Tina Memo No. 2018-005
Presented at the Royal Entomology Society, St. Albans, April 2018.

Procrustes Analysis of Shape and How to Fix it.

N.A.Thacker.

Last updated
17 / 5 / 2018

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Procrustes Analysis of Shape: and How to Fix it.

N. A. Thacker, ISBE, University of Manchester

Abstract

I was asked if I wanted to give a talk on shape analysis from a computer vision perspective.

This presentation will cover a brief summary of the approach to morphometric shape analysis based upon landmark points, often referred to as Procrustes Analysis. It will explain the statistical problems of the method along with some popular misconceptions. In particular, the misconceptions relate to the idea that using unscaled Euclidean distances when comparing point positions makes derived models more biologically meaningful. Whilst the statistical problems are entirely due to the inapprorpiate Likelihood function, used for object aligment and data modelling, which follows from this.

A modification of the method, (recently published in Frontiers in Zoology), will be presented which incorporates point uncertainty distributions, which are estimated as part of the model. Results will be presented for a variety of real datasets, showing how the conventional problems with analysis, and particularly those associated with statistical consistency, model selection and semi- or psuedo- landmarks are solved.

Computer Vision and Pattern Recognition

- Pattern Recognition; classifiers trained on exemplar data generate class labels.

  e.g. ANNs (deep learning), nearest neighbour classifiers.

  Methods lack uncertainty estimation, not suitable for quantitative analysis (measurement or counting), unless supplemented by extensive simulation (which is difficult).

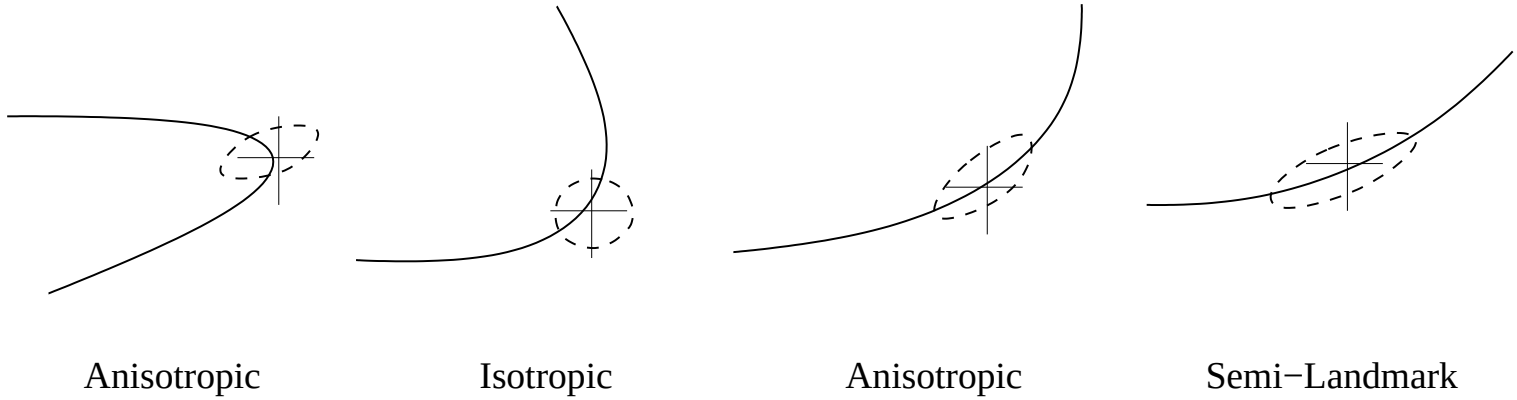
- Computer vision; Model based methods constructed from specific image features.

  e.g. shape and appearance models, Procrustes analysis.

  Can only be applied to groups of very similar objects. Methods based upon Likelihood can give error estimates and are usable for quantitative science (measurement of differences and classifications). Simulation is needed but tractable.

Semi-Landmarks

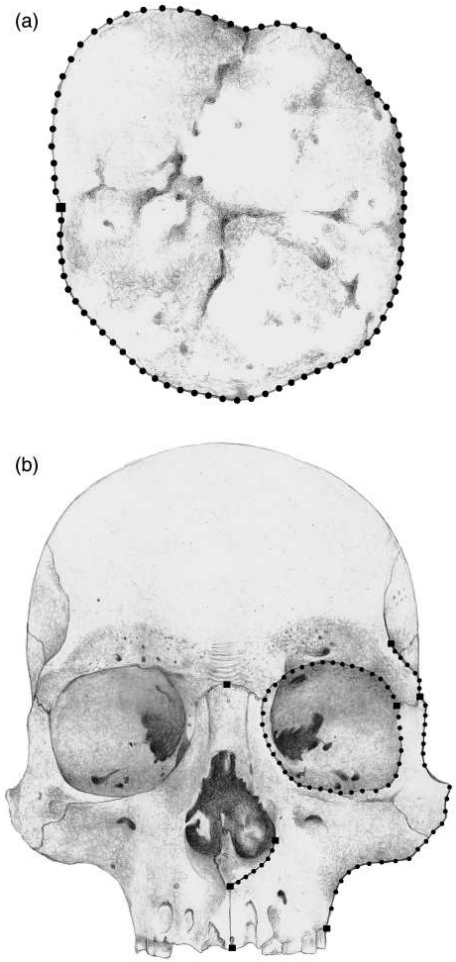
Some landmarks are more difficult for humans to localise than others.



