Multi-aspect procedures for paired data with application to biometric morphing

Luigi Salmaso, Chiara Brombin, Pierfrancesco Galzignato, Giuseppe Ferronato

Multi-aspect procedures for paired data with application to biometric morphing

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2010-0059.R2
Manuscript Type:	Original Paper
Date Submitted by the Author:	05-Sep-2010
Complete List of Authors:	Salmaso, Luigi; University of Padova Brombin, Chiara; Univ. Padova Galzignato, Pierfrancesco; Univ. Padova Ferronato, Giuseppe; Univ. Padova
Keywords:	permutation tests, testing mean and variance, morphometry
Abstract:	According to Fisher, different tests of significance are appropriate to test different features of the same null hypothesis, thus leading to the Multi-Aspect (MA) testing issue. When dealing with paired data, usually inferences concern differences between the means. However, there are some circumstances in which it is of interest to test for differences between the variances. Here we present a nonparametric permutation solution to this problem. Our goal is to develop MA techniques for paired data, thus finding powerful tests, such that both differences in mean and in variance are separately identified.

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

Multi-aspect procedures for paired data with application to biometric morphing

Chiara Brombin* Luigi Salmaso* Giuseppe Ferronato†
Pier-Francesco Galzignato†

Abstract

As is common in case-control studies, treatments have an influence not only on mean values, but also on variance or distributional aspects. That is why several statistics, each one suitable for a specific aspect, are usually assessed (Salmaso and Solari, 2005). According to Fisher (1947, p. 185), different tests of significance are appropriate to test different features of the same null hypothesis (Lehmann, 1993), thus leading to the Multi-Aspect (MA) testing issue (Pesarin and Salmaso , 2010). When dealing with paired data, usually inferences concern differences between the means. However, there are some circumstances in which it is of interest to test for differences between the variances (McCulloch, 1987). Here we present a nonparametric permutation solution to this problem. Our goal is to develop MA techniques for paired data, thus finding powerful tests, such that both differences in mean and in variance are separately identified. The inferential procedures proposed in the paper and assessed throughout a simulation study are then applied to a real case-study in rhinoseptoplasty surgery.

Keywords

Finite-sample consistency; Multi-Aspect procedure; Combination-based permutation test; Shape analysis.

*Department of Management and Engineering, University of Padova, Italy, e-mail: salmaso@gest.unipd.it, Phone: +39 0444 99 87 20, Fax: +39 0444 99 88 84

†Department of Maxillofacial Surgery, University of Padova, Italy

1 Introduction

The methodological developments presented in this work curiously arise from graphical results obtained in an application of biometric morphing to a rhinoseptoplasty case study, in which 14 patients were evaluated before and after surgery. Biometric morphing is an innovative technique that combines procedures typical of shape analysis and image processing, and makes it possible to visualize average morphologic outcomes after a surgical intervention (Pahuta et al., 2009). Average results, induced by the surgery itself, have been graphically represented using this procedure, following the guidelines given in Pahuta et al. (2009). However, we were also interested in quantifying shape changes using inferential procedures. As said elsewhere (Brombin and Salmaso, 2009), inferential methods in shape analysis are parametric in nature and may not be very powerful unless a large number of cases is available. On the other hand, permutation tests represent an appealing alternative since they are distribution-free, allow for quite efficient solutions when the number of cases is lower than the number of covariates, they may be tailored for sensitivity to specific treatment alternatives and provide one-sided as well as two-sided tests of hypotheses (Blair et al., 1994). For this reason an extension of the NonParametric Combination (NPC) methodology (Pesarin and Salmaso , 2010) to statistical shape analysis has been proposed in Brombin (2009), where the two-independent-sample case has been discussed extensively. Here we examine the behaviour of the proposed tests in the two-dependent-sample case. In actual fact, when dealing with paired data, inferences usually concern differences between the means. However, as can be seen from the literature, there are also diverse circumstances in which it is of interest to test for differences between the variances (McCulloch, 1987). Examples may be found in psychology (Rothstein et al., 1981) or when it is of interest to evaluate the reliability/variability of measurements taken in two different laboratories (Snedecor and Cochran, 1980). Moreover, in case control designs, treatments may have an influence not only on mean values, but also on scatter or distributional aspects, thus leading to the Multi-Aspect (MA) testing issue (Pesarin and Salmaso , 2010). All these observations led us to consider the MA approach for paired data. In particular our goal was to find "good" tests, capable of separating mean aspect from scatter aspect (i.e. of detecting differences in means and differences in variances separately).
The paper is set out as follows. In Section 2 we present our case study, recalling some notions from shape analysis and biometric morphing. In Section 3 we introduce NPC methodology focusing on MA procedures for the paired data case, illustrating results from the literature along with our proposal. We also present results of a simulation study carried out to evaluate the per-

formance of the proposed inferential procedure. The results of the proposed MA procedure applied to the case study at hand are shown in Section 4, along with our concluding remarks.

