10-2006

Allometric Extension for Multivariate Regression Models

Thaddeus Tarpey
Wright State University - Main Campus, thaddeus.tarpey@wright.edu

Christopher T. Ivey

Journal of Data Science **4**(2006), 479-495

Allometric Extension for Multivariate Regression

Thaddeus Tarpey[1] and Christopher T. Ivey[2]
[1] *Wright State University and*
[2] *California State University, Chico*

Abstract: In multivariate regression, interest lies on how the response vector depends on a set of covariates. A multivariate regression model is proposed where the covariates explain variation in the response only in the direction of the first principal component axis. This model is not only parsimonious, but it provides an easy interpretation in allometric growth studies where the first principal component of the log-transformed data corresponds to constants of allometric growth. The proposed model naturally generalizes the two–group allometric extension model to the situation where groups differ according to a set of covariates. A bootstrap test for the model is proposed and a study on plant growth in the Florida Everglades is used to illustrate the model.

Key words: Bootstrap test, canonical correlation, principal components, reduced rank regression.

1. Introduction

In multivariate regression, interesting and interpretable results are possible if the model can incorporate the dependencies between the response variables as well as the dependencies of the response variables on the predictor variables. Full rank multivariate regression models are often used in practice where each response variable is fitted separately (e.g., Johnson and Wichern, 1998, p.420). If the most of the variability in the p-dimensional fitted values from a multivariate regression lie in a lower dimensional space, then reduced-rank regression models provide a more parsimonious modeling of the data (e.g., Anderson, 1951, 1999; Reinsel and Velu, 1998) where the fitted values are constrained to lie in a lower dimensional space. In this paper, we consider a rank one multivariate regression model where the line containing the fitted values coincides with the first principal component axis of the response variables.

This work was motivated by modeling the dependency of allometric growth of plants in the Florida Everglades on soil characteristics. Allometric growth models are used to quantify the relationship between the size and shape of organisms (e.g., Hills 1982; Jolicoeur, 1963; Klingenberg, 1996; Klingenberg and Froese,

1991; Mosimann, 1970). The allometric model stipulates that parts of the plant grow at constant relative rates. Rates of allometric growth, as suggested by (Jolicoeur, 1963), are estimated using the eigenvector of the covariance matrix of the log–transformed measurements associated with the largest eigenvalue. The first eigenvector is then regarded as the allometric direction for modeling rates of growth of the different parts of the plant.

In studies of two related groups (e.g. males and females) an allometric extension model (Hills, 1982; Bartoletti, Flury and Nel, 1999) holds when one group is an extension of the other group along a common allometric axis. In Section 2 we show that the multivariate regression model with fitted values constrained to lie along the first principal component axis of the response distribution provides a natural generalization to the two-group allometric extension model to the situation where groups differ according to a set of covariates. In Section 3 we discuss the estimation of the allometric extension model for regression. A bootstrap test for the allometric extension model in regression is proposed in Section 4 and the test is applied to the plant data in Section 5. The paper is concluded in Section 6.

2. Allometric Extension for Multivariate Regression

In growth studies of two groups (e.g. males and females), one group may be an extension of the other group along a common allometric axis. The two–group allometric extension model states that the two groups share a common first principal component axis and the two groups are lined up along this common axis (Hills, 1982). The allometric extension model generalizes easily to more than two groups (e.g. closely related species). Schnute (1984) defines a closely related model based on a mixture model for bivariate data yielding a trend line which is similar to the line of allometric direction.

Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ denote the mean vectors for the two groups with corresponding covariance matrices $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$. Let $\boldsymbol{\beta}_1^{(1)}$ and $\boldsymbol{\beta}_2^{(2)}$ denote the normalized eigenvectors of each respective covariance matrix associated with the largest eigenvalue. Formally, the allometric extension model can be stated as follows (Flury, 1997, Section 8.7):

$$\boldsymbol{\beta}_1 := \boldsymbol{\beta}_1^{(1)} = \pm\boldsymbol{\beta}_2^{(2)} \tag{2.1}$$

and

$$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \delta\boldsymbol{\beta}_1 \text{ for some } \delta \in \Re. \tag{2.2}$$

The following notation allows the generalization of the two-group allometric extension model to the multivariate regression setting: let $\mathbf{X}$ denote a q–variate random vector of regressors and let $\mathbf{Y}$ denote a p–variate response vector. The multivariate regression analogue of allometric extension will be defined in terms

of the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$. It will be assumed throughout this paper that the same set of regressor variables are used to model each response variable. Form the $(p+q)\times 1$ random vector $(\mathbf{Y}', \mathbf{X}')'$ with mean $\boldsymbol{\mu} = (\boldsymbol{\mu}_y', \boldsymbol{\mu}_x')'$ and partitioned covariance matrix

$$\begin{pmatrix} \boldsymbol{\psi}_{yy} & \boldsymbol{\psi}_{yx} \\ \boldsymbol{\psi}_{xy} & \boldsymbol{\psi}_{xx} \end{pmatrix}.$$

We will assume that the conditional expectation of $\mathbf{Y}$ given $\mathbf{X}$ is linear as is the case for the multivariate normal distribution and, more generally, elliptical distributions. The multivariate linear regression model can be expressed as:

$$E[\mathbf{Y}|\mathbf{X}] = \boldsymbol{\mu}_y + \boldsymbol{\psi}_{yx}\boldsymbol{\psi}_{xx}^{-1}(\mathbf{X} - \boldsymbol{\mu}_x). \tag{2.3}$$

To generalize (2.2) to the multivariate regression setting, we require that the conditional means $E[\mathbf{Y}|\mathbf{X} = \mathbf{X}_1]$ and $E[\mathbf{Y}|\mathbf{X} = \mathbf{x}_2]$ for arbitrary values $\mathbf{x}_1$ and $\mathbf{x}_2$ of $\mathbf{X}$ line up along a common axis and this axis needs to coincide with the first principal component axis of $\mathbf{Y}$.