When extracting information from images this can be understood in terms of the uncertainty of an auto-correlation based image patch localisation.

Features localised on a smooth curve have no well defined position (in either 2D/3D)- semi-landmarks.

Semi-Landmarks are Everywhere in Biological Data.



Procrustes analysis cannot deal properly with semi-landmarks.

Procrustes

- Align each example shape to centroid of 2D/3D landmarks.
- Rotate and scale individual shape data to L.S. position of 2D/3D landmark positions.
- Form least squares average of mark-up locations.
- Analyse variation around the mean using Principle Component Analysis (PCA).

One view of this approach is that it is a geometrically motivated approximation to tangent spaces.

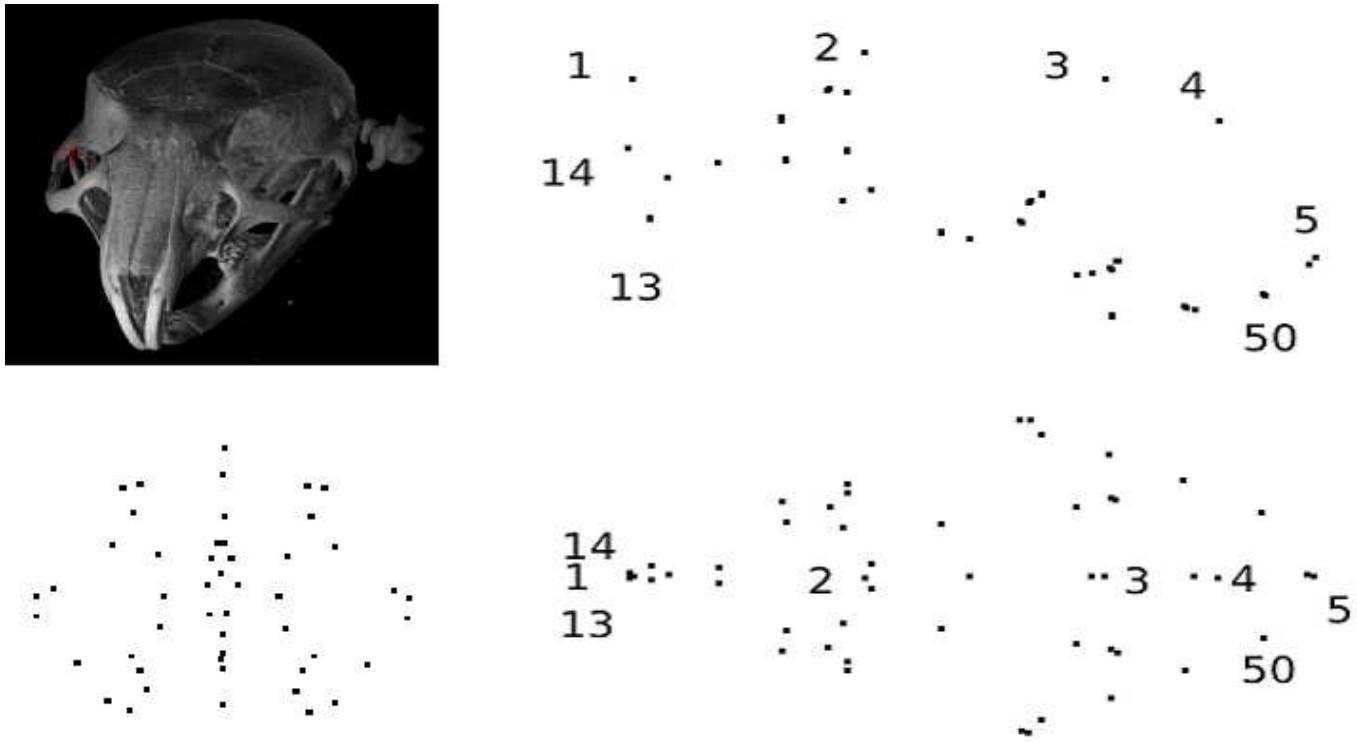
Why can't it deal with semi-landmarks?

Alternatively, this approach can be derived from Likelihood on an assumption of homogenous uniform independent errors on each landmark.

Known Problems;

