Exploratory Factor Analysis in Morphometry

A. M. C. Machado[1,3], J. C. Gee[2], and M. F. M. Campos[1]

[1] DCC, Federal University of Minas Gerais, Belo Horizonte, Brazil
{alexei,mario}@dcc.ufmg.br
[2] Department of Radiology, University of Pennsylvania, Philadelphia, USA
gee@rad.upenn.edu
[3] DCC, Pontifical Catholic University of Minas Gerais, Belo Horizonte, Brazil

Abstract. In this paper, we present an exploratory factor analytic approach to morphometry in which a high-dimensional set of shape-related variables is examined with the purpose of finding clusters with strong correlation. This clustering can potentially identify regions that have anatomic significance and thus lend insight to the morphometric investigation. The analysis is based on information about size difference between the differential volume about points in a template image and their corresponding volumes in a subject image, where the correspondence is established by non-rigid registration. The Jacobian determinant field of the registration transformation is modeled by a reduced set of factors, whose cardinality is determined by an algorithm that iteratively eliminates factors that are not informative. The results show the method's ability to identify gender-related morphological differences without supervision.

1 Introduction

This work presents a novel method for exploring the relationship among morphometric variables and the possible anatomic significance of these relationships. Our approach is based on the analysis of high-dimensional sets of vector variables obtained from non-rigidly deforming a template image so as to align its anatomy with the subject anatomy of a group, depicted in MRI studies. The resultant individualized templates provide an anatomic labeling of the subject data. In addition, information about regional shape can be extracted from the alignment transformations. This information—in the form of differential size differences among subject anatomies—is statistically analyzed to yield a reduced set of common *factors* that correspond to new variables with possible anatomic significance. In this way, the methodology more naturally facilitates hypothesis-driven explorations of regional differences in the data and lends deeper insight to the morphometric investigation.

An important and commonly used method for shape representation is the principal component analysis (PCA) of variance [10]. In this approach, a new set of variables is determined as linear combinations of the original variables, in such a way that they account for most part of the variance presented in the sample.

C. Taylor, A. Colchester (Eds.): MICCAI'99, LNCS 1679, pp. 378–385, 1999.
© Springer-Verlag Berlin Heidelberg 1999

Marcus [9] and Cootes *et al.*[2] showed how PCA could be used to construct models based on manually chosen landmarks. The analysis was focused on the variable domain, since the number of subjects in the sample was usually larger than the number of variables being measured. Generalizing the use of PCA to high-dimensional variables, Le Briquer and Gee[7] applied the method directly to the mappings, registering an atlas to the subject anatomies.

Although PCA is efficient in summarizing variability in a dataset, the principal components are generally difficult to interpret. Factor analysis differs from PCA in that it provides a basis transformation from the variables into a factor domain that preserves the correlation among original variables. Factor analysis accounts for the covariance among these variables instead of representing principal modes of variance. In this sense, the information about the dependency among variables is preserved.

The use of factor analysis in morphometry has been examined by Marcus [9] and Reyment and Jöreskog [11] in the context of analyzing general scalar variables, presenting a wide discussion on the factor analysis of landmark data. Scalar features such as landmark distances and curvatures were considered in the analysis of ostracod species, where the displacements between landmarks were modeled as deformations of a thin-plate spline.

In this paper, we extend the exploratory factor analytic approach to high-resolution MRI morphometry, where the set of measurement variables are of very high dimension. The method is applied to the study of the corpus callosum and compared to previous published results [3,8].

2 Methods

The displacement fields representing the spatial warps between the reference and the subjects' images are obtained from a Bayesian generalization of the Bajcsy and Kovačič elastic matching technique [1] and its finite element implementation [5]. The process comprises a global and local registration which account, respectively, for rigid transformations and continuous one-to-one mapping according to the elasticity theory principles.

2.1 Atlas-Based MRI Morphometry

The amount of scaling applied to an infinitesimal area around a point $\mathbf{x}$ in the template, when it is deformed to match a subject, can be evaluated by computing the Jacobian determinant of the mapping function between corresponding points. The pointwise Jacobian determinants for two populations can be compared by computing an *effect size*:

$$e(x) = [\mu_1(J(x)) - \mu_2(J(x))]/\sqrt{\sigma^2(J(x))}, \tag{1}$$

where $\mu_p(J(x))$ is the mean pointwise Jacobian determinant of a population p and $\sigma^2(J(x))$ is the unbiased variance for populations 1 and 2 combined. These comparisons show shape differences between populations, reflected in the amount of compression or dilatation of a region-of-interest in the reference image.