Definition. Let $\boldsymbol{\beta}_1$ denote the eigenvector of $\boldsymbol{\psi}_{yy}$ associated with the largest eigenvalue. The conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ follows an *allometric extension model for multivariate regression* if

$$E[\mathbf{Y}|\mathbf{X} = \mathbf{x}_1] - E[\mathbf{Y}|\mathbf{X} = \mathbf{x}_2] = \delta\boldsymbol{\beta}_1 \tag{2.4}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are arbitrary values of $\mathbf{X}$ and $\delta \in \Re$ is a constant that may depend on both $\mathbf{x}_1$ and $\mathbf{x}_2$.

The allometric extension model for multivariate regression states that the fitted values of the regression of $\mathbf{Y}$ on $\mathbf{X}$ must all line up on the first principal component axis of $\mathbf{Y}$. In other words, the variability in $\mathbf{Y}$ explained by $\mathbf{X}$ is along the first principal component axis of $\mathbf{Y}$ only.

A convenient characterization of the allometric extension model is given by the following result. Suppose $\boldsymbol{\psi}_{yy}$ and $\boldsymbol{\psi}_{xx}$ are both of full rank. Let $\mathbf{H} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p]$ denote the orthogonal matrix of eigenvectors of $\boldsymbol{\psi}_{yy}$ corresponding to the ordered (largest to smallest) eigenvalues. Partition $\mathbf{H}$ as $\mathbf{H} = [\boldsymbol{\beta}_1 : \mathbf{H}_2]$ where $\mathbf{H}_2$ is the $p \times (p-1)$ matrix of the second through pth eigenvectors. It is easy to show that the allometric extension model for multiple regression is equivalent to

$$\mathbf{H}_2'\boldsymbol{\psi}_{yx} = \mathbf{0}. \tag{2.5}$$

From (2.5) it follows that

$$\boldsymbol{\Psi}_{yx} = \boldsymbol{\beta}_1\mathbf{a}',$$

for some $\mathbf{a} \in \Re^q$. If the allometric extension model for multivariate regression holds, then the components of $\mathbf{X}$ are uncorrelated with all but the first principal component of $\mathbf{Y}$.

Note that the allometric extension model for multivariate regression is a special case of a reduced rank regression with rank equal to one with the additional constraint that the line containing the fitted values from the multivariate regression coincides with the first principal component axis of $\mathbf{Y}$. In particular, from (2.5), it follows that

$$\text{Cov}(E[\mathbf{Y}|\mathbf{X}]) = (\mathbf{a}'\boldsymbol{\Psi}_{xx}^{-1}\mathbf{a})\boldsymbol{\beta}_1\boldsymbol{\beta}_1'.$$

Thus, the covariance matrix for the conditional mean of $\mathbf{Y}$ given $\mathbf{X}$ has only one non-zero eigenvalue with corresponding eigenvector $\boldsymbol{\beta}_1$. If $\mathbf{Y}$ and $\mathbf{X}$ have a joint normal distribution, then the covariance matrix of the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ is $\boldsymbol{\psi}_{yy} - \boldsymbol{\psi}_{yx}\boldsymbol{\psi}_{xx}^{-1}\boldsymbol{\psi}_{xy}$. If the allometric extension model for multivariate regression holds as well, then the covariance matrix for the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ can be written

$$\boldsymbol{\Psi}_{yy} - (\mathbf{a}'\boldsymbol{\Psi}_{xx}^{-1}\mathbf{a})\boldsymbol{\beta}_1\boldsymbol{\beta}_1'. \tag{2.6}$$

From (2.6) it follows that $\boldsymbol{\beta}_1$ is an eigenvector for the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ which provides a generalization of (2.1) from the two-group model.

The allometric extension model for regression is illustrated in Figure 1.

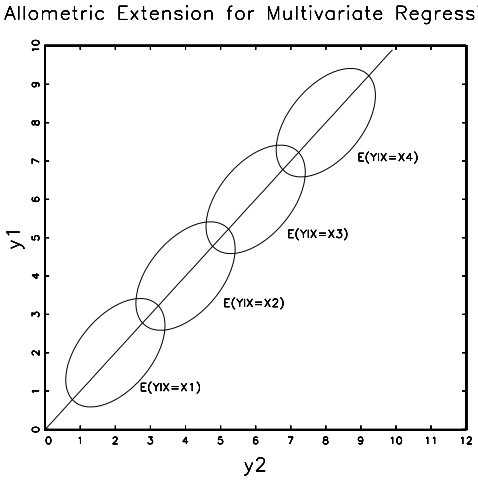

Figure 1: An illustration of the allometric extension model for multivariate regression.