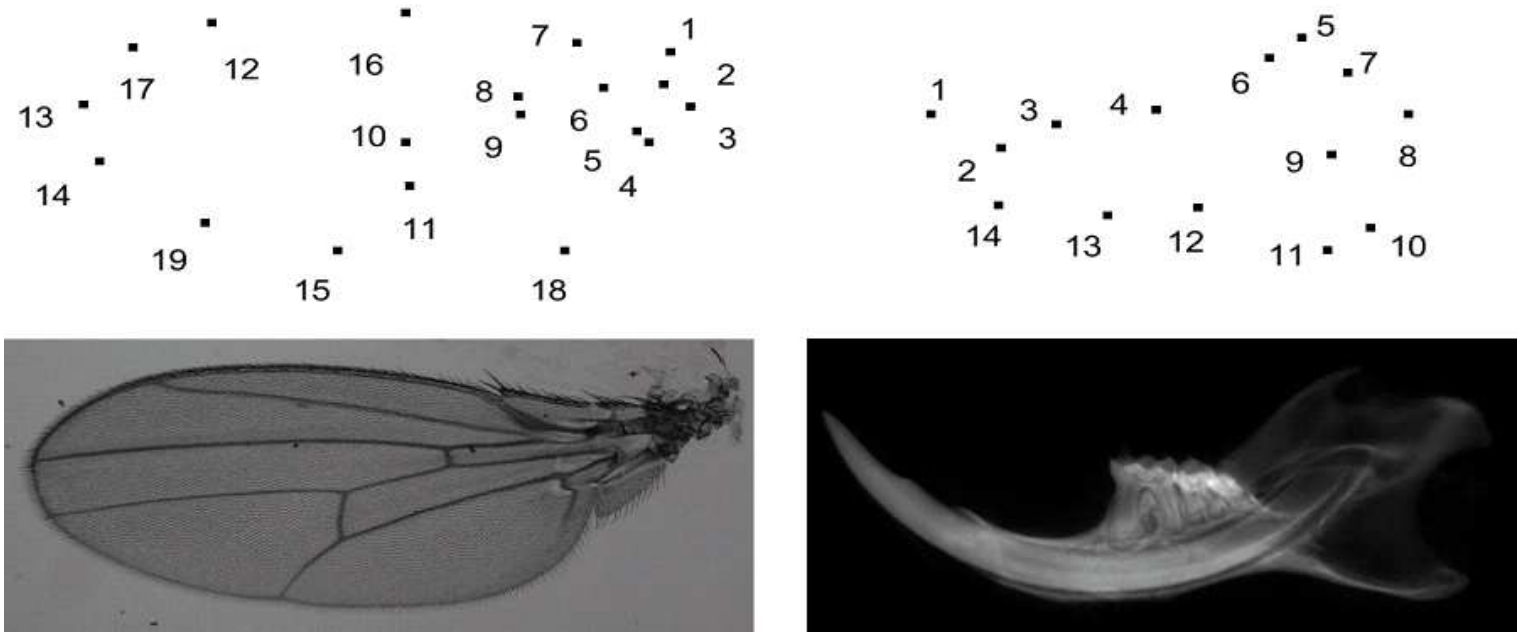
- it takes no account of true landmark localisation uncertainty,
- it is not statistically self-consistent,

Example Data: Mouse Skull



Typical landmarks for a sample volume (top-left) from the 3D mouse skull (MS) data when projected on the xy (top-right), zy (bottom-left) and xz (bottom-right) planes.

Example Data: Fly Wings and Mouse Mandibles



Typical landmarks corresponding to sample images of fly wings (left) and mouse mandibles (right); for the fly wing data, landmarks 1-15 correspond to the original data sets FL1, FL2, FR1 and FR2, while landmarks 16-19 were added later (to FL1) in order to experiment with semi-landmarks (P-FL1).

Data Summary

- Mouse mandible micro-CT images and consists of 337 samples with 14 landmarks per sample.
- Fly wing data left and right wings (L and R) of 200 female flies with repeats, each of these four data sets has 15 landmarks per sample we also added four semi-landmarks to each sample of the original data set
- Mouse skull (MS) 3D data of mark-ups from an automatic tool used to localise landmarks based on few given manual mark-up examples 42 samples with 50 landmarks per sample.

The animal datasets used in this paper have been approved according to German ethical standards. They were registered under number V312-72241.123-34 (97-8/07) and approved by the ethics commission of the Ministerium fur Landwirtschaft, Umwelt und landliche Raume on 27.12.2007.

Model Building

Repeatability data used to determine the most appropriate model order by adding new linear components until the residuals σ match expected measurement accuracy σ_m.