2.2 Factor Analysis

The purpose of factor analysis is to reduce the dimensionality of a problem by exploring the correlation among its variables. A set of p original variables, $\mathbf{y}$, is represented as linear combinations of m new variables called factors:

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{y} = (y_1, \ldots, y_p)^T$ are the variables observed from a sample of a population with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ and covariance matrix $\boldsymbol{\Sigma}$; $\mathbf{f} = (f_1, \ldots, f_m)^T$ are the factors; $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)^T$ are the error terms which account for the portion of $\mathbf{y}$ that is not common to the other variables; and $\boldsymbol{\Lambda} = ((\lambda_{11}, \ldots, \lambda_{1m}), \ldots, (\lambda_{p1}, \ldots, \lambda_{pm}))^T$ is the loading matrix. The coefficients λ_{ij}, called *loadings*, express the covariance between variable y_i and factor f_j.

The model and purpose of factor analysis lead to some assumptions regarding the behavior of $\mathbf{f}$ and $\boldsymbol{\epsilon}$. Since the expected value $\mathcal{E}(\mathbf{y} - \boldsymbol{\mu})$ is the null vector, $\mathcal{E}(\mathbf{f})$ and $\mathcal{E}(\boldsymbol{\epsilon})$ must also be $\mathbf{0}$. In order for the factors to account for all the correlation among the variables $\mathbf{y}$, the covariance among error terms and factors must be 0 and the covariances among error terms are represented by the diagonal matrix $\boldsymbol{\Psi} = diag(\psi_1, \ldots, \psi_p)$. The diagonal entries of $\boldsymbol{\Psi}$, ψ_i, are called *specific variances*. It is also assumed that the covariance matrix for factors is the identity matrix, so that the variance of y_i can be expressed as

$$var(y_i) = \textstyle\sum_{j=1}^{m} \lambda_{ij}^2 + \psi_i. \tag{3}$$

The summation in (3) is called the *communality* or *common variance*. Based on the independence between $\boldsymbol{\Lambda}\mathbf{f}$ and $\boldsymbol{\epsilon}$, and considering $cov(\mathbf{f}) = \mathbf{I}\mathbf{i}$, the relationship between $\boldsymbol{\Sigma}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ can be written as

$$\boldsymbol{\Sigma} = cov(\boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\epsilon}) = cov(\boldsymbol{\Lambda}\mathbf{f}) + cov(\boldsymbol{\epsilon}) = \boldsymbol{\Lambda} cov(\mathbf{f}) \boldsymbol{\Lambda}^T + \boldsymbol{\Psi} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}. \tag{4}$$

An interesting characteristic of the loading matrix $\boldsymbol{\Lambda}$ is that it can be multiplied by an orthogonal matrix and still be able to represent the covariance among factors and original variables. Since any orthogonal matrix $\mathbf{Q}$ multiplied by its transpose leads to the identity matrix, the basic model for factor analysis in (2) can be written as

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{Q}\mathbf{Q}^T\mathbf{f} + \boldsymbol{\epsilon} = \boldsymbol{\mu} + \boldsymbol{\Lambda}^*\mathbf{f}^* + \boldsymbol{\epsilon},$$

where $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{Q}$ and $\mathbf{f}^* = \mathbf{Q}^T\mathbf{f}$. If $\boldsymbol{\Sigma}$ in (4) is expressed in terms of $\boldsymbol{\Lambda}^*$, we have

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*T} + \boldsymbol{\Psi} = \boldsymbol{\Lambda}\mathbf{Q}(\boldsymbol{\Lambda}\mathbf{Q})^T + \boldsymbol{\Psi} = \boldsymbol{\Lambda}\mathbf{Q}\mathbf{Q}^T\boldsymbol{\Lambda}^T + \boldsymbol{\Psi} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}, \tag{5}$$

showing that the original loading matrix $\boldsymbol{\Lambda}$ and the rotated matrix $\boldsymbol{\Lambda}^*$ yield the same representation of $\boldsymbol{\Sigma}$. The rotation of loadings plays an important role in factor interpretation, as it is possible to obtain a rotated matrix that assigns a few high loadings for each variable, keeping the other loadings small. If such matrix is obtained, each variable will be related to a single factor (or at least to a few ones), which can potentially be given a morphological interpretation.