3. Estimating the Allometric Regression Model

A straightforward approach to estimating the parameters of the allometric extension model is to fit a full-rank regression in the usual way and then project the fitted values onto the first principal component axis of the sample covariance matrix of $\mathbf{Y}$. Let $\mathbf{S}$ denote the sample covariance matrix and partition $\mathbf{S}$ as

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{pmatrix}.$$

and let $\hat{\boldsymbol{\beta}}_1$ denote the eigenvector of $\mathbf{S}_{yy}$ associated with the largest eigenvalue of $\mathbf{S}_{yy}$.

Consider the regression model

$$\mathbf{Y} = \mathbf{C}\mathbf{X} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ has been centered at zero. If a reduced-rank regression of rank one holds, then the coefficient matrix $\mathbf{C}$ can be factored as

$$\mathbf{C} = \mathbf{C}_1\mathbf{C}_2,$$

where $\mathbf{C}_1$ is a $p \times 1$ vector and $\mathbf{C}_2$ is a $1 \times q$ vector. $\mathbf{C}_1$ can be estimated using the eigenvector of $\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ associated with the largest eigenvalue. $\mathbf{C}_2$ is then estimated using $\hat{\mathbf{C}}_1{}'\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}$ (e.g., Reinsel and Velu, 1998). In other words, fitting the reduced-rank regression model amounts to projecting the usual least-squares coefficient matrix $\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}$ onto the space spanned by the first eigenvector of $\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$.

For the allometric extension model for regression, the coefficient matrix can be estimating by projecting the usual least-squares coefficient matrix onto the space spanned by $\hat{\boldsymbol{\beta}}_1$, the first eigenvector of $\mathbf{S}_{yy}$. Define the projection matrix $\hat{\mathbf{P}}$ as

$$\hat{\mathbf{P}} = \hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_1'.$$

Then the estimated coefficient matrix for the allometric extension model is given by

$$\hat{\mathbf{P}}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}. \tag{3.1}$$

An alternative estimation approach is to use normal theory maximum likelihood. However, there do not appear to be closed form solutions and the likelihood equations are extremely messy (details are not given here).

4. Testing for Allometric Extension

This section describes a bootstrap test that can be used to determine if an allometric extension model for multivariate regression is consistent with the data or not. The null hypothesis of the test is that the allometric extension model holds for the distribution. In an allometric study where the first principal component corresponds to the direction of allometric growth, the bootstrap test will then allow the investigator to determine if the regressor variables affect growth of the plant or animal along this allometric axis, or if the regressor variables disturb the allometric growth of the plant by altering the coefficients of allometric growth.

Using the notation from Section 2, define the $p \times p$ orthogonal matrix $\mathbf{H} = [\boldsymbol{\beta}_1, \mathbf{H}_2]$ whose columns are the eigenvectors of $\boldsymbol{\Psi}_{yy}$. If $\mathbf{Z}_2$ denotes the second through pth principal components of $\mathbf{Y}$, then $\mathbf{Z}_2 = \mathbf{H}_2'\mathbf{Y}$ (assuming $\mathbf{Y}$ has been centered at zero). From (2.5), the hypothesis of allometric extension for multivariate regression is equivalent to the components of $\mathbf{Z}_2$ being uncorrelated to the components of $\mathbf{X}$:

$$H_0 : \mathbf{H}_2'\boldsymbol{\Psi}_{yx} = \mathbf{0}.$$

If $\mathbf{H}_2$ were known, the normal theory likelihood ratio test (e.g., Anderson's book, p.394) is equivalent to

$$\frac{|\mathbf{H}_2'\mathbf{S}_{yy}\mathbf{H}_2||\mathbf{S}_{xx}|}{|\begin{pmatrix} \mathbf{H}_2'\mathbf{S}_{yy}\mathbf{H}_2 & \mathbf{H}_2'\mathbf{S}_{yx} \\ \mathbf{S}_{xy}\mathbf{H}_2 & \mathbf{S}_{xx} \end{pmatrix}|}, \tag{4.1}$$

and the asymptotic null distribution of n times the natural logarithm of (4.1) would be chi-square with degrees of freedom equal to $(p-1)q$. In practice the eigenvectors of $\boldsymbol{\Psi}_{yy}$ must be estimated. Let $\hat{\mathbf{H}}_2$ denote the matrix of eigenvectors of $\mathbf{S}_{yy}$ associated with the second through pth ordered eigenvalues. Multiplying the natural logarithm of (4.1) by n after replacing $\mathbf{H}_2$ with $\hat{\mathbf{H}}_2$ leads to the following test statistic (after simplifying using standard results for partitioned matrices):

$$\hat{\lambda} = n \ln\Big[\frac{|\hat{\mathbf{H}}_2'\mathbf{S}_{yy}\hat{\mathbf{H}}_2|}{|\hat{\mathbf{H}}_2'(\mathbf{S}_{yy} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy})\hat{\mathbf{H}}_2|}\Big]. \tag{4.2}$$

The finite sample and asymptotic distributions of (4.2) are unknown due to the fact that $\mathbf{H}_2$ needs to be estimated.

A straightforward alternative to the parametric approach is to use a bootstrap test which does not require the normality assumption. The basic idea is to compute (4.2) using the raw data. Next, transform the raw data so that the null hypothesis is approximately valid (Hall and Wilson, 1991). When testing a hypothesis about the mean, this transformation involves centering the raw data at the hypothesized value of the mean (see e.g., Westfall and Young, 1993). In the

current setting, the null hypothesis concerns the covariance structure. In order to transform the data so that the null hypothesis of allometric extension holds, the data needs to be re-scaled (instead of centered) – details are below. Finally, the re-scaled data is resampled and (4.2) is computed for the bootstrap samples. An (approximate) p-value for the bootstrap test is computed as the the proportion of the bootstraped test statistics that exceed the test statistic computed from the raw data.