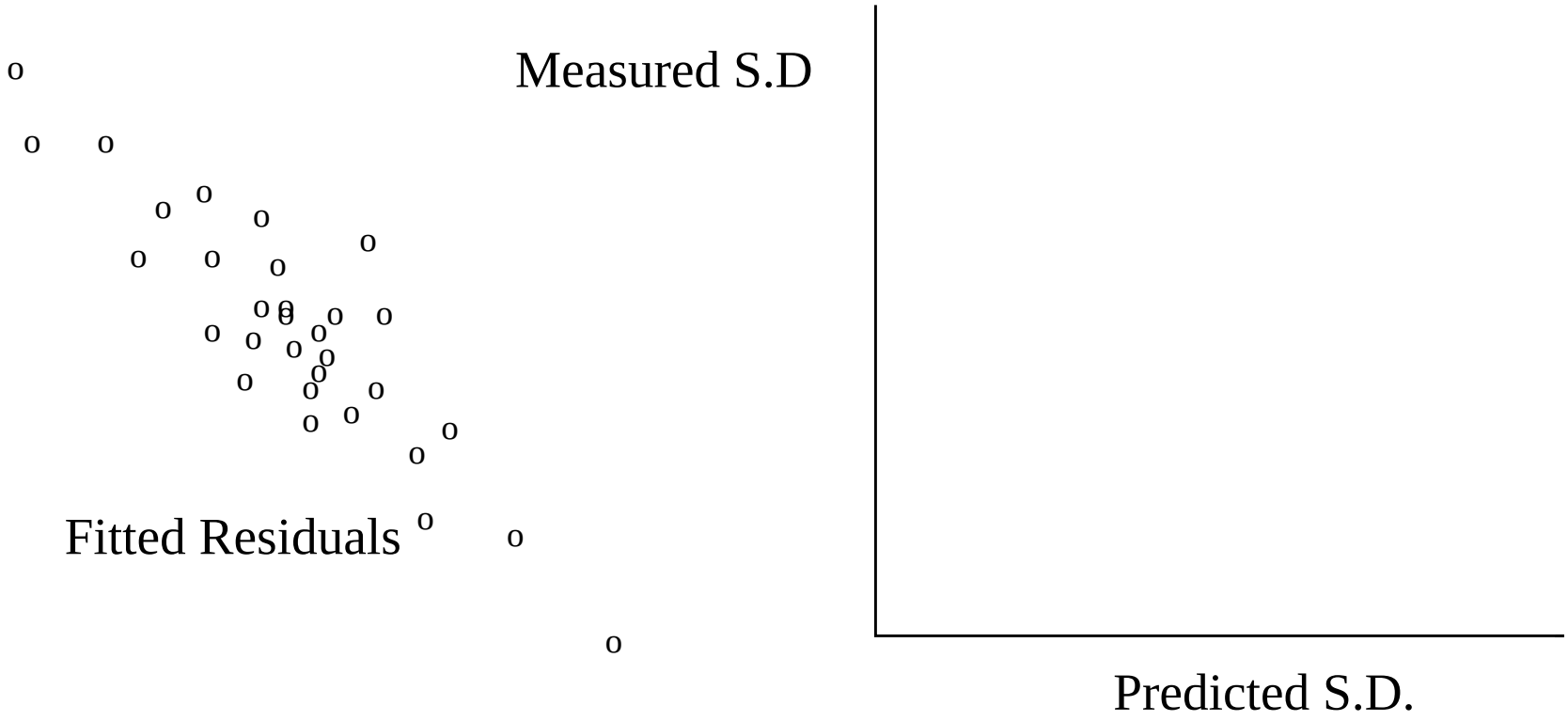
$$\sigma^2 = \sigma_m^2 + \sigma_b^2$$

3D Mouse Skull data we use 14 model components

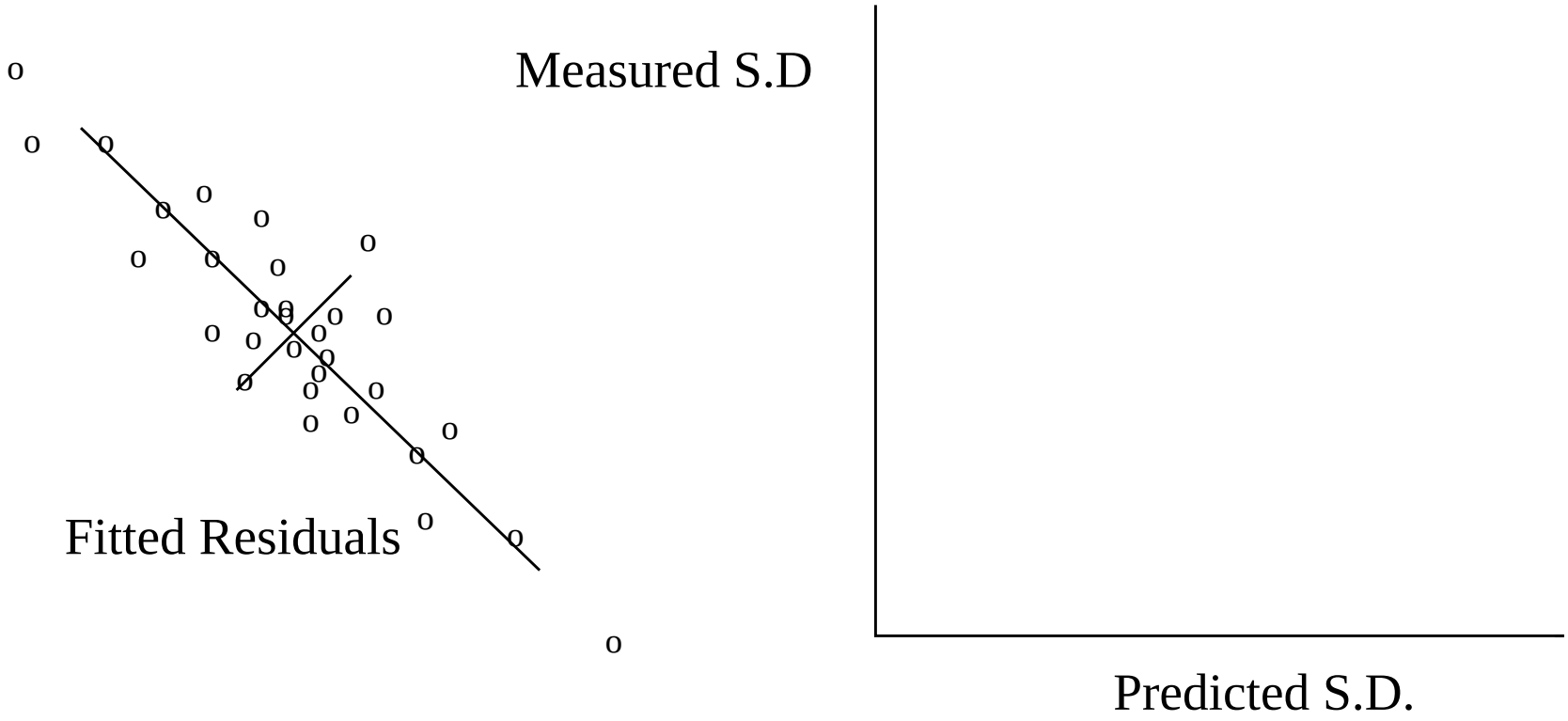
2D Mouse Mandible data we use 6 components