2 The rhinoseptoplasty case study

2.1 Materials and methods

14 patients (4 men and 10 women) were evaluated before and after rhinoseptoplasty, a plastic surgery procedure dealing with both functional and cosmetic issues. A deviated septum may lead to a misshapen nose and even impede breathing. The primary purpose of the surgery is the recovery of normal nasal functions, with aesthetic issues considered later in order to harmonize the new, fully-functional nose with the rest of the face. Expected results of this surgery are improvements in terms of nose functionality, i.e. better breathing, as well as a straighter, reshaped nose (e.g. if a hump is present, it is removed, the tip is sculpted and the width of the nose is narrowed).
Follow-up ranged from 18 to 36 months (median of 24 months) when changes in soft tissue cover (i.e. skin) are assumed to be stabilized. Pre and post-surgery photographs were taken of each patient. All those analyzed gave their informed consent to participation in the study (including photographic documentation for scientific purposes only). Images of patients' faces were taken from 5 angles:

- frontal view,
- left and right sides at 45 degrees (oblique views),
- left and right sides at 90 degrees (lateral views).

For convenience, we used the right side profile only.
We emphasize that all those analyzed gave their informed consent to participation in the experiment (including the photography).
Patients' heads were positioned in the Frankfurt Plane (FP, see Farkas, 1994) thus evaluating subjects in a pose that approximates the position naturally assumed by the head in human beings. In order to avoid interobserver (operator-dependent) variability and bias, all photographs were taken by the same operator and for the same reason, and landmarks in the nasal profile were also placed by the same operator. The nose's underlying skeletal structure is commonly revealed by the overlying skin cover, producing specific topographic key landmarks and surgical guidepoints. For reliability of

linear measurements, a knowledge of the precise location of the landmarks on the surface of the nose is essential. When carrying out shape analysis, we followed the "landmark-based" approach that allows us to represent shapes by a discrete sampling of the object contours (Dryden and Mardia, 1998; Small, 1996).

Bookstein (1997) recommended the use of landmarks for the analysis of biological features, and also constraining the choice of landmarks to important features of the organism or of the biological structure that is the object of the study (Slice et al., 1996). According to the definition given by Dryden and Mardia (1998), a landmark is a point of correspondence on each object that matches between and within populations/a population. Landmark points are loci characterized by Cartesian coordinates as well as names designed to imply true homology, i.e. biological correspondence, from form to form (Dryden and Mardia, 1998). We remind readers that in geometric morphometrics the term homologous has no other meaning than that the same name is used for corresponding parts in different species or developmental stages (Slice et al., 1996). Moreover, these points represent a foundation for the explanations of the biological processes, and still nowadays many of the explanations of form accepted as epigenetically valid adduce deformations of the landmark locations Bookstein (1986).

We wish to highlight that despite the effectiveness of the "landmark-based" representation of shapes, automatic detection of landmarks is not always straightforward and the resulting shape analysis is determined by the choice of landmarks (Srivastava et al., 2005).

To define the shape of each profile, we selected 5 anatomical landmarks, following the guidelines given in Farkas (1994) (see Figure 1 (a)):

1. Nasion (n): the point in the midline of both the nasal root and the nasofrontal suture;

2. Rhinion (rhi): the osteocartilaginous junction at the hump of the nose;

3. Pronasale (prn) : the most protruded point of the nasal tip;

4. Subnasale (sn): the midpoint of the angle at the columella base where the lower border of the nasal septum and the surface of the upper lip meet;

5. Labiale superius (ls): the midpoint of the upper vermilion line.

2.2 Evaluation of shape configurations by means of biometric morphing and NPC methodology

Biometric morphing is an original procedure chiefly exploited in a paper by Pahuta et al. (2009). Using this procedure, pictures of a population of subjects can be combined into a single image explaining the average morphology of a precise anatomic region of interest. Biometric morphing combines typical statistical shape analysis and image processing procedures and can be defined as the application of thin-plate spline geometric morphometrics (TPSGM) to morphing. Resulting images have been called "morphs"(Pahuta et al., 2009). In actual fact, morphing is an effective image-processing tool that transforms (or morphs) one image into another in a seamless transition, while TPSGM allows us to quantify actual shape variation, thus taking the geometry of image deformation into account. The fundamental steps of the procedure to obtain "morphs"are described at great length in Pahuta et al. (2009) and we refer the reader to this paper for further details. Here we provide a brief summary of the algorithm which involves three steps:

1. Choose and digitize landmarks and semilandmarks (tpsDig software was used, Rohlf, 2007).

2. Perform Generalized Procrustes Analysis (GPA) to align patients. Pre- and post-surgery samples are processed separately (tpsRelw software was used, Rohlf, 2008).

3. Unwarp the images on the basis of the transformation from original registered points to target GPA points. Average the unwarped images to generate an average "morph" (tpsSuper software was used, Rohlf, 2004).