In order to describe how to re-scale the data so that the null hypothesis of allometric extension holds, let $\mathbf{BDB}'$ denote the spectral decomposition of $\boldsymbol{\psi}$, the covariance matrix of $(\mathbf{Y}', \mathbf{X}')'$. The matrix $\mathbf{D}$ is a diagonal matrix of eigenvalues of $\boldsymbol{\psi}$ and $\mathbf{B}$ is the orthogonal matrix of eigenvectors of $\boldsymbol{\psi}$. Let $\mathbf{P} = \boldsymbol{\beta}_1\boldsymbol{\beta}_1'$ denote the projection matrix for the eigenvector $\boldsymbol{\beta}_1$ associated with the largest eigenvalue of $\boldsymbol{\psi}_{yy}$. Then from (3.1), the covariance matrix

$$\boldsymbol{\psi}_0 = \begin{pmatrix} \boldsymbol{\psi}_{yy} & \mathbf{P}\Psi_{yx} \\ \boldsymbol{\psi}_{xy}\mathbf{P} & \boldsymbol{\psi}_{xx} \end{pmatrix}$$

is a covariance matrix that satisfies the allometric extension model. Let $\boldsymbol{\psi}_0 = \mathbf{B}_0\mathbf{D}_0\mathbf{B}_0'$ denote the spectral decomposition of $\boldsymbol{\psi}_0$. Then we can transform the original variables $(\mathbf{Y}', \mathbf{X}')'$ to have a covariance matrix equal to $\boldsymbol{\psi}_0$ by noting the following:

$$\begin{aligned}
&\mathrm{Cov}[(\mathbf{B}_0'\mathbf{D}_0^{1/2})(\mathbf{D}^{-1/2}\mathbf{B}')\begin{pmatrix}\mathbf{Y}\\ \mathbf{X}\end{pmatrix}] \\
&= (\mathbf{B}_0'\mathbf{D}_0^{1/2})(\mathbf{D}^{-1/2}\mathbf{B}')\mathrm{Cov}[\begin{pmatrix}\mathbf{Y}\\ \mathbf{X}\end{pmatrix}](\mathbf{B}\mathbf{D}^{-1/2})(\mathbf{D}_0^{1/2}\mathbf{B}_0) \\
&= (\mathbf{B}_0'\mathbf{D}_0^{1/2})(\mathbf{D}^{-1/2}\mathbf{B}')\boldsymbol{\psi}(\mathbf{B}\mathbf{D}^{-1/2})(\mathbf{D}_0^{1/2}\mathbf{B}_0) \\
&= (\mathbf{B}_0'\mathbf{D}_0^{1/2})(\mathbf{D}^{-1/2}\mathbf{B}')(\mathbf{B}\mathbf{D}\mathbf{B}')(\mathbf{B}'\mathbf{D}^{-1/2})(\mathbf{D}_0^{1/2}\mathbf{B}_0) \\
&= (\mathbf{B}_0'\mathbf{D}_0^{1/2})\mathbf{I}(\mathbf{D}_0^{1/2}\mathbf{B}_0) = \boldsymbol{\psi}_0.
\end{aligned}$$

Therefore, left multiplying the $(p+q)$-dimensional vector $(\mathbf{Y}', \mathbf{X}')'$ by $(\mathbf{B}_0'\mathbf{D}_0^{1/2})$ $(\mathbf{D}^{-1/2}\mathbf{B}')$ yields a distribution satisfying the allometric extension model.

To perform the bootstrap testing, one simply re-scales the raw data using the sample counterparts of the above matrices and then resamples the data. In particular, let $(\mathbf{y}_i', \mathbf{x}_i')', i = 1, \ldots, n$, denote the raw data and let $\mathbf{S}$ denote the sample covariance matrix. Let $\hat{\mathbf{B}}\hat{\mathbf{D}}\hat{\mathbf{B}}'$ denote the spectral decomposition of $\mathbf{S}$ where $\hat{\mathbf{B}}$ is an orthogonal $(p+q)\times(p+q)$ matrix and $\hat{\mathbf{D}}$ is a diagonal matrix of ordered eigenvalues of $\mathbf{S}$. Let

$$\mathbf{S}_0 = \begin{pmatrix} \mathbf{S}_{yy} & \hat{\mathbf{P}}\mathbf{S}_{yx} \\ \mathbf{S}_{xy}\hat{\mathbf{P}} & \mathbf{S}_{xx} \end{pmatrix}.$$

Denote the spectral decomposition of $\mathbf{S}_0$ by

$$\mathbf{S}_0 = \hat{\mathbf{B}}_0 \hat{\mathbf{D}}_0 \hat{\mathbf{B}}_0'.$$

The bootstrap testing procedure is given in the following steps:

Bootstrap Testing Procedure

1. Compute $\hat{\lambda}$, the test statistic using (4.2).
2. Re-scale the raw data so that the null hypothesis of allometric extension (approximately) holds by computing:

$$\begin{pmatrix} \mathbf{y}_{i0} \\ \mathbf{x}_{i0} \end{pmatrix} = \hat{\mathbf{B}}_0' \hat{\mathbf{D}}_0^{1/2} \hat{\mathbf{D}}^{-1/2} \hat{\mathbf{B}}' \begin{pmatrix} \mathbf{y}_i \\ \mathbf{x}_i \end{pmatrix},$$

   for $i = 1, \ldots, n$.
3. Obtain a bootstrap sample (sampling with replacement) from the re-scaled data $(\mathbf{y}_{i0}', \mathbf{x}_{i0}')'$ and compute the test statistic (4.2), call it $\hat{\lambda}^*$.
4. Repeat step (3) N times for some large value of N (say $N = 10,000$) generating bootstrap test statistics $\lambda_j^*,\ j = 1, \ldots, N$.
5. A bootstrap p-value is then computed by the proportion of λ_j^* that exceed λ.

