uniform</td>
<td>100</td>
<td>.93</td>
<td>1.28</td>
<td>1.60</td>
<td>2.05</td>
<td>3.57</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SCN</td>
<td>96</td>
<td>.79</td>
<td>1.11</td>
<td>1.45</td>
<td>2.41</td>
<td>4.92</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LCN</td>
<td>100</td>
<td>.79</td>
<td>1.22</td>
<td>1.62</td>
<td>2.09</td>
<td>3.56</td>
<td>1.37</td>
</tr>
<tr>
<td>90</td>
<td></td>
<td></td>
<td>Normal</td>
<td>91</td>
<td>1.70</td>
<td>1.67</td>
<td>2.77</td>
<td>2.87</td>
<td>6.67</td>
<td>2.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Uniform</td>
<td>98</td>
<td>1.27</td>
<td>1.79</td>
<td>2.31</td>
<td>3.22</td>
<td>6.44</td>
<td>2.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SCN</td>
<td>92</td>
<td>1.05</td>
<td>1.69</td>
<td>2.28</td>
<td>3.26</td>
<td>6.24</td>
<td>1.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LCN</td>
<td>99</td>
<td>1.40</td>
<td>1.95</td>
<td>2.38</td>
<td>3.08</td>
<td>4.89</td>
<td>1.78</td>
</tr>
<tr>
<td>1.28</td>
<td>.75</td>
<td></td>
<td>Normal</td>
<td>99</td>
<td>1.45</td>
<td>1.87</td>
<td>2.52</td>
<td>3.27</td>
<td>6.17</td>
<td>2.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Uniform</td>
<td>100</td>
<td>1.56</td>
<td>1.92</td>
<td>2.40</td>
<td>3.06</td>
<td>4.33</td>
<td>2.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SCN</td>
<td>97</td>
<td>1.22</td>
<td>1.70</td>
<td>2.32</td>
<td>3.17</td>
<td>10.96</td>
<td>2.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LCN</td>
<td>100</td>
<td>1.36</td>
<td>1.87</td>
<td>2.30</td>
<td>3.25</td>
<td>6.07</td>
<td>2.18</td>
</tr>
<tr>
<td>90</td>
<td></td>
<td></td>
<td>Normal</td>
<td>96</td>
<td>1.88</td>
<td>2.72</td>
<td>3.85</td>
<td>4.43</td>
<td>8.06</td>
<td>2.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Uniform</td>
<td>99</td>
<td>1.98</td>
<td>2.65</td>
<td>3.60</td>
<td>4.85</td>
<td>13.88</td>
<td>2.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SCN</td>
<td>88</td>
<td>1.72</td>
<td>2.36</td>
<td>3.26</td>
<td>4.52</td>
<td>31.46</td>
<td>2.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LCN</td>
<td>99</td>
<td>1.82</td>
<td>2.54</td>
<td>3.08</td>
<td>3.92</td>
<td>7.79</td>
<td>2.59</td>
</tr>
<tr>
<td>99</td>
<td></td>
<td></td>
<td>Normal</td>
<td>78</td>
<td>3.05</td>
<td>4.50</td>
<td>6.53</td>
<td>9.88</td>
<td>37.17</td>
<td>3.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Uniform</td>
<td>77</td>
<td>2.90</td>
<td>5.15</td>
<td>7.61</td>
<td>11.68</td>
<td>35.03</td>
<td>3.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SCN</td>
<td>69</td>
<td>2.74</td>
<td>3.47</td>
<td>4.88</td>
<td>10.47</td>
<td>115.2</td>
<td>4.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LCN</td>
<td>87</td>
<td>3.27</td>
<td>4.42</td>
<td>5.84</td>
<td>9.09</td>
<td>26.09</td>
<td>3.28</td>
</tr>
<tr>
<td>31</td>
<td>.675</td>
<td>.75</td>
<td>Normal</td>
<td>100</td>
<td>1.10</td>
<td>1.31</td>
<td>1.53</td>
<td>1.82</td>
<td>2.38</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Uniform</td>
<td>100</td>
<td>1.08</td>
<td>1.27</td>
<td>1.51</td>
<td>1.74</td>
<td>2.06</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SCN</td>
<td>100</td>
<td>.96</td>
<td>1.21</td>
<td>1.40</td>
<td>1.77</td>