All the tps softwares mentioned above are available online at `http://life.bio.sunysb.edu/morph/`. While in Pahuta et al. (2009) the results of facial surgery are described by means of biometric morphing, our interest focuses on the statistical quantification of these visible changes in facial features. Our data are the landmark coordinates of aligned patients. Some graphical outputs obtained applying biometric morphing techniques are shown in Figure 1 (b-e). Moreover, average results are illustrated in the post-surgery morph: nasal dorsum is straight and the size and shape of the nasal tip has changed (see Figure 2). In Figure 1 (b-c), the mean shape is represented as grids of deformation and clear differences in terms of location are displayed, while Figure 1 (d-e) reveals possible differences in terms of spread around landmarks. These pictures constitute our starting points for

the formalization and then solution of the Multi-Aspect problem with paired data.

3 Multi-Aspect procedures for paired data

3.1 A brief overview

As previously stated, methodological issues discussed here arise from Figure 1 (d-e). On the one hand grids of deformation (Figure 1, b-c) reveal differences in mean between pre and post profiles, on the other Figure 1 (d-e) highlights different scatter structures, i.e. it seems that the surgery induces differences even in terms of spread around the chosen landmarks. This led us to investigate whether or not this evidence is "statistically"true or if it is just a graphical effect.

According to Fisher (1947, p. 185), different tests of significance are appropriate to test different features of the same null hypothesis (Lehmann, 1993), thus leading to the Multi-Aspect (MA) testing issue (Pesarin and Salmaso , 2010).

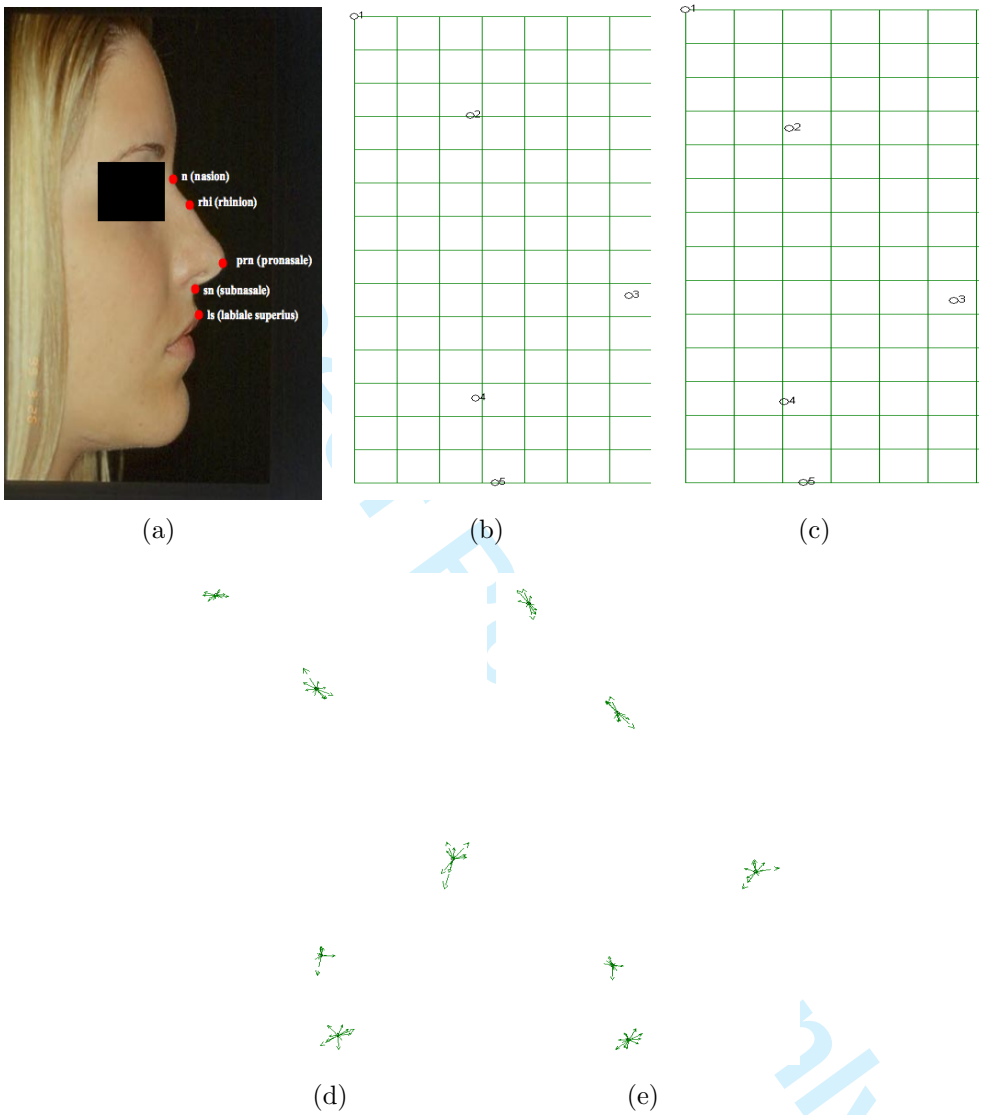
Here we provide some details concerning the Multi Aspect (MA) procedure in the two independent samples case. For further details, we refer the reader to Salmaso and Solari (2005) and Pesarin and Salmaso (2010). Let us consider a two-independent-sample problem in which the side-assumptions are that the treatment may act on the first two moments of responses belonging to the first group. Without loss of generality, let us assume that the data set and response model behave as $X_{1i} = \mu + \Delta_{1i} + \epsilon_{1i}, X_{2i} = \mu + \epsilon_{2i}$, $i = 1, \ldots, n_j$, $j = 1, 2$, where μ is a population nuisance constant, ϵ_{ji} are exchangeable random errors such that $\mu + \epsilon_{ji} > 0$ in probability, and $\Delta_{1i} \geq 0$ are non-negative stochastic effects which may depend on $\mu + \epsilon_{1i}$, and in addition satisfy the second-order condition $(\mu + \Delta_{1i} + \epsilon_{1i})^2 \geq (\mu + \epsilon_{1i})^2$, $i = 1, \ldots, n_1$.