2.3 Analysis of Effect Size

The effect size expression shown in (1) can be formulated in the factor domain, where for each point x_i, $y_i = J(x_i)$. In this way, the comparison between two populations can be performed based exclusively on the reduced set of factor values (*scores*) which completely represent each subject in the new basis. From the factor analysis model in (2) and the definition of variance in (3), we have that

$$\begin{aligned} e(x_i) &= [\mu_1(y_i) - \mu_2(y_i)]/\sqrt{\sigma^2(y_i)} \\ &= [\frac{1}{n1}\sum_{k=1}^{n1}(\sum_{j=1}^{m}\lambda_{ij}f_{jk} + \epsilon_{ik} + \mu(y_i)) - \frac{1}{n2}\sum_{k=1}^{n2}(\sum_{j=1}^{m}\lambda_{ij}f_{jk} + \epsilon_{ik} + \mu(y_i))]/\sqrt{\sigma^2(y_i)} \\ &= [\sum_{j=1}^{m}\lambda_{ij}(\mu_1(f_j) - \mu_2(f_j)) + \mu_1(\epsilon_i) - \mu_2(\epsilon_i)]/(\sum_{j=1}^{m}\lambda_{ij}^2 + \psi_i)^{1/2}, \end{aligned}$$

where $\mu_1(y_i)$ and $\mu_2(y_i)$ are the mean Jacobian determinant values for the two populations; $\mu(y_i)$ and $\sigma^2(y_i)$ are the mean and variance for populations 1 and 2 combined; $n1$ and $n2$ are the respective number of subjects in each sample; m is the number of factors; λ_{ij} is the loading for variable i and factor j; f_{jk} is the j-th factor score of subject k; $\mu_p(f_j)$ is the mean value of factor j in the population p; $\mu_p(\epsilon_i)$ is the mean value of the error term i; and ψ_i the associated variance.

2.4 Implementation

In factor analysis, it is necessary to first define the number of factors to be considered. We present next an iterative algorithm that determines the number of factors based exclusively on the characteristics of the data set, instead of subjective considerations. The first step in the process is the computation of the covariance matrix $\mathbf{\Sigma}$. Since the purpose of factor analysis is to represent the covariance among variables, the expression for $\mathbf{\Sigma}$ presented in (5) is simplified to $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T$, as $\mathbf{\Psi}$ is diagonal and does not affect the covariance values. $\mathbf{\Sigma}$ is then decomposed into

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{\Theta}\mathbf{L}^T = (\mathbf{L}\mathbf{\Theta}^{1/2})(\mathbf{L}\mathbf{\Theta}^{1/2})^T,$$

where $\mathbf{\Theta}^{1/2} = diag(\sqrt{\theta_1}, \ldots, \sqrt{\theta_p})$ is the diagonal matrix with the square root of the eigenvalues of $\mathbf{\Sigma}$ and $\mathbf{L}$ is the matrix of the corresponding eigenvectors. The loading matrix can thus be estimated based on the sample covariance matrix as $\mathbf{\Lambda} = \mathbf{L}\mathbf{\Theta}^{1/2}$. The number of factors (number of columns in $\mathbf{\Lambda}$) can be initialized to the number of eigenvalues greater then 1, since they account for the variance in at least one variable.

The computed loading matrix is then rotated so that each variable will exhibit high loading for only a few factors. This can be achieved by finding a sequence of rotations that maximizes the variance of the squared Loadings in each column of $\mathbf{\Lambda}$ (*varimax* algorithm) [6]. The resultant loading values represent

the correlation between variables and factors. The following algorithm reduces the initial number of factors by discarding factors which do not have high correlation with at least two variables. Since the absolute value for correlation ranges from 0 to 1, we consider factors to be informative when they have loadings with absolute value greater or equal to 0.5. Convergence is achieved when the number of factor m_t at iteration t equals the number of factors m_{t-1} computed in the previous iteration. The algorithm is summarized below:

Begin
  Compute sample covariance matrix $\mathbf{\Sigma}$*;*
  Decompose $\mathbf{\Sigma}$ *into its eigenvectors* $\mathbf{L}$ *and eigenvalues* $\mathbf{\Theta}$*;*
  Set initial number of factors m_t *to the number of eigenvalues with value greater then 1;*
  Repeat until $m_t = m_{t-1}$
    $m_{t-1} \leftarrow m_t$*;*
    Estimate loadings as $\mathbf{\Lambda} = \mathbf{L}\mathbf{\Theta}^{1/2}$*;*
    Rotate loadings based on varimax *algorithm;*
    $m_t \leftarrow 0$*;*
    For $j \leftarrow 1$ *to* m_{t-1} *do*
      $nvar \leftarrow 0$*;*
      For $i \leftarrow 1$ *to* p *do if* $\lambda_{ij} \geq 0.5$ *then* $nvar \leftarrow nvar + 1$*;*
      If $nvar > 1$ *then* $m_t \leftarrow m_t + 1$*;*
End

3 Experimental Results

The set of MRI images used in the study is composed of 12 male and 16 female normal controls recruited as part of an ongoing study on schizofrenia being conducted at the Mental Health Clinical Research Center of the University of Pennsylvania. The subjects are right-handed with average age of 27 years (σ=5.8) for the male group and 28 years (σ=9.4) for the female. The images were acquired on a GE 1.5 Tesla instrument, using a spoiled GRASS pulse sequence optimized for high resolution, near isotropic volumes (flip angle = 35^o, TR = 35 ms, TE = 6 ms, field of view = 24 cm, 0.9375×0.9375 mm^2 in-plane resolution, 1.0 mm slice thickness, no gap). The images were obtained in the axial plane and the midsagittal slice extracted and reformatted into 256×256 8-bit images (Fig. 1).

The subject images were rigidly registered to a female template by identifying 3 pairs of corresponding points and applying least-square optimization. The callosa were segmented using the K-means clustering algorithm and extracted by manual delineation. Local registration was performed by elastically matching the template to each globally aligned subject callosum. The resulting displacement fields were used to compute the determinant Jacobian values to which factor analysis was applied.

Fig. 2 shows the identified factors. The algorithm took 5 iterations to converge from 27 to 18 factors which were highly correlated with at least 2 variables. The

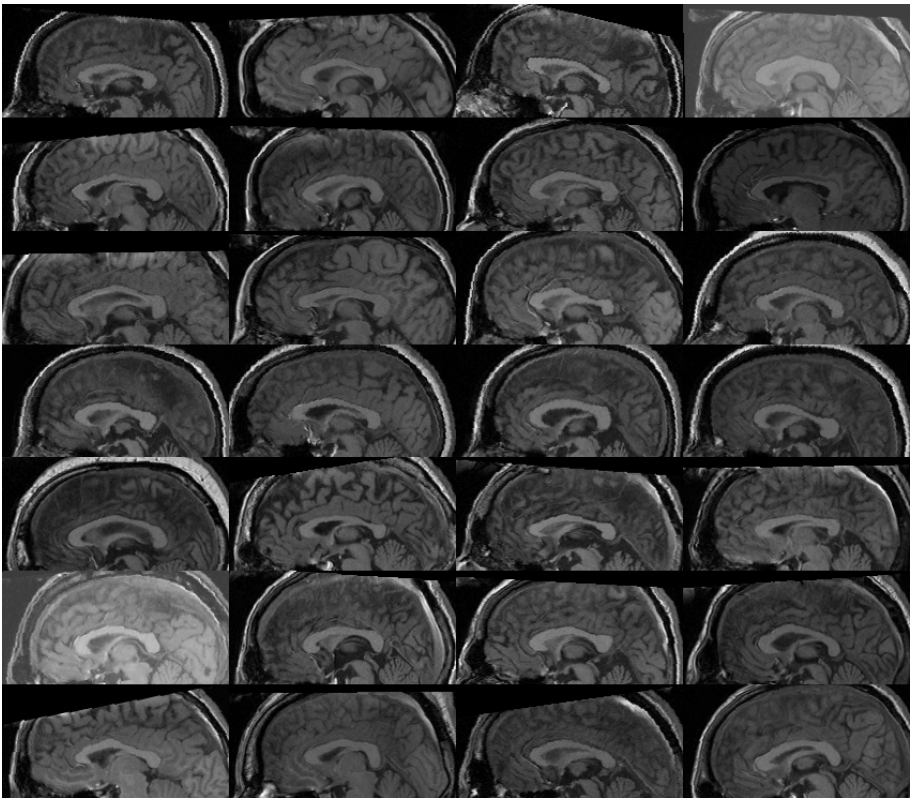

Fig. 1. Female (top) and male (bottom) subjects.