This bootstrap procedure is quite easy to implement using standard statistical software because it only requires the computation of the eigenvectors and eigenvalues of the sample covariance matrix. Code for the bootstrap test is available using the R-software (Ihaka and Gentelman, 1996) from the authors upon request.

A simulation study was conducted to evaluate the performance of the bootstrap testing procedure. Multivariate normal data was simulated using the Gauss software for a wide variety of parameterizations and sample sizes. The power of the bootstrap test was estimated based on 1000 simulated data sets. For each simulated data set, the bootstrap test described above was performed for $N = 1000$ bootstrap samples. Ideally, the p-value distribution when simulating under the null hypothesis of an allometric extension model should be uniform $(0, 1)$. The actual p-value distribution under the null hypothesis is approximately uniform with deviations from uniformity in the direction of a conservative test (i.e., the probability of a type I error is smaller than specified using the usual significance levels). Under the null hypothesis, the p-value distribution coincided more closely

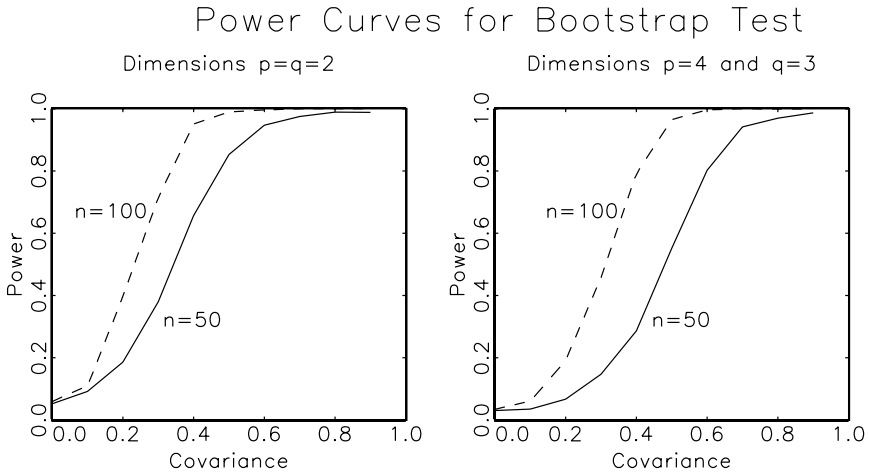

Figure 2: Power curves for bootstrap test. Simulation results for the bootstrap test of the allometric extension model.

with a uniform $(0, 1)$ distribution as the sample size increased or as the dimensionality of the data decreased, or as the correlation between the first principal component of the response and the regressors increased. The power of the bootstrap test was evaluated simulating data from distributions where the allometric extension model does not hold. The results of the power study are summarized by the power curves shown in Figure 2. In all cases, the power was computed using a significance level 0.05. The left panel of Figure 2 shows the results of simulating a bivariate response ($p = 2$) with two regressors ($q = 2$). The right panel of Figure 2 is based on $p = 4$ response variables and $q = 3$ regressors. In each panel, power curves for sample sizes of $n = 50$ and 100 are shown. The first principal component of the response had a variance of 4 in the left panel and a variance of 6 in the second panel. The other principal component(s) of the response(s) had a variances of 2 and 3 in the left and right panels respectively. The regressors had unit variance in all cases and the covariance of the first principal component with the regressors was 0.5 in all cases. The power curves in Figure 2 were generated by increasing the covariance from 0 (where the null hypothesis of allometric extension holds) to 0.9 in increments of 0.1 between the regressors and the 2nd through pth principal components of the response. As this covariance increases, the model moves further away from the allometric extension model and therefore the power of the test increases which is evident from the curves in Figure 2. The bootstrap test is slightly less powerful in the higher dimensional simulation (right panel of Figure 2) compared to the lower dimensional simulation. These power curves show that the bootstrap test is quite powerful even when the correlation

between the regressors and the 2nd–pth principal components of the response are moderate. For example, in the left panel of Figure 2 with $n = 50$, the power of the test is 0.947 when the 2nd principal component has a correlation of 0.42 with each of the regressors whereas, in the right panel, the power is 0.942 when the correlation between the 2nd–4th principal components of the response have a correlation of 0.4 with the regressors. Further details on the simulation results are available from the authors.

As mentioned in the previous section, the maximum likelihood estimators of the covariance matrix in the allometric extension model for multivariate regression are intractable and the likelihood equations are very messy. Therefore, a likelihood ratio test for the allometric extension model is not pursued.

5. Application

The ecological balance of nutrient-limited areas such as the Florida Everglades can be jeopardized by anthropogenic nutrient enrichment. In order to detect and monitor the presence of nutrient contamination in the Everglades, a study was conducted to examine variation in plant morphology that results from soil characteristics. In particular, the negative effects of phosphorus enrichment in the Everglades are a concern. This studied examined *Sagittaria lancifolia* and *Cladium jamaicense* (or "sawgrass") plants which are common in the Florida Everglades. Survey sites were randomly located throughout the Florida Everglades and a sample of plants and soil measurements were obtained at each site in work that involved the second author. Data were collected during the wet and dry seasons.