Suppose the hypotheses are $H_0 : \{X_1 \stackrel{d}{=} X_2\}$ against $H_1 : \{X_1 \stackrel{d}{\neq} X_2\}$ and that we are essentially interested in the first two moments, so that the hypotheses become equivalent to $H_0 : \{(\mu_{11} = \mu_{12}) \bigcap (\mu_{21} = \mu_{22})\}$ and $H_1 : \{(\mu_{11} \neq \mu_{12}) \bigcup (\mu_{21} \neq \mu_{22})\}$, where $\mu_{rj} = \mathbb{E}(X_j^r)$ is the rth moment of the jth variable.

Let X_{ji}^*, $i = 1, \ldots, n_j$, $j = 1, 2$, denote a permutation of the original data.

A Multi-Aspect (MA) approach deals with one partial permutation test to each concurrent aspect, $T_1^* = \sum_i X_{1i}^*$ and $T_2^* = \sum_i X_{1i}^{*2}$, followed by their nonparametric combination.

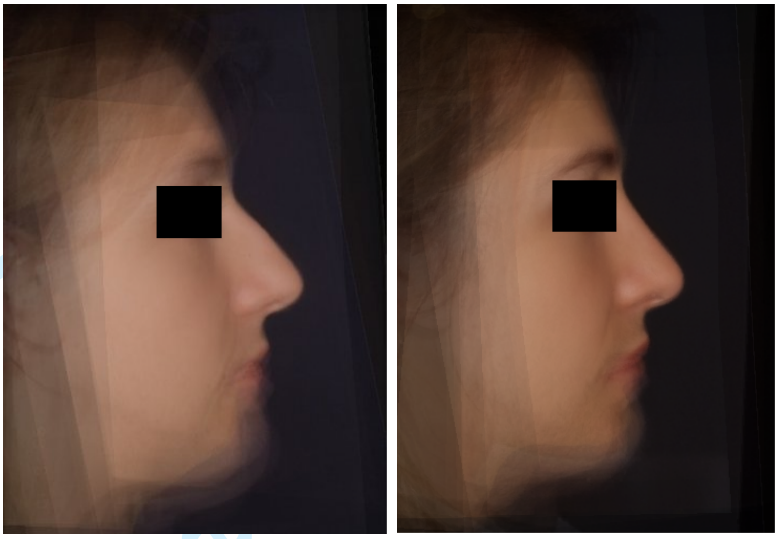
To summarize, the MA procedure embodies 3 steps:



(a) (b) (c)

(d) (e)

Figure 1: Selected landmarks (a). Pre- and post-surgery consensus, mean shape represented as grids of deformation (b)- (c) and as vector (scatter/variance) around each point (d)-(e).



(a) (b)

Figure 2: Nasal profile morphs, pre (a) and post surgery (b).

1. definition of the aspects of interest and selection of a suitable test statistic for each aspect;
2. organization of the aspects in a hierarchical structure;
3. choice of a proper combining function to combine within and between aspects.

The MA approach aims to supply a global evaluation on the basis of a set of partial tests, allowing also for the contrary, providing partial judgments on the strength of the global test of a finite number of dependent tests.
Partial and global tests are exact, unbiased and consistent and MA is robust under very mild conditions (Salmaso and Solari, 2005).
The extension of this algorithm to the shape analysis field is straightforward since what is needed is to perform T_1^* and T_2^* aspect tests for each coordinate of a single landmark and then consider their combination. We refer the reader to Brombin (2009) and Brombin and Salmaso (2009) for the full treatment of this issue.
However, when dealing with paired data, i.e. the two dependent samples case, things become more complicated, especially in the first step of the MA procedure since, when choosing the test statistic suitable for the aspect, dependencies between measurements must be properly taken into account.
When dealing with paired data, inferences may involve tests for evaluating