images show the absolute loading values that are greater then 0.5, for each informative factor. The gray levels are proportional to the absolute correlation, where pure white corresponds to complete correlation and the gray intensity shown in the background corresponds to a value of 0.5. These results can be compared with the findings of Gee *et al.* [4], who examined the same data set by computing the effect size between female and male populations. The comparison of Fig. 2 with Fig. 3 reveals an interesting relationship between the factors and the areas in which female and male morphology differ. The second factor shown in Fig. 2 coincides with the splenium and is is related to the major gender-related difference in the callosal anatomy, that has been observed in previous work [3,8]. The inferior portion of the splenium can be divided into 2 parts: the right half does not seem to discriminate the two populations, as can be seen in Fig. 3b, whereas the left region is actually larger in the female group. These two regions are clearly separated into two factors (factors 14 and 11, respectively) as can be seen in Fig. 2. Other callosal parts that contribute to shape difference

between females and males are indicated by factors 5, 9 and 15. Factor 15 is of particular interest, since it appears isolated in the right-most portion of the splenium (Fig. 3a), representing a region where the effect size is greater than 0.75.

Fig. 2. Factor analysis of callosal morphology. For each factor, voxels that are highly correlated are highlighted in white. Factors are numbered from 1 to 18, left to right and top to bottom.

(a) (b)

Fig. 3. Effect size analysis of callosal morphology. Voxels are highlighted where the effect size for area differences (female − male) are greater than 0.75 (a) and 0.5 (b).

4 Conclusion

A novel approach to morphometry was presented, in which the relationship among anatomic substructures are explored. The method is based on the factorial analytic model, where the covariance among variables is represented in a new basis of lower dimension. This enables more parsimonious descriptions and allows exploratory analysis of correlations which may reveal relationships between regions of interest that have not yet been observed. Factors can be visually identified as regions that embed strong correlation. The issue of determining the number of factors can be related to the desired degree of detailing in the analysis — fewer factors are expected to encompass greater regions with coarse correlation, whereas larger factor sets may represent smaller regions with stronger correlation. An algorithm was presented, which iteratively reduces the number of factors based on their contribution to the covariance representation of variable clusters. The ability of factor analysis to explore the relationship between parts of structures is a powerful tool for morphometry and a vast field for future work.

References

1. R. Bajcsy and S. Kovačič. Multiresolution elastic matching. *Computer Vision, Graphics and Image Processing*, 46:1–21, 1989. 379
2. T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models: Their training and applications. In *Computer Vision and Image Understanding*, pp. 38–59, 1995. 379
3. C. Davatzikos, M. Vaillant, S. Resnick, J. Prince, S. Letovsky, and R. Bryan. A computerized approach for morphological analysis of the corpus callosum. *Journal of Computer Assisted Tomography*, 20(1):88–97, 1996. 379, 383
4. J. C. Gee, B. Fabella, S. Fernandes, B. Turetsky, R. C. Gur, and R. E. Gur. New experimental results in atlas-based brain morphometry. In *Proceedings of the SPIE Medical Imaging 1999: Image Processing*, San Diego, 1999. Bellingham. 383
5. J. C. Gee and D. R. Haynor. Numerical methods for high-dimensional warps. In A. Toga, editor, *Brain Warping*. Academic Press, San Diego, 1999. 379
6. H. Harman. *Modern Factor Analysis*. University of Chicago Press, 1976. 381
7. L. Le Briquer and J. C. Gee. Design of a statistical model of brain shape. In *XV International Conference on Information Processing in Medical Imaging*, pp. 477–482, 1997. 379
8. A. Machado and J. C. Gee. Atlas warping for brain morphometry. In *Proceedings of the SPIE Medical Imaging 1998: Image Processing*, Bellingham, pp. 642-651, 1998. 379, 383
9. L. Marcus. Traditional morphometrics. In *Proceedings of the Michigan Morphometrics Workshop*, pp. 77–122. The University of Michigan Museum of Zoology, 1990. 379
10. A. Rencher. *Methods of Multivariate Analysis*. John Wiley & Sons, 1995. 378
11. R. Reyment and K. Jöreskog. *Applied Factor Analysis in the Natural Sciences*. Cambridge University Press, 1996. 379