2D Fly Wing data we use 2-3 components

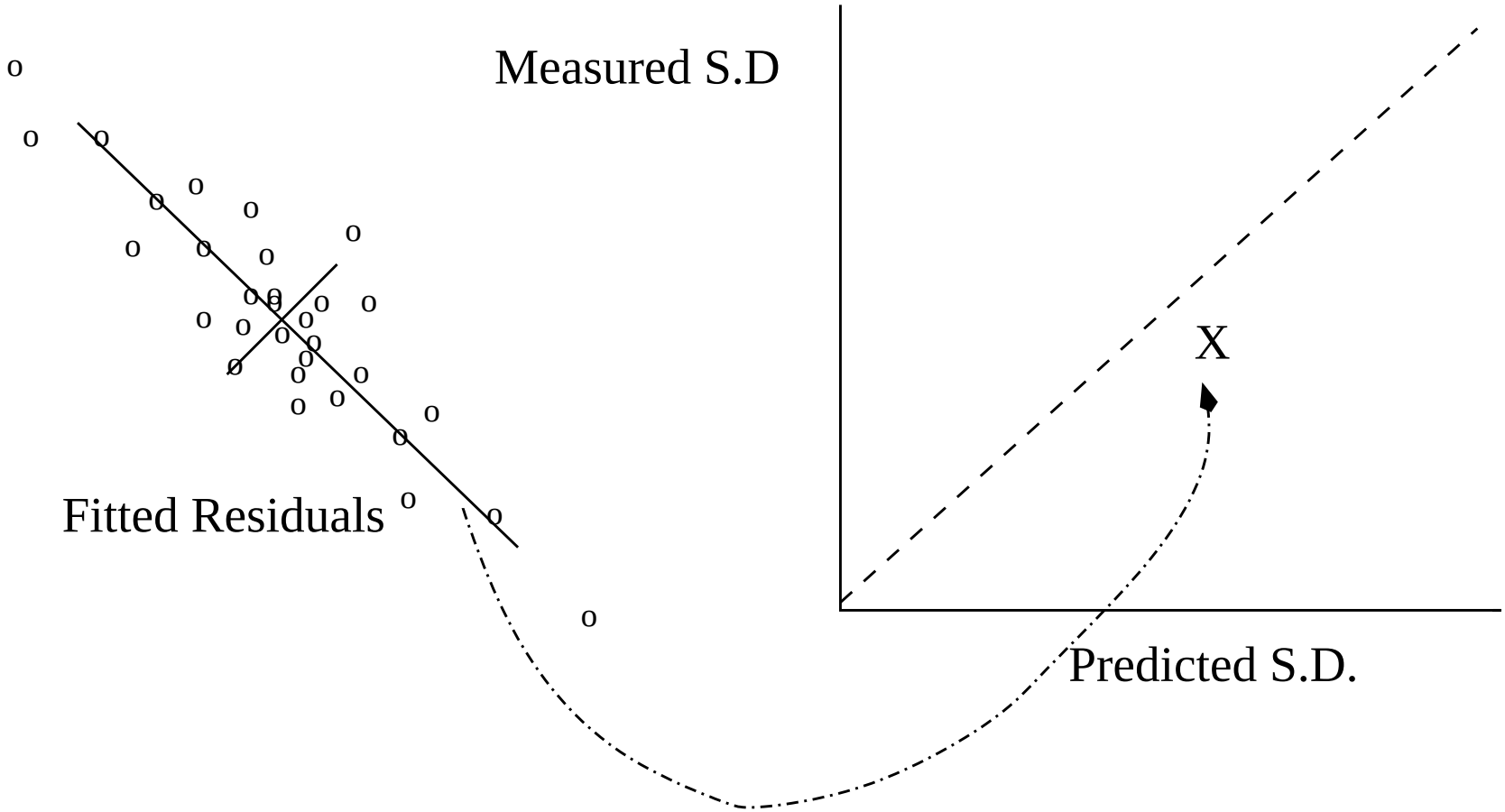
Monte-Carlo Evaluation of Procrustes Residuals.