<td>3.11</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LCN</td>
<td>100</td>
<td>1.01</td>
<td>1.34</td>
<td>1.53</td>
<td>1.79</td>
<td>2.60</td>
<td>1.37</td>
</tr>
<tr>
<td>90</td>
<td></td>
<td></td>
<td>Normal</td>
<td>100</td>
<td>1.48</td>
<td>1.83</td>
<td>2.22</td>
<td>2.53</td>
<td>3.60</td>
<td>2.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Uniform</td>
<td>100</td>
<td>1.55</td>
<td>1.85</td>
<td>2.16</td>
<td>2.46</td>
<td>2.92</td>
<td>2.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SCN</td>
<td>100</td>
<td>1.37</td>
<td>1.72</td>
<td>2.15</td>
<td>2.62</td>
<td>4.03</td>
<td>1.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LCN</td>
<td>100</td>
<td>1.69</td>
<td>2.06</td>
<td>2.30</td>
<td>2.58</td>
<td>3.73</td>
<td>1.78</td>
</tr>
<tr>
<td>99</td>
<td></td>
<td></td>
<td>Normal</td>
<td>98</td>
<td>3.38</td>
<td>3.02</td>
<td>3.56</td>
<td>4.20</td>
<td>5.34</td>
<td>2.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Uniform</td>
<td>99</td>
<td>3.39</td>
<td>3.06</td>
<td>3.30</td>
<td>4.04</td>
<td>5.81</td>
<td>2.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SCN</td>
<td>98</td>
<td>1.99</td>
<td>2.48</td>
<td>3.05</td>
<td>4.15</td>
<td>8.91</td>
<td>3.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LCN</td>
<td>98</td>
<td>2.70</td>
<td>3.25</td>
<td>3.75</td>
<td>4.46</td>
<td>5.89</td>
<td>2.47</td>
</tr>
<tr>
<td>1.28</td>
<td>.75</td>
<td></td>
<td>Normal</td>
<td>100</td>
<td>1.81</td>
<td>2.11</td>
<td>2.42</td>
<td>2.77</td>
<td>3.84</td>
<td>2.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Uniform</td>
<td>100</td>
<td>1.84</td>
<td>2.16</td>
<td>2.43</td>
<td>2.76</td>
<td>3.29</td>
<td>2.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SCN</td>
<td>100</td>
<td>1.69</td>
<td>2.03</td>
<td>2.33</td>
<td>2.71</td>
<td>5.58</td>
<td>2.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LCN</td>
<td>100</td>
<td>1.73</td>
<td>2.02</td>
<td>2.23</td>
<td>2.47</td>
<td>3.18</td>
<td>2.18</td>
</tr>
<tr>
<td>90</td>
<td></td>
<td></td>
<td>Normal</td>
<td>100</td>
<td>2.35</td>
<td>2.70</td>
<td>2.95</td>
<td>3.66</td>
<td>5.16</td>
<td>2.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Uniform</td>
<td>100</td>
<td>2.10</td>
<td>2.65</td>
<td>3.07</td>
<td>3.53</td>
<td>4.04</td>
<td>2.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SCN</td>
<td>100</td>
<td>2.01</td>
<td>2.39</td>
<td>3.10</td>
<td>3.81</td>
<td>5.41</td>
<td>2.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LCN</td>
<td>100</td>
<td>2.17</td>
<td>2.61</td>
<td>3.06</td>
<td>3.49</td>
<td>4.16</td>
<td>2.59</td>
</tr>
<tr>
<td>99</td>
<td></td>
<td></td>
<td>Normal</td>
<td>100</td>
<td>3.08</td>
<td>3.87</td>
<td>4.53</td>
<td>5.51</td>
<td>10.54</td>
<td>3.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Uniform</td>
<td>100</td>
<td>3.18</td>
<td>3.86</td>
<td>4.49</td>
<td>5.10</td>
<td>6.11</td>
<td>3.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SCN</td>
<td>100</td>
<td>2.56</td>
<td>3.48</td>
<td>4.40</td>
<td>5.50</td>
<td>12.57</td>
<td>4.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LCN</td>
<td>100</td>
<td>3.33</td>
<td>4.01</td>
<td>4.61</td>
<td>5.30</td>
<td>6.76</td>
<td>3.28</td>
</tr>
</tbody>
</table>