possible differences between the means or the variances. With reference to the last aspect, we start by presenting an interesting result from the literature. Pitman (1939) and Morgan (1939) proposed a test for equality of variances for paired, normally distributed data based on the correlation between the sums and the differences within pairs (McCulloch, 1987).
Let (X_1, Y_1), (X_2, Y_2), $\ldots (X_n, Y_n)$ denote i.i.d. pairs of observations and let $D_i = X_i - Y_i$, $i = 1, \ldots, n$ and $S_i = X_i + Y_i$, $i = 1, \ldots, n$.
Pitman (1939) and Morgan (1939) noted that $\mathrm{Cov}(S_i, D_i) = \sigma_X^2 - \sigma_Y^2$ and thus a test of $H_0 : \sigma_X^2 = \sigma_Y^2$ is equivalent to $H_0 : \rho_{DS} = 0$, where ρ_{DS} denotes the population correlation between D_i and S_i.
For the sake of simplicity, we call this test Pitman's test. When (X_i, Y_i) are bivariate normal, (D_i, S_i) are also bivariate normal and therefore a test of $H_0 : \rho_{DS} = 0$ can be made referring r_{DS}, the Pearson product-moment correlation coefficient between D_i and S_i, to the usual tables for significance.
Along with Pitman's test, which uses Pearson's product-moment correlation coefficient as a test statistic, other tests have been proposed in the literature, such as

- the jackknife test based on the log of the ratio of the sample variances $\log \frac{S_X^2}{S_Y^2}$ (Rothstein et al., 1981),
- the test based on $|\log |S_X^2| - \log |S_Y^2||$ in the presence of multivariate paired data,
- the test based on Spearman's rank correlation coefficient on D_i and S_i (McCulloch, 1987).

In particular the last proposal has several advantages over the jackknife procedure, since it is easily computed, its null distribution is largely tabulated and it has known asymptotic relative efficiency to Pitman's test under a variety of distributions (McCulloch, 1987).
Moreover, we recall that Pitman's test based on Pearson's product-moment correlation coefficient is not robust to departures from normality. The type I error size is larger than the nominal size for heavy-tailed distributions and smaller for light-tailed ones (McCulloch, 1987). For this reason, robust alternatives to Pitman's test, using the framework of the one-sample t-test have been explored in Grambsch (1994).
The main disadvantage of the test based on $|\log |S_X^2| - \log |S_Y^2||$ is that it can be calculated only if $(n - k - 1) > 0$, hence when the sample size is larger than the number of the observed variables.
For this reason, we suggest calculating, for each variable, the test $T_j =$

$\log(S^2_{X_j}/S^2_{Y_j})$, $j = 1, \ldots, k$ and then to combine partial test statistics using NPC methodology (in particular using the Direct combining function, Pesarin and Salmaso , 2010).
To summarize, our goal is to find powerful tests for multivariate problems, capable of separating mean from scatter aspect (i.e. to separately detect differences in means and differences in variances).

3.2 Simulation study

In the simulation study, we compare the power behaviour of

1. the test T_1, evaluating differences in mean (location aspect);
2. the test T_2, evaluating differences in variance (scatter aspect) and obtained by combining the k test statistics $T_{2,j} = \log(var(X_j)/var(Y_j))$, $j = 1, \ldots, k$, where k is the number of variables.
3. the test T_3, evaluating differences in variance (scatter aspect) and obtained by combining the k test statistics $T_{3,j} = \rho(D_j, S_j)$, $j = 1, \ldots, k$, using Spearman's rank correlation coefficient.

It is of interest to test $H_0 : \mathbf{X} \stackrel{d}{=} \mathbf{Y}$, where $\mathbf{X} = \boldsymbol{\mu} + \mathbf{Z_X}$, $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\delta} + \boldsymbol{\sigma}\mathbf{Z_Y}$, where $\boldsymbol{\sigma} = \boldsymbol{\sigma}_X/\boldsymbol{\sigma}_Y$ and $\mathbf{Z_X}$ and $\mathbf{Z_Y}$ are *i.i.d.* random deviates such that $\mathbb{E}(\mathbf{Z}) = \mathbf{0}$.
We evaluate their power behaviour when the correlation between variables representing the two situations (pre and post treatment) increases, when δ effect increases or when group variances σ^2_X and σ^2_Y are different (see Table 1).
Furthermore, we examine the case $n < k$, i.e. when the number of subjects n under study is less than the number of variables k (see Table 2).
With reference to data matrix generation, we recall that the entire data set was generated from a $2k$-variate normal distribution. The first k variables X represent data in the first situation and the other k variables Y represent data in the second situation (e.g. pre and post-treatment variables).
The mean vector μ is set equal to 0. A parameter δ was specified for Y variables, representing the location aspect or treatment effect.
$B = 1000$ iterations and $CMC = 1000$ permutations were carried out.
Results under the null and the alternative hypotheses are shown in Tables 1-3. Examining the results, we may state that the proposed tests are asymptotically separated, even in the presence of a small correlation between pre and post-treatment measurements that was expected to confound mean and

scatter aspects (see Table 3). The approximation is satisfactory under normality assumptions, even in the presence of small sample sizes.
We believe that the proposed approach is appealing since more detailed hypothesis testing is useful to obtain richer information.

