AN M-POPULATION COVARIANCE STABILIZATION: PROC VARSSTB

J. E. Dunn, University of Arkansas
Mike Zodrow, Duke Power Company

Key Words and Phrases: Power transformations; Multivariate; Normalization; Variance stabilization; Maximum likelihood.

ABSTRACT
Building on the notion of data based power transformations, commonly referred to as Box-Cox transformations, a SAS procedure has been written for simultaneous normalization and covariance matrix stabilization in the m-sample, multi-variate case. Test runs on two quite diverse operating systems for Fisher's iris data and for a more elaborate example indicate that the iterative procedure is economically feasible in terms of CPU time to convergence as well as being relatively insensitive to starting values.

1. INTRODUCTION
In order to approximate the simultaneous assumption of normality and homogeneity of variances in the univariate, m-sample case, Box and Cox (1964) proposed a univariate, data based power transformation defined by

\[ z = \begin{cases} 
\frac{(y^\lambda - 1)}{\lambda} & \text{for } \lambda \neq 0 \\
\ln y & \text{for } \lambda = 0 
\end{cases} \]

where \( \lambda \) is an additional parameter to be estimated by maximum likelihood under the assumption that \( z \) will be normally distributed.

Andrews et al. (1971) extended this transformation to the one-sample, multi-variate setting with the primary objective of either enhancing or assessing both marginal and joint normality of the underlying distribution. Their method simply consisted of applying the transformation defined by equation (1) separately to each of the \( p \) responses. Arguing that the requirement of covariance matrix stabilization as well as normalization is inherent to many familiar multivariate procedures such as classification, discriminant analysis and MANOVA, Dunn and Tubbs (1980) made the following extension to the m-sample, multivariate problem:

Suppose that a random sample \( y_{ij1}, \ldots, y_{ijn} \)
of size \( n_i \) is drawn independently from each of \( m \) \( p \)-variate populations, where \( y_{ij} = (y_{ij1}, \ldots, y_{ijp}) \) for \( i = 1, \ldots, m \) and \( j = 1, \ldots, n_i \). We shall require an estimate of the parameter set \( \lambda_1, \ldots, \lambda_p \) such that

\[ z_{ijk} \sim N(\mu_{ij}, \Sigma), \]  

where

\[ z_{ijk} = \begin{cases} 
\frac{(y_{ijk}^\lambda - 1)}{\lambda} & \text{if } \lambda \neq 0 \\
\ln y_{ijk} & \text{if } \lambda = 0 
\end{cases} \]  

for \( i = 1, \ldots, m; j = 1, \ldots, n_i; \) and \( k = 1, \ldots, p \). Note that both covariance matrix stabilization, i.e., \( z_1 = \ldots = z_p = Z \) (\( Z \) unspecified), and normalization are implicit in the likelihood function corresponding to equation (2). Details of the iteration leading to maximum likelihood estimates of \( \lambda_1, \ldots, \lambda_p \) are summarized from Dunn and Tubbs (1980) in the following section. The purpose of this report is to describe and illustrate a new SAS procedure which is an implementation of this approach. Clearly, it will include the Box and Cox case with \( p = 1 \), and the Andrews, et al. case with \( m = 1 \).

2. MAXIMUM LIKELIHOOD ESTIMATION
Following Dunn and Tubbs (1980), the joint likelihood for the data set can be re-expressed as the concentrated likelihood function.

\[ L(\lambda_1, \ldots, \lambda_p) = \exp \left( -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{p} (y_{ijk} - \mu_{ij})' \Sigma^{-1} (y_{ijk} - \mu_{ij}) \right) \theta^{k/2} \]

where

\[ \theta = \prod_{i=1}^{m} n_i^{-1} \prod_{k=1}^{p} \left( \frac{1}{\lambda_k} \right)^{\lambda_k/2} \]