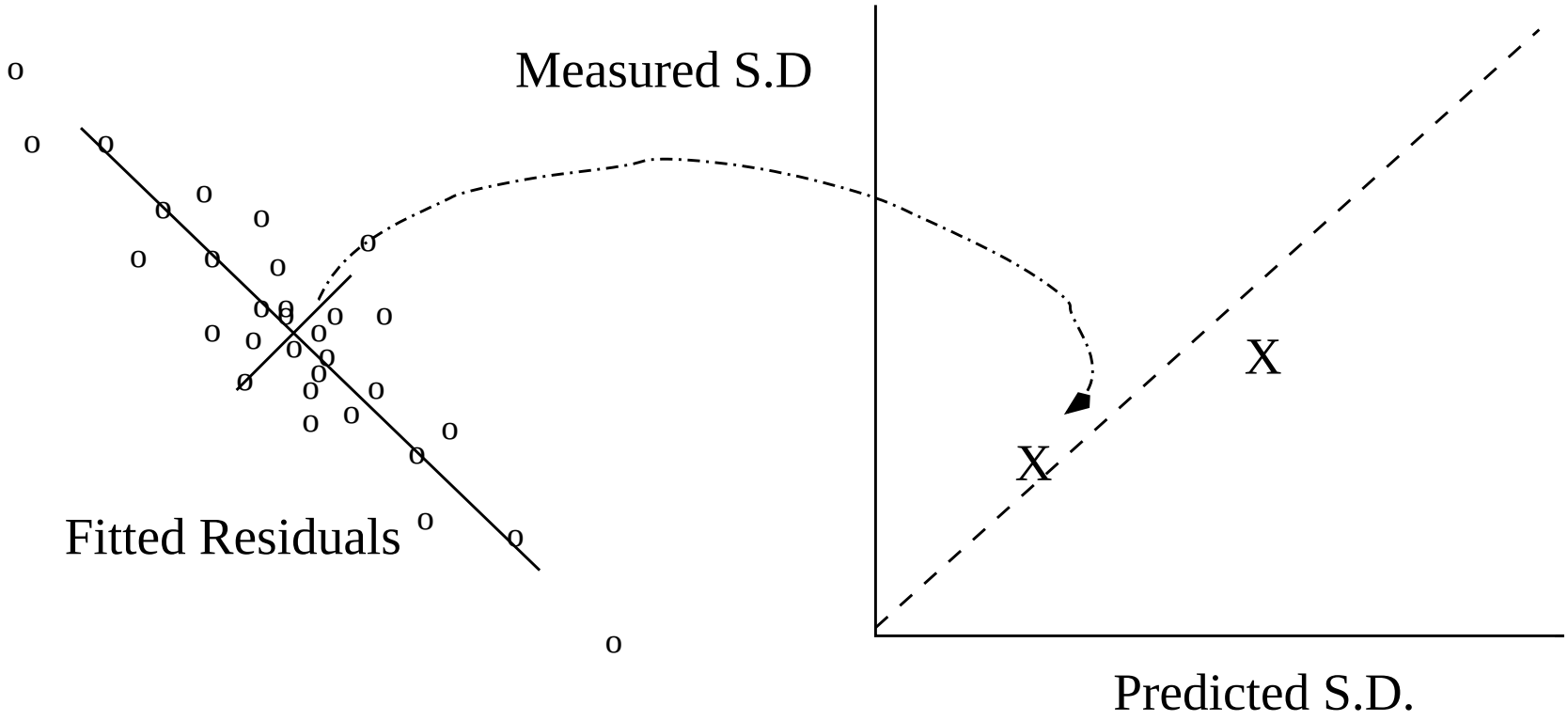
Monte-Carlo Evaluation of Procrustes Residuals.