4 Concluding remarks

Applying the same routine developed for the simulation study to the real data, we find differences in location at the nasion, pronasale and subnasale landmarks and difference in variance only at the nasion landmark.
Inferential results are consistent with the expectation of the doctors involved in this study and with the clinical literature in general.
However, we should emphasize that results obtained in this application could be altered by the GPA superimposition. Actually, we could regard GPA superimposition as a method for standardizing shapes. It is well known that different results may be obtained using standardized or original data (Brombin, 2009). GPA privileges the shape, but it may alter the dependency structures and, as a result, this produces permutationally non-equivalent tests within the permutation testing framework.
However, taking into account all these considerations, we may not be able to quantify exactly the variation in spread, because of the use of GPA superimposition itself, but if we find significant differences in variance after superimposition, this would reflect the presence of potential differences even in the raw data. Hence, we can conclude that surgery generates differences both in terms of mean and in variance and that these differences, expected from the surgeons, are statistically significant.
As stated, inferential methods for paired data usually aim to assess differences between the means. However, it may be of interest to test for differences between the variances (McCulloch, 1987). We have presented an extension of the MA techniques to the paired data case, in a nonparametric permutation framework. Results obtained in the simulation study are promising since the proposed tests are asymptotically separated, allowing us to discriminate differences in mean from those in variance. The approximation is satisfactory under normality assumptions, even in the presence of small sample sizes. We wish to point out that MA procedures provide more information on the underlying data structure than the traditional approach that merely assesses the equality in mean of pre and post-treatment measurements.

Table 1: Some simulation results. Simulations under H_0 ($n = 20$, $k = 10$, $\delta = 0$, $\sigma_X^2 = \sigma_Y^2 = 1$, $B = MC = 1000$) (a). Simulation under $H_{1,\delta}$, changes only in mean ($n = 20$, $k = 10$, $\delta = 1$, $\sigma_X^2 = \sigma_Y^2 = 1$, $B = MC = 1000$) (b). Simulation under H_{1,σ^2}, changes only in variance ($n = 20$, $k = 10$, $\delta = 0$, $\sigma_X^2 = 0.5, \sigma_Y^2 = 0.3$, $B = MC = 1000$) (c). Simulation under H_{1,δ,σ^2}, changes in both the aspects ($n = 20$, $k = 10$, $\delta = 0.5$, $\sigma_X^2 = 0.1, \sigma_Y^2 = 0.8$, $B = MC = 1000$) (d). Bold numbers indicate that the test is under the alternative, otherwise the test is under the null hypothesis.

(a)

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T_1	0.012	0.048	0.095	0.207	0.300	0.501
T_2	0.007	0.053	0.081	0.185	0.300	0.503
T_3	0.009	0.048	0.091	0.181	0.285	0.486

(b)

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T_1	**0.951**	**0.997**	**0.999**	**1.000**	**1.000**	**1.000**
T_2	0.017	0.052	0.100	0.191	0.285	0.477
T_3	0.013	0.047	0.092	0.189	0.291	0.468

(c)

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T_1	0.007	0.046	0.104	0.198	0.305	0.520
T_2	**0.792**	**0.956**	**0.986**	**0.996**	**0.998**	**1.000**
T_3	**0.734**	**0.925**	**0.967**	**0.991**	**0.997**	**1.000**

(d)

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T_1	**0.329**	**0.614**	**0.731**	**0.882**	**0.929**	**0.976**
T_2	**0.187**	**0.448**	**0.589**	**0.739**	**0.834**	**0.939**
T_3	**0.157**	**0.399**	**0.542**	**0.722**	**0.822**	**0.922**

Table 2: Other simulation results, when the number of cases n is lower than the number of variables k. Simulation under H_0 ($n = 10$, $k = 15$, $\delta = 0$, $\sigma_X^2 = \sigma_Y^2 = 1$, $B = MC = 1000$) (a). Simulation under $H_{1,\delta}$, changes only in mean ($n = 10$, $k = 15$, $\delta = 1$, $\sigma_X^2 = \sigma_Y^2 = 1$, $B = MC = 1000$) (b). Simulation under H_{1,σ^2}, changes only in variance ($n = 10$, $k = 15$, $\delta = 0$, $\sigma_X^2 = 0.5$, $\sigma_Y^2 = 0.3$, $B = MC = 1000$) (c). Simulation under H_{1,δ,σ^2}, changes in both the aspects ($n = 10$, $k = 15$, $\delta = 0.5$, $\sigma_X^2 = 1$, $\sigma_Y^2 = 0.8$, $B = MC = 1000$) (d). Bold numbers indicate that the test is under the alternative, otherwise the test is under the null hypothesis.