Dunn and Tubbs (1980) have successfully applied the Fletcher-Powell conjugate gradient method (Cooper and Steinberg, 1970) to obtain a minimization of equation (4), which clearly is equivalent to maximizing equation (3). This procedure requires the
gradient of the response function, \( \nabla \Psi = (d_1, \ldots, d_p) \), where \( d_h \equiv d \Psi / d \lambda_h \). It can be shown that

\[
\frac{d_h}{d \lambda_h} = \left( \frac{\partial \lambda_h}{\partial \lambda_h} \right) \frac{1}{\left( \sum \lambda_h \right)^2} \int_{k=1}^{p} \frac{\partial \lambda_k}{\partial \lambda_h} \left( \frac{\partial \lambda_k}{\partial \lambda_h} \right) \left( \sum \lambda_k \right) \]

where \( \lambda = (\lambda_{ik}) \) is defined by equation (5) and \( a_{ij} \) is the cofactor of \( g_{ij} \).

It also follows that

\[
\frac{d_{kh}}{d \lambda_h} = b_1 \text{ if } k \neq h, \\
\frac{d_{kh}}{d \lambda_h} = b_2 \text{ if } k = h,
\]

where

\[
k_1 = \sum_{i=1}^{n} \left( y_{ijk} \lambda_i \right) \left( y_{ijh} \lambda_h \right) \left( y_{ijh} \lambda_h \right) \left( y_{ijh} \lambda_h \right) \\
k_2 = \sum_{i=1}^{n} \left( y_{ijn} \lambda_i \right) \left( y_{ijn} \lambda_i \right) \left( y_{ijn} \lambda_i \right) \left( y_{ijn} \lambda_i \right) \left( y_{ijn} \lambda_i \right) \left( y_{ijn} \lambda_i \right)
\]

with

\[
y_{ijk} = \sum_{j=1}^{l} \lambda_i \lambda_j y_{ijk} \\
y_{ijh} = \sum_{j=1}^{l} \lambda_i \lambda_j y_{ijh} \\
\lambda = \sum_{i=1}^{n} \lambda_i \\
\lambda = \sum_{i=1}^{n} \lambda_i
\]

This algorithm has been implemented as a new Fortran-based SAS procedure called VARSTB which is described in the following section. Here, if \( \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p) \) represents an iterated sequence of estimates of \( \lambda_k \) (\( k = 1, \ldots, p \)), then the rate of convergence is defined in terms of

\[
\text{Stress} = \max_k \left| \frac{\lambda_k}{\lambda_k} \right| - 1
\]

and the iteration is assumed to have converged if \( \text{Stress} < \varepsilon \).

3. SAS PROC VARSTB

The procedure is a multivariate normalization and covariance matrix stabilization program utilizing simultaneous power transformations on the variables. The data set must first be sorted by the \text{CLASSES} variable. By default, initial powers are set to 1, i.e., equivalent to no transformation, and the convergence criterion for the Fletcher-Powell method is preset to \( 10^{-10} \). The user has the option to set both the initial powers and the convergence criterion.

Specifications

PROC VARSTB options;

VAR variables;
CLASS variable;

Both the VAR and the CLASS are statements needed in addition to the PROC VARSTB statement:

PROC VARSTB Statement

PROC VARSTB options;

The following options can appear on the PROC statement:

DATA = SAS dataset
Names the data set to be analyzed. If omitted, the most recently created SAS dataset is used.

STRESS = n
S = n
Specifies the convergence criterion for the Fletcher-Powell method. The default value is \( 10^{-10} \).

Power = n
P = n
Specifies the starting values for initial powers. The default is 1.

VAR Statement

VAR variables;
The VAR statement lists the numeric variables to be analyzed. If omitted, all numeric variables not specified in any other statement are used.

CLASS Statement

CLASS variable;
The CLASS statement specifies the name of a variable, either character or numeric, that defines the classes to be analyzed. A CLASS statement must be present. The data set must have been sorted by the CLASS variable before PROC VARSTB is used on it.
NOTE:

If an observation has a missing value for any of the continuous variables, it is omitted from the analysis. Maximum number of sampled populations is 20 as well as the maximum number of variables. The number of variables times the number of observations must be less than 4000. The procedure also prints the initial and final pooled estimate of the common covariance matrix. (As one would get from PROC DISCRIM before and after the power transformations.)