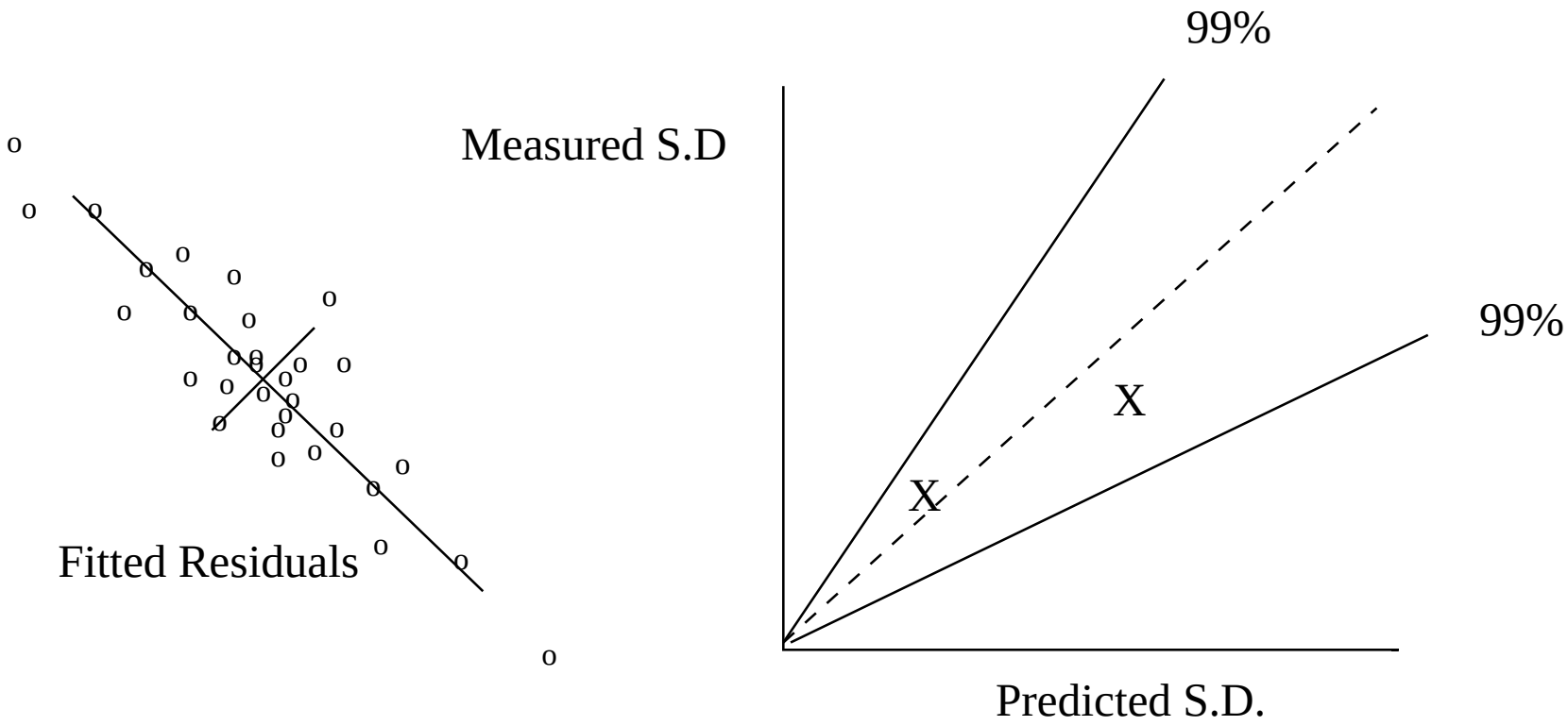
Monte-Carlo Evaluation of Procrustes Residuals.



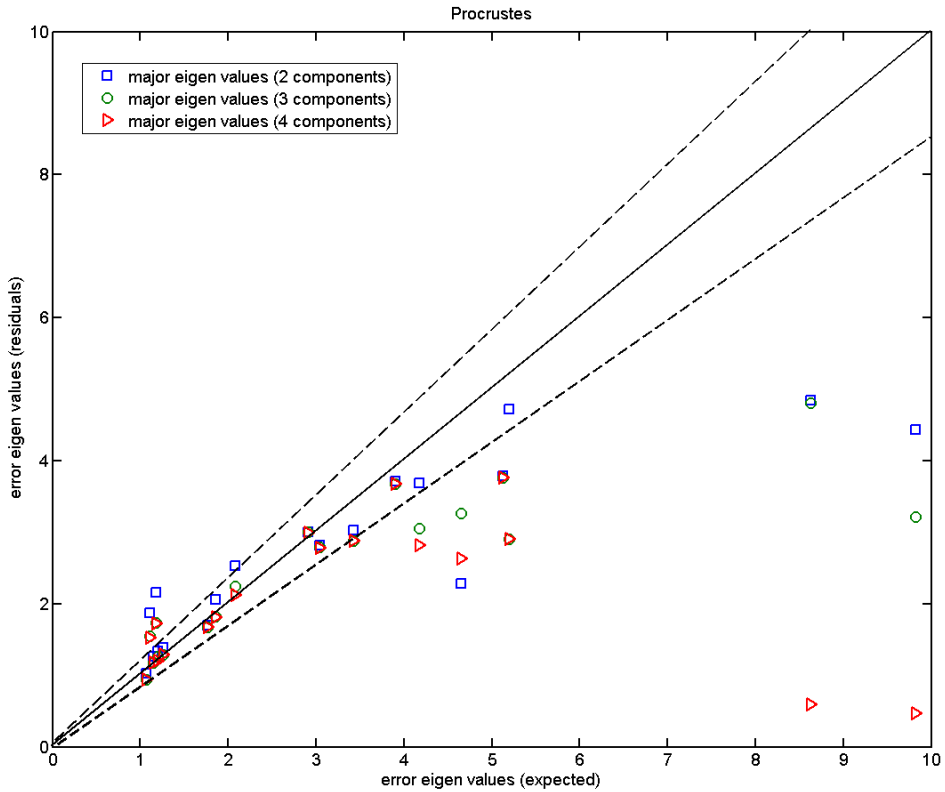
Monte-Carlo Evaluation of Procrustes Residuals.



Monte-Carlo Evaluation of Procrustes Residuals.