(a)

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T_1	0.000	0.032	0.087	0.192	0.297	0.480
T_2	0.010	0.058	0.105	0.187	0.292	0.495
T_3	0.010	0.056	0.094	0.182	0.300	0.494

(b)

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T_1	**0.000**	**0.844**	**0.962**	**0.995**	**0.999**	**1.000**
T_2	0.018	0.065	0.120	0.205	0.318	0.511
T_3	0.011	0.053	0.093	0.218	0.320	0.502

(c)

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T_1	0.000	0.030	0.084	0.183	0.290	0.486
T_2	**0.524**	**0.825**	**0.913**	**0.964**	**0.983**	**0.997**
T_3	**0.436**	**0.776**	**0.880**	**0.950**	**0.972**	**0.993**

(d)

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T_1	**0.000**	**0.237**	**0.419**	**0.671**	**0.792**	**0.918**
T_2	**0.107**	**0.293**	**0.448**	**0.639**	**0.758**	**0.891**
T_3	**0.088**	**0.259**	**0.404**	**0.605**	**0.750**	**0.885**

Table 3: Other simulation results, introducing some correlations between pre and post treatment variables ($0.33 \leq \rho \leq 0.77$). Simulation under H_0 ($n = 15$, $k = 10$, $\delta = 0$, $\sigma_{XY} = 0.5$, $\sigma_X^2 = \sigma_Y^2 = 1$, $B = MC = 1000$) (a). Simulation under $H_{1,\delta}$ ($n = 15$, $k = 10$, $\delta = 1$, $\sigma_{XY} = 0.5$, $\sigma_X^2 = \sigma_Y^2 = 1$, $B = MC = 1000$) (b). Simulation under H_{1,σ^2}, changes only in variance ($n = 15$, $k = 10$, $\delta = 0$, $\sigma_{XY} = 0.3$, $\sigma_X^2 = 0.5$, $\sigma_Y^2 = 0.3$, $B = MC = 1000$) (c). Simulation under H_{1,δ,σ^2}, changes in both the aspects ($n = 15$, $k = 10$, $\delta = 0.5$, $\sigma_{XY} = 0.3$, $\sigma_X^2 = 1$, $\sigma_Y^2 = 0.8$, $B = MC = 1000$) (d). Bold numbers indicate that the test is under the alternative, otherwise the test is under the null hypothesis.

(a)

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T_1	0.009	0.051	0.102	0.221	0.318	0.530
T_2	0.009	0.046	0.091	0.192	0.291	0.486
T_3	0.008	0.047	0.102	0.204	0.304	0.501

(b)

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T_1	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
T_2	0.007	0.039	0.086	0.192	0.274	0.494
T_3	0.012	0.056	0.100	0.211	0.294	0.509

(c)

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T_1	0.012	0.049	0.103	0.199	0.298	0.465
T_2	**0.948**	**0.996**	**0.999**	**1.000**	**1.000**	**1.000**
T_3	**0.919**	**0.992**	**0.996**	**1.000**	**1.000**	**1.000**

(d)

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T_1	**0.564**	**0.837**	**0.912**	**0.971**	**0.987**	**0.998**
T_2	**0.124**	**0.328**	**0.471**	**0.667**	**0.788**	**0.905**
T_3	**0.108**	**0.310**	**0.442**	**0.629**	**0.751**	**0.891**

Table 4: Partial p-values for each landmark coordinate (a) and for each landmark (b), combining using Fisher's combining function.

(a)

	T_1	T_2	T_3
n_x	0.0167	0.1237	0.3834
n_y	0.0142	0.0091	0.0794
rhi_x	0.5282	0.4058	0.4859
rhi_y	0.0285	0.6094	0.3180
prn_x	0.0167	0.8623	0.7489
prn_y	0.1515	0.0549	0.0394
sn_x	0.0167	0.8911	0.8681
sn_y	0.0142	0.1629	0.1771
ls_x	0.0167	0.2818	0.1019
ls_y	0.8617	0.4207	0.3907

(b)

	T_1	T_2	T_3
n	0.0005	0.0093	0.1307
rhi	0.0640	0.5380	0.4045
prn	0.0066	0.1986	0.1311
sn	0.0005	0.4233	0.4338
ls	0.0573	0.3699	0.1629

Acknowledgements

Authors wish to thank the University of Padova (CPDA088513/08) for providing the financial support for this research. The authors wish also to thank the anonymous Referees for valuable comments and suggestions which led to an improved version of this paper.

References

Blair, R. C., Higgins, J. J., Karniski, W., Kromrey, J. D. (1994). A study of multivariate permutation tests which may replace Hotelling's t^2 test in prescribed circumstances. *Multivariate Behavioral Research* **29**, 141-163.

Bookstein, F. L. (1986). Size and shape spaces for landmark data in two dimensions. *Statistical Science* **1**, 181-242.

Bookstein, F. L. (1997). Shape and the information in medical images: A decade of the morphometric synthesis. *Computer Vision and Image Understanding* **66**, 9-118.

Brombin, C. (2009). *A nonparametric permutation approach to statistical shape analysis*. Ph.D. thesis, University of Padova, Italy.