4. EXAMPLES

Two examples were run on each of two quite diverse operating systems in order to benchmark the timing requirements of VARSTB.

System 1 consisted of CMS SAS release 82.3 (without shared segments) at the University of Arkansas, and run under VM-370 (VM-SF option) on an Amdahl 470, V6-II (8 meg) machine. System 2 consisted of TSO SAS release 82.3 at Duke Power Company, and was run under OS/VMS on an IBM 3081.

Example 1. Fisher's well-known iris data (1936) consists of 50 observations from each of three varieties of iris. Each observation consists of measurements on four flower parts, namely sepal and petal length and width.

Morrison (1967) used this data as an example in multivariate analysis of variance for which he states, "We shall, of course, assume tacitly that these populations are multivariate normal with equal covariance matrices". If a check is performed, however, Bartlett's test for equal covariance matrices gives $X^2 = 144.0$ with 20 degrees of freedom ($p < 0.0001$).

Starting with all exponents identically one and using the following SAS code,

```
DATA IRIS;
  Input Species $ Sepl Sepw Petl Petw; Cards;
  PROC Sort; By Species;
  PROC VARSTB S=.00000001;
  CLASS Species;
  VAR Sepl Sepw Petl Petw;

VARSTB converged to estimates $\hat{\lambda}_1 = -0.43053$, $\hat{\lambda}_2 = 0.51697$, $\hat{\lambda}_3 = 0.39843$, and $\hat{\lambda}_4 = 0.55464$ in four iterations with a final stress of $6.76 \times 10^{-5}$. Timing was 3.74 seconds for System 1 and 2.07 seconds for System 2.
```

The necessity for transformation is substantiated by testing $H_0: \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$, using the generalized likelihood ratio test

$$X^2 = N \ln \left[ \Psi(1, 1, 1, 1)/\Psi(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4) \right]$$

$$= 150 \ln \left( \frac{1.54974 \times 10^{12}}{1.30736 \times 10^{12}} \right) = 25.51$$

with 4 degrees of freedom ($p < 0.00005$).

Graphically, the effect of the transformation is illustrated by Figure 1 which shows 95% confidence ellipses for the first two linear discriminant functions before and after transformation using Hotelling's $T^2$ statistic.

Retesting for equal covariance matrices, Bartlett's statistic was reduced to $X^2 = 65.2$. Even though $p < 0.0001$, the reduction of Bartlett's statistic and the graphical results suggest that, despite the fact that the transformed variables are still heteroscedastic, one can more tactfully assume equality of covariance matrices after the transformation.

Example 2. Gipson (1972) reported the following ten measurements on female canine skulls:

(1) Total length of skull
(2) Upper second molar to bulla
(3) Zygomatic width
(4) Width of brain core at parieto-temporal suture
(5) Width across molars
(6) Orbit to avelolus at first molar
(7) Upper canine to second molar
(8) Crown length of upper fourth premolar
(9) Minimum width of fourth premolar
(10) Width of upper canine

Identification of the skulls consisted of assignment to one of the following six groups, where the sample sizes are also shown:

<table>
<thead>
<tr>
<th>Population</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coyote</td>
<td>20</td>
</tr>
<tr>
<td>Coyote-dog hybrid</td>
<td>19</td>
</tr>
<tr>
<td>Dog</td>
<td>13</td>
</tr>
<tr>
<td>Prairie wolf</td>
<td>20</td>
</tr>
<tr>
<td>Red wolf</td>
<td>25</td>
</tr>
<tr>
<td>Timber wolf</td>
<td>14</td>
</tr>
</tbody>
</table>