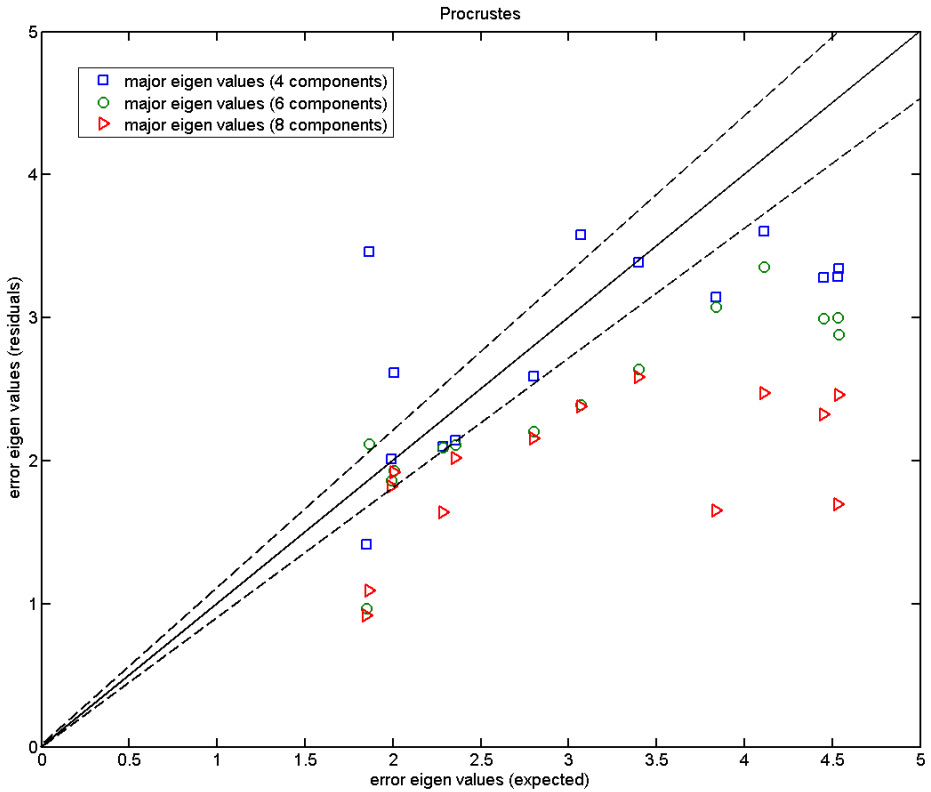


Monte-Carlo Evaluation of Procrustes Residuals.



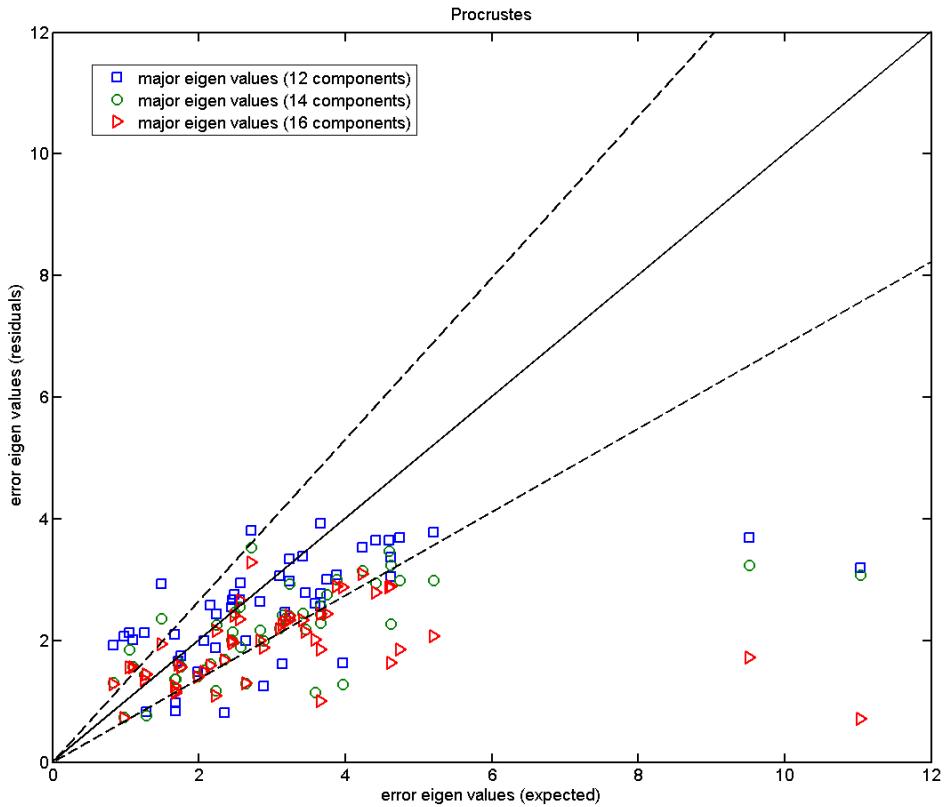
Fly wing data (P-FL1): error eigenvalues computed using the residuals after Procrustes alignment on the Monte-Carlo data, against the expected ones which were used when generating the simulated data; for 2, 3 and 4 model components; the two dashed lines show the $\pm 2.8\sigma$ range.

Monte-Carlo Evaluation of Procrustes Residuals.



Mouse mandible data (MM1): error eigenvalues computed using the residuals after Procrustes alignment on the Monte-Carlo data, against the expected ones which were used when generating the simulated data; for 4, 6 and 8 model components; the two dashed lines show the $\pm 2.8\sigma$ range.

Monte-Carlo Evaluation of Procrustes Residuals.