Brombin, C., Salmaso, L. (2009). Multi-aspect permutation tests in shape analysis with small sample size. *Computational Statistics & Data Analysis* **53**, 3921-3931.

Dryden, I. L., Mardia, K. V. (1998). *Statistical Shape Analysis*. John Wiley & Sons, London.

Farkas, L. G. (1994). *Anthropometry of the Head and Face* (2nd ed.). Raven Press, New York.

Fisher, R. A. (1947). *The Design of Experiments* (4th ed.). Hafner Press, New York.

Grambsch, P. M. (1994). Simple robust tests for scale differences in paired data. *Biometrika* **81**, 359-372.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association* **88**, 1242-1249.

McCulloch, C. (1987). Tests for equality of variances with paired data. *Communications in Statistics-Theory and Methods* **16**, 1377-1391.

Morgan, W. A. (1939). A test for the significance of the difference between the two variances in a sample from a normal bivariate distribution. *Biometrika* **31**, 13-19.

Pahuta, M. A., Rohlf, F. J., Antonyshyn, O. M. (2009). Biometric morphing: A novel technique for the analysis of morphologic outcomes after facial surgery. *Annals of plastic syrgery* **62**, 48-53.

Pesarin, F. and Salmaso L. (2010). *Permutation tests for complex data: theory, applications and software*. Wiley, Chichester-New York.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika* **31**, 9-12.

Rohlf, F. J. (2004). *tpsSuper, Superimposition, image unwarping and averaging, version 1.14*. Department of Ecology and Evolution, State University of New York at Stony Brook.

Rohlf, F. J. (2007). *tpsDig2, digitize landmarks and outlines, version 2.12*. Department of Ecology and Evolution, State University of New York at Stony Brook.

Rohlf, F. J. (2008). *tpsRelw, Relative warps analysis, version 1.46*. Department of Ecology and Evolution, State University of New York at Stony Brook.

Rothstein, S. M., Bell, W. D., Patrick, J. A., Miller, H. (1981). A jackknife test of homogeneity of variance with paired replicates data. *Psychometrika* **46**, 35-40.

Salmaso, L., Solari, A. (2005). Multiple Aspect Testing for case-control designs. *Metrika* **12**, 1-10.

Slice, D. E., Bookstein, F. L., Marcus, L. F., Rohlf, F. J. (1996). A glossary for geometric morphometrics. *Advances in Morphometrics* **284**, 531-551.

Small, C. G. (1996). *The Statistical Theory of Shape*. Springer, New York.

Snedecor, G. W., Cochran, W. G. (1980). *Statistical Methods*, 7th ed. Ames, IA: Iowa State University Press.

Srivastava, A., Joshi, S., Mio, W., Liu, X. (2005). Statistical shape analysis: Clustering, learning and testing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**, 590-602.

Responses to the Reviewer 1

A.1 page 3, last paragraph: the manuscript says both that the answer positions in the Frankfurt plane and that they were in the normal rest position for each subject. This seems inconsistent.

R.1 We have modified the sentence in the paper.

A.2 page 4, line 6: being prominent is not so important, what is important is that they are repeatable homologous landmarks of biological interest. regarding the word homologous, most biologists would differ with the statement you attribute to Slice et al. (1996). While statistical analyses can be performed with an arbitrary assignment of homology to landmarks, biological interpretation is dependent upon this being done in a meaningful way.

R.2 We have replaced "prominent" with "important" in the paper.
When working with shape analysis, it is always of primary importance to well pose the biological question and, as a consequence, to find a sensible biological interpretation. Here, the adjective "prominent" refers to the fact that landmark points must be important/significant from a biological point of view, as said later on. Actually we have emphasized that "these points represent a foundation for the explanations of the biological processes, and still nowadays many of the explanations of form accepted as epigenetically valid adduce deformations of the landmark locations Bookstein (1986)".
Moreover, we say that "To define the shape of each profile, we selected 5 anatomical landmarks, following the guidelines given in Farkas (1994)". Since Leslie Farkas is ascribed as the pioneer and the most important contributor in the anthropometrics of the human face, we believe that this may guarantee that a meaningful biological interpretation may be properly done on these points.

A.3 page 11, line 24: I am not sure of the relevance of the assertion that the differences are visible to the naked eye. It would be nice to have some other method of quantifying the amount of the difference.

R.3 We have modified the sentence in the paper. In fact, we have quantified differences between pre and post surgery profiles by means of statistical shape analysis and permutation tests. However, the hump reduction may be appreciated just looking at pre and post surgery nasal profile morphs. In practice, results in rhinoplasty operations are observable after the first two weeks and definitive about two years after the intervention.

A.4 pages 12 – 14: I'm not sure of the meaning of the color coding. It is not mentioned in the caption.

R.4 Bold numbers indicate that the test is under the alternative (hence, we are evaluating power), otherwise we are evaluating the type I error rate (i.e., the test is under the null hypothesis).