Data on $n_1 = 287$ *Sagittaria lancifolia* plants were collected during the wet season and $n_2 = 298$ plants during the dry season. For the sawgrass, there were 517 and 615 observations in the wet and dry seasons respectively. The response variables were the log-transformed length and width of the leaves:

$$\begin{aligned} y_1 &= \log(\text{Lamina length}) \\ y_2 &= \log(\text{Lamina width}). \end{aligned}$$

For the sawgrass, an addition response variable was measured:

$$y_3 = \log(\text{rhizome diameter}).$$

Using the previous notation, $p = 2$ for the *Sagittaria lancifolia* plants. The $q = 2$ soil variables are

$$x_1 = \text{Total phosphorus in soil (units are micrograms/gram)}$$

x_2 = Percent ash-free dry weight of soil.

Ash-free dry weight is related to the amount of organic matter in the soil and is useful for distinguishing mineral soils, called marl, and peat soils that are higher in organic matter. The ash-free dry weight relates to the accessibility of nutrient availability to plants. Variables x_1 and x_2 were standardized since they are measured on different scales. Also, the response variables $\mathbf{y}$ were all centered at zero. Phosphorus is a nutrient that has a fertilizing effect on plants leading to higher growth. Similarly, a regression of the leaf dimensions on ash-free weight also shows a positive relationship.

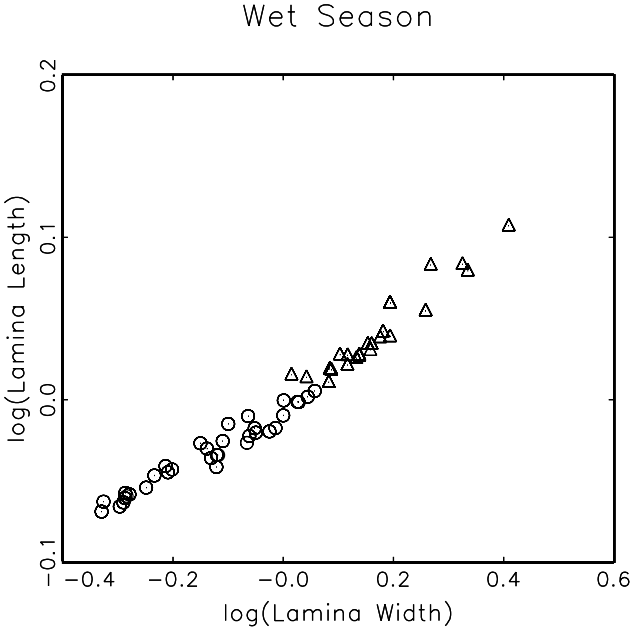

Figure 3: Fitted values of logarithms of leaf length and width from a regression on soil ash-free weight and soil total phosphorus of *Sagittaria lancifolia* plants. Plants were divided into two–groups: above average values of phosphorus (triangles) and below average values of phosphorus (circles).

We present first the results for the *Sagittaria lancifolia* plants. Figure 3 shows fitted values of log(Lamina length) and log(Lamina width) from a full-rank regression on x_1 and x_2 for the wet season data. The fitted values tend to cluster along a line indicating that a reduced-rank regression model with rank 1 may hold for the data. The plants were divided into two–groups based on large (above average) and small (below average) phosphorus values. The large

phosphorus group is plotted with a triangle and the small phosphorus group is plotted with a circle. The two different plotting symbols indicate that the large group appears to be an allometric extension of the small group (Bartoletti *et al.*, 1999). However, the large and small grouping is artificial. Because phosphorus and the ash-free dry weight are continuous variables, the allometric extension model for regression needs to be tested.

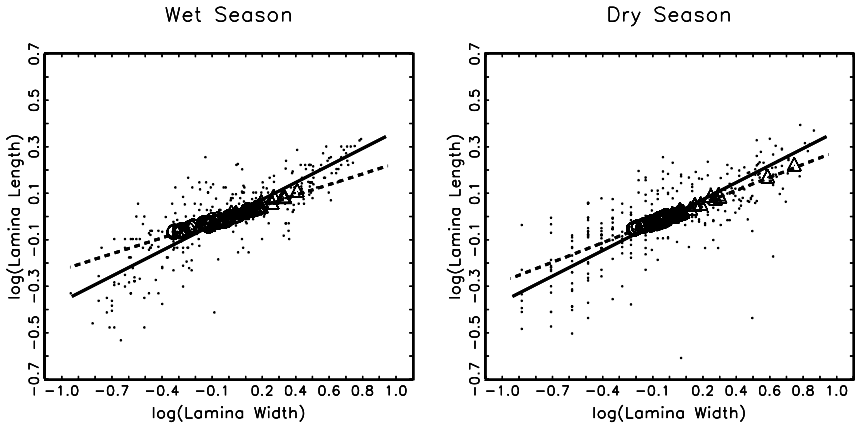

Figure 4: Scatterplots of the log-leaf data for the wet and dry seasons for the *Sagittaria lancifolia* plant data. The solid lines in the left and right frames are the estimated first principal component axes for the raw data and the dashed lines are the first principal component axes for the fitted values.