Data for the 35 canine skulls were input into SAS with the following code:

```
DATA IRIS;
  Input Species $ Sepl Sepw Petl Petw; Cards;
  PROC Sort; By Species;
  PROC VARSTB S=.00000001;
  CLASS Species;
  VAR Sepl Sepw Petl Petw;
```
Since Gipson's original manuscript indicated a sizeable collection of unknowns to be classified as well as an interest in selecting the most discriminatory variables, covariance matrix stabilization was initiated. Without transformation, Bartlett's test yielded $X^2 = 544.5$ with 275 degrees of freedom ($p < 0.0001$). From initial values identically 0.75, VARSTB converged to estimates

$$\begin{align*}
\lambda_1 &= 0.43849 \\
\lambda_2 &= 0.25610 \\
\lambda_3 &= 0.36664 \\
\lambda_4 &= 2.15277 \\
\lambda_5 &= -0.16702
\end{align*}$$

in 12 iterations with a final stress of $6.59 \times 10^{-11}$. Convergence required 18.77 seconds for System 1 and 10.98 seconds for System 2. A generalized likelihood ratio test of $H_0: \lambda_1 = \ldots = \lambda_{10} = 1$ yielded $X^2 = 11.3$ with 10 degrees of freedom ($p > 0.32$), indicating that the transformation was not particularly effective even though Bartlett's statistic was reduced to $X^2 = 449.0$. An examination of the estimated generalized variances

<table>
<thead>
<tr>
<th>$\Sigma$, i.e., after VARSTB</th>
<th>$\Sigma$</th>
<th>$\Sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coyote</td>
<td>5.53</td>
<td>4.33 $\times 10^{-25}$</td>
</tr>
<tr>
<td>Coyote-dog</td>
<td>66.02</td>
<td>1.43 $\times 10^{-23}$</td>
</tr>
<tr>
<td>Hybrid</td>
<td>331042</td>
<td>8.64 $\times 10^{-22}$</td>
</tr>
<tr>
<td>Prairie wolf</td>
<td>340.36</td>
<td>1.76 $\times 10^{-25}$</td>
</tr>
<tr>
<td>Red wolf</td>
<td>464.05</td>
<td>1.96 $\times 10^{-24}$</td>
</tr>
<tr>
<td>Timber wolf</td>
<td>478.19</td>
<td>4.78 $\times 10^{-25}$</td>
</tr>
<tr>
<td>Foothed</td>
<td>202805</td>
<td>4.71 $\times 10^{-22}$</td>
</tr>
</tbody>
</table>

indicates that even though some stabilization has occurred with respect to the coyote collection, the domestic dog skull collection still represents a relatively heterogeneous group (consistent with one of Darwin's observations).

In an effort to find an improved transformation, it was noted that Gipson (1972) had recorded an estimate of the age of each specimen based on tooth wear. An analysis of covariance indicated that variables (3), (4), and (10) showed a significant regression ($p < 0.10$) on both age and log (age). Though some lack of parallelism ($p = 0.01$) was evident for variable (4), an attempt was made to establish covariance matrix stability in a projection space orthogonal to the effects of aging as described by Burmaby (1966). Mathematically, if $y$ is a p x 1 vector of response variables which is related to a k x 1 vector $x$ of ancillary variables through a regression model $y = \alpha + \beta x$, then

$$z = [I - \beta(\beta')^{-1}\beta'] y$$

represents the p-dimensional coordinates of the projection of $y$ onto a p - k dimensional subspace $Z$ which is orthogonal to the trajectories implied by the rate constants in $\beta$. The important point is that regardless of where $y$ is on its growth trajectory, its projection onto $Z$ is unchanged. This is illustrated by Figure 2 using two trajectories $Y_1$ and $Y_2$ (representing, say, growth with increasing age) in p = 3 dimensional space. For k = 1 ancillary variable, $\beta$ is identified as a single column vector $b$. Note that the trajectories have been initiated well away from the origin at points labeled $s_1$ and $s_2$. However, instead of comparing the two trajectories at $s_1$ and $s_2$, i.e., $y$ adjusted for $x$, it may be more meaningful to compare $z_1$ with $z_2$ in the p - k = 2 dimensional subspace $Z$ which is orthogonal to $\beta$.