**Note:** \( N \) is the number of samples out of 1000 in which solutions \( x^* \) are obtained. SCN—scale contaminated normal; LCN—location contaminated normal.

## Acknowledgments

We are grateful to Arthur Dempster, the editor, and the referees for their valuable comments, which led to an improved version of this article, and to Robert Odeh for the extension of the tables for screening that were used in the Monte Carlo studies.

**Technometrics, November 1985, Vol. 27, No. 4**
Components of the predictive density of $Y_{n+1}$, given $X_{n+1}$, and the data $(x, y)$ are

1. the joint posterior density of $(\mu, \Sigma)$, obtained by combining the noninformative prior with the multivariate normal likelihood [using (8.1.7), (8.2.6), and (8.2.8) of Box and Tiao (1973)], yielding

\[
p(\mu, \Sigma | x, y) \propto \left[ \sigma^2 \sigma^2(1 - \rho^2) \right]^{-n+3/2} \times \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 / \sigma^2 + \frac{1}{2} \sum_{i=1}^{n+1} (y_i - \mu_y - s(x_i - \mu_x))^2 / \sigma^2_{x} (1 - \rho^2) \right\},
\]

where $-\infty < \mu_x < \infty$, $-\infty < \mu_y < \infty$, $\sigma^2 > 0$, $\sigma^2_x > 0$, $-1 < \rho < 1$;

2. the density $p(Y_{n+1} | X_{n+1} = x_{n+1}, \mu, \Sigma)$, which is $N(\mu_y + \rho \sigma_{x} (x_{n+1} - \mu_x) / \sigma_{x}, \sigma^2_{x} (1 - \rho^2))$.

By combining components 1 and 2 and by reparameterizing with $w = \sigma^2_x$, $r = \sigma^2_x (1 - \rho^2)$, and $s = \rho \sigma_{x} / \sigma_{x}$, the predictive density can be written as

\[
p(Y_{n+1} | X_{n+1} = x_{n+1}, x, y) \propto \int w^{-(n+1)/2} r^{-(n+4)/2} \times \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu_x)^2 / w + \frac{1}{r} \sum_{i=1}^{n+1} (y_i - \mu_y - s(x_i - \mu_x))^2 \right\} d\mu_x d\mu_y dr dw ds,
\]

where $-\infty < \mu_x < \infty$, $-\infty < \mu_y < \infty$, $w > 0$, $r > 0$, and $-\infty < s < \infty$. Applying (A.2.1.2) of Box and Tiao (1973) to integrate out $w$ and $r$, respectively, (A.1) reduces to

\[
p(Y_{n+1} | X_{n+1} = x_{n+1}, x, y) \propto \int \left\{ \sum_{i=1}^{n+1} (y_i - \mu_y - s(x_i - \mu_x))^2 \right\}^{-(n+2)/2} d\mu_x d\mu_y ds.
\]

Adding and subtracting $(n+1)(\bar{y}_{n+1} - s\bar{x}_{n+1})^2$ to \( \sum_{i=1}^{n+1} (y_i - \mu_y - s(x_i - \mu_x))^2 \) gives

\[
\sum_{i=1}^{n+1} (y_i - \mu_y - s(x_i - \bar{x}_{n+1})^2 = (n+1)c^2_{xy}
\]

where $c^2_{xy} = \sum_{i=1}^{n+1} (y_i - \bar{y}_{n+1} - s(x_i - \bar{x}_{n+1}))^2 / (n+1)$. Substituting (A.3) for (A.2), $\mu_x$ can be easily integrated out. By repeating the same technique to integrate out $\mu_x$ and $s$, respectively, the predictive density in (2.1) finally reduces to

\[
p(Y_{n+1} | X_{n+1} = x_{n+1}, x, y) \propto \int \left\{ 1 + \frac{1}{n} \left[ \frac{1}{n} \left[ \frac{(x_{n+1} - \bar{x}_{n+1})^2}{s_x^2 (1 - r^2)} + \frac{(y_{n+1} - \bar{y}_{n+1})^2}{s_y^2} \right] \right] \right\}^{-(n+2)/2} d\mu_x d\mu_y ds.
\]

The density in (A.4) is easily recognized as a $t$ density, as given in (2.2),

\[\text{REFERENCES}\]