The bootstrap test was run on the data for the wet and dry seasons. 50,000 bootstrap samples were obtained for each test using the bootstrap testing procedure outlined in Section 4. For the wet season, only one of the 50,000 bootstrap samples yielded a test statistic bigger than the observed test statistic for the raw data giving a p-value of $p = 0.00002$. Thus, for the wet season, the allometric extension model for regression is clearly rejected. However, for the dry season, the bootstrap p-value is $p = 0.25413$ indicating that the allometric extension model for regression is consistent with the dry season data.

Figure 4 provides an illustration of the results. Each frame of Figure 4 shows the raw log-leaf data for the wet (left frame) and dry (right frame) seasons. The solid line in each frame is the estimated first principal component axis for the raw plant data. The dashed line is the estimated first principal component axis for the fitted values. Also plotted are the fitted values (indicated with large circles and triangles corresponding to small and large levels of phosphorus). Under the allometric extension model, the dashed and solid principal component axes would exactly coincide. The discrepancy between the solid and dashed line is greater for the wet season than for the dry season which gives some indication why the

allometric extension model is less plausible for the wet season than for the dry season as evidenced by the bootstrap test.

The bootstrap test for allometric extension was also run for the sawgrass data which had three dependent variables (logarithms of leaf length, width and rhizome). For the wet season, the bootstrap test yielded an estimated p-value of $p = 0.10314$; for the dry season, the p-value is $p = 0.00000$. Therefore, the allometric extension model for multivariate regression is consistent for the wet season data but not for the dry season data, which is the reverse of the results for the *Sagittaria lancifolia* plants.

The coefficients for the first principal component z_1 for the wet season sawgrass data are $0.516, 0.555, 0.653$ which gives estimates of allometric growth rates. The second principal component z_2 can be regarded as a comparison of length and width with rhizome:

$$z_2 = -0.723\log(\text{length}) - 0.126\log(\text{width}) + 0.679\log(\text{rhizome}).$$

The third principal component z_3 compares width with length and rhizome:

$$z_3 = 0.459\log(\text{length}) - 0.823\log(\text{width}) + 0.336\log(\text{rhizome}).$$

Because the allometric extension model is consistent with the wet season sawgrass data, it follows that phosphorus and ash-free dry weight explain variability in the growth of the plants along the z_1 allometric axis only.

Table 1: Percentage of leaf variability explained by regression models.

Plant	Model	Wet Season	Dry Season
Sagittaria lancifolia	Full Rank Regression	17.29%	15.26%
	Reduced Rank (rank = 1)	17.26%	15.24%
	Allometric Extension	16.99%	15.21%
Cladium jamaicense (Sawgrass)	Full Rank Regression	36.37%	32.36%
	Reduced Rank (rank = 1)	36.34%	32.19%
	Allometric Extension	36.32%	31.80%

The percentage of variability in the responses accounted for by the regressors in a multivariate regression can be computed as

$$R^2 = \frac{\text{tr}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy})}{\text{tr}(\mathbf{S}_{yy})} \times 100\%.$$

Table 1 gives the percentage of variability in the log-leaf measurements explained by the full unrestricted regression model, a reduced rank regression model with

rank equal to one and for the allometric extension model for both plants and both season. The allometric extension model is the most restrictive of the three models and therefore must have the lowest R^2 value. However, from Table 1, the proportion of variability explained for the allometric extension model is essentially the same as the full unrestricted model and the reduced rank regression model, particularly in the cases where the bootstrap test indicated that the allometric extension model holds (dry season for *Sagittaria lancifolia* and wet season for sawgrass).

Below is the estimated covariance matrix for the dry season *Sagittaria lancifolia* plant data of $(\mathbf{Z}', \mathbf{X}')'$ where $\mathbf{Z} = (Z_1, Z_2)'$ are the two principal components of the log-leaf measurements and $\mathbf{X} = (X_1, X_2)'$ corresponds to the standardized ash-free weight and phosphorus measurements:

$$\begin{pmatrix} 0.213 & 0.000 & 0.186 & 0.091 \\ 0.000 & 0.013 & -0.008 & -0.010 \\ 0.186 & -0.008 & 1.000 & 0.476 \\ 0.091 & -0.010 & 0.476 & 1.000 \end{pmatrix}.$$

The pairwise covariances between x_1 and x_2 with the second principal component of the leaf measurements appear quite small as would be expected for a model consistent with the allometric extension model.

It is interesting to also perform a likelihood ratio test for a reduced rank regression (rrr) model of rank $r = 1$. The likelihood ratio test for testing if the rank of the coefficient matrix is of rank r is

$$\hat{\lambda}_{rrr} = \Big(\frac{|\mathbf{S}(\text{residual})|}{|\mathbf{S}_{yy}|} \Big)^{-n/2}$$

where $\mathbf{S}$(residuals) is the residual sum of squares matrix from the reduced rank model (Reinsel and Velu, 1998, p.50). Asymptotically, $-2\log(\hat{\lambda}_{rrr})$ follows a chi–squared distribution on $(p-r)(q-r)$ degrees of freedom. For the *Sagittaria lancifolia* plant data we have $(p-1)(q-1) = 1$ degree of freedom. The observed test statistics for the wet and dry seasons are $\chi^2 = 0.834$ and 0.489 respectively indicating that a reduced rank regression with rank equal to one is consistent for both the wet and dry season leaf data. As we have seen however, the more parsimonious model of allometric extension is consistent for the dry season data. Even though we rejected the allometric extension model for the wet season, the less restrictive reduced rank regression of rank equal to one is not rejected. Also, Figure 4 shows that the allometric extension model for the wet season, although rejected by the likelihood ratio test, nonetheless provides a fairly good approximation to the observed data.