From the analysis of covariance, the estimated regression coefficients on age are summarized in the vector

$$\begin{align*}
\beta' &= (0, 0, 1.22514, -0.42778, 0, 0, 0, 0, 0, 0.18830) \\
\beta'' &= (0, 0, 3.69420, -1.17076, 0, 0, 0, 0, 0, 0.46026)
\end{align*}$$

Note that with the exception of variables (3), (4), and (10), all other coordinates defined by equation (6) remained unchanged. Applying Bartlett's test to coordinates in the respective 9-dimensional projection spaces, i.e. by arbitrarily dropping any one of the coordinates $z_3$, $z_4$, or $z_{10}$ generated by equation (6), yielded results shown in the first column of the following table, where 225 degrees of freedom are associated with each $X^2$:
Applying VARSTB to the first nine coordinates in the projection spaces, then retesting for equal covariance matrices, yielded the results in the second column.

Even though Bartlett's test is not strictly valid here due to disruption of the independence assumption caused by use of empirically estimated transformations, comparison of the \( X^2 \) statistics as a rough guide suggests that use of the projection space orthogonal to log (age) is more effective of the two. This is reinforced by the fact that the criterion function in VARSTB attained a minimum of \( 1.91 \times 10^{-20} \) in the space orthogonal to log (age) and a minimum of \( 3.32 \times 10^{-20} \) in the space orthogonal to age. Estimated generalized variances in the subspace orthogonal to log (age) were:

<table>
<thead>
<tr>
<th>Species</th>
<th>Estimated Generalized Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coyote</td>
<td>( 5.30 \times 10^{-26} )</td>
</tr>
<tr>
<td>Coyote-dog hybrid</td>
<td>( 1.30 \times 10^{-24} )</td>
</tr>
<tr>
<td>Dog</td>
<td>( 7.10 \times 10^{-23} )</td>
</tr>
<tr>
<td>Prairie wolf</td>
<td>( 2.31 \times 10^{-25} )</td>
</tr>
<tr>
<td>Red wolf</td>
<td>( 7.90 \times 10^{-26} )</td>
</tr>
<tr>
<td>Timber wolf</td>
<td>( 2.36 \times 10^{-23} )</td>
</tr>
<tr>
<td>Pooled</td>
<td>( 5.30 \times 10^{-26} )</td>
</tr>
</tbody>
</table>

indicating that the heterogeneity problem within the dog sample was not completely resolved by the combined transformations. However, power transformations now appear to be effective in this subspace in that a test of \( H_0: \lambda_1 = \ldots = \lambda_p = 1 \) yielded \( X^2 = 154.8 \) with 9 degrees of freedom (\( p < 0.0001 \)).

5. DISCUSSION

Use of the derived power transformations was not completely effective in stabilizing the covariance structure of either example. Yet it seems that VARSTB allowed progress to be made toward that end. Even though the option for a quadratic classifier in PROC DISCRIM would seem to negate the need for covariance matrix stabilization, still the robustness of this classifier depends on the estimates of the individual covariance matrices. Thus, the accuracy of the quadratic classifier is likely to suffer for the small sample sizes represented here, compared to a linear classifier based on a pooled estimate. Certainly the assumption of a common metric as well as normality is implicit to both PROC STEPDISC and all MANOVA tests, with no convenient options available. The VARSTB approach seems optimal in that it iterates toward both criteria.

Extension to the more general class of transformations \( (y_i + \theta_i)^{\lambda_i} - 1)/\lambda_i \) \((i = 1, \ldots, p)\) may be possible using similar methods. However, at present it seems reasonable to approximate the optimal choice of \((\theta_1, \ldots, \theta_p)\) by use of a grid search over \((\theta_1, \ldots, \theta_p)\), using the available procedure.

ACKNOWLEDGEMENTS

Partial support for this research was provided to the first author under NSF Contract No. ISP8011447 (EPSCOR). Special thanks are due to Tim Mantooth who wrote the original version of VARSTB and to Dr. Phil Gipson for making his data available to us. A copy of VARSTB and/or the canine data may be obtained by contacting the second author.

REFERENCES


Figure 1
The result of a pooled covariance matrix estimate is superimposed on results from individually estimated covariance matrices.

Before VARSTB transformations

Discriminant Function 1

Setosa Versicolor Virginica

Discriminant Function 2

After VARSTB transformations

Discriminant Function 1

Setosa Versicolor Virginica

Discriminant Function 2

Discriminant Function 3