6. Discussion

In the plant examples of Section 5, departures from the allometric extension model may be of interest in identifying the effects of soil nutrient over-enrichment. For instance, if nutrients in the soil and/or water alter the usual growth rates of the plant, then the allometric extension model for multivariate regression will fail to hold. Since the allometric extension model appears consistent with the data from the dry season for the *Sagittaria lancifolia* plants and since phosphorus has a fertilizing effect, we would expect to see larger leaves following the same growth pattern in areas suffering from phosphorus contamination compared to uncontaminated sites. Strong departures from the allometric extension in this context may be evidence that the phosphorus is causing differences in the both the size and shape of the leaves, possibly a deformation.

The model considered here is quite simple and many generalizations can be pursued. A more restrictive one-dimensional model is that of isometric growth where the fitted values are constrained to lie on the line determined by the unit vector $(1, 1, \ldots, 1)'/\sqrt{p}$. Another interesting variation of the problem is to consider models where the allometric extension model holds for a subset of the response variables in a multivariate regression (Ivey *et al.*, 2004). For higher dimensional data, one could easily modify the bootstrap test of Section 4 to test if the fitted values lie in a lower dimensional subspace spanned by the first few eigenvectors of the covariance matrix of the response distribution. In fact, one can regard these models as special cases of common principal component models (Flury, 1988) where some or all of the eigenvectors of $\boldsymbol{\Psi}_{yx}\boldsymbol{\Psi}_{xx}^{-1}\boldsymbol{\Psi}_{xy}$ are constrained to coincide with some of the eigenvectors of $\boldsymbol{\Psi}_{yy}$. Common principal component models for regression then would postulate that the covariance matrix for $E[\mathbf{Y}|\mathbf{X}]$ could be expressed as

$$\boldsymbol{\Psi}_{yx}\boldsymbol{\Psi}_{xx}^{-1}\boldsymbol{\Psi}_{xy} = \sum_{j=1}^{r} (\mathbf{a}_j'\boldsymbol{\Psi}_{xx}^{-1}\mathbf{a}_j)\boldsymbol{\beta}_j\boldsymbol{\beta}_j',$$

where $\boldsymbol{\beta}_j$ are eigenvectors of $\boldsymbol{\Psi}_{yy}$, the $\mathbf{a}_j \in \Re^q$, and r is the rank of the model.

It would also be of interest to consider nonlinear generalizations of the reduced rank models. Hastie and Stuetzle (1989) introduced principal curves as a nonlinear generalization of principal component axes. If there is nonlinear structure in the response distribution, then there may very well be nonlinear structure in the fitted values as well after regressing the response on regressors. If the fitted values fall on a curve, then one would have a nonlinear reduced rank regression model of rank equal to one. If this curve coincides with the first principal curve of the response distribution, then a nonlinear allometric extension model for regression holds for the data.

References

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics* **5**, 248-264.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley.

Anderson, T. W. (1999). Asymptotic theory for canonical correlation analysis *J. of Multivariate Analysis* **70**, 1-29.

Bartoletti, S., Flury, B. and Nel, D.G. (1999). Allometric Extension, *Biometrika* **55**, 1210-1214.

Flury, B. (1988). *Common Principal Components and Related Multivariate Models*. Wiley.

Flury, B. (1997). *A First Course in Multivariate Statistics*. Springer.

Hall, P. and Wilson, S. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757-762.

Hastie, T. and Stuetzle, W. (1989). Principal Curves. *J. Amer. Statist. Asso.* **84**, 502-516.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics, *J. of Computational and Graphical Statistics* **5**, 299-314.

Ivey, C. T. and Tarpey, T. Jones, R. D. and Richards, J. H. (2006). Contrasting morphological responses of two Everglades macrophytes to soil characteristics. Submitted for publication.

Johnson, R. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis, 4-th ed.* Prentice Hall.

Jolicoeur, P. (1963). The multivariate generalization of the allometry equation. *Biometrics* **19**, 497-499.

Klingenberg, C. P. (1996). Multivariate allometry. In *Advances in Morphometrics* (Edited by M. Corti, A. Loy, J. P. Naylor and D. E. Slice), 23-49. Plenum.

Klingenberg, C. P. and Froese, R. (1991). A multivariate comparison of allometric growth patterns. *J. Systematic Zoology* **40**, 410-419.

Mosimann, J, E, (1970). Size allometry: Size and shape variables with characterizations of the lognormal and generalized gamma distributions. *J. Amer. Statist. Asso.* **65**, 930-945.

Reinsel, G. and Velu, R. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer.

Schnute, J. (1984). Linear mixtures: A new approach to bivariate trend lines *J. of the American Statistical Association* **79**, 1-8.

Westfall, Westfall, P. and Young, S., 1993, *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustement.* Wiley

Received April 19, 2005; accepted June 20, 2005.

Thaddeus Tarpey
Wright State University
Department of Mathematics and Statistics
120 MM Building
Dayton, OH 45435, USA thaddeus.tarpey@wright.edu

Christopher T. Ivey
DEpartment of Biological Sciences
California State University
Chico, CA 95929-0515
ctivey@csuchoco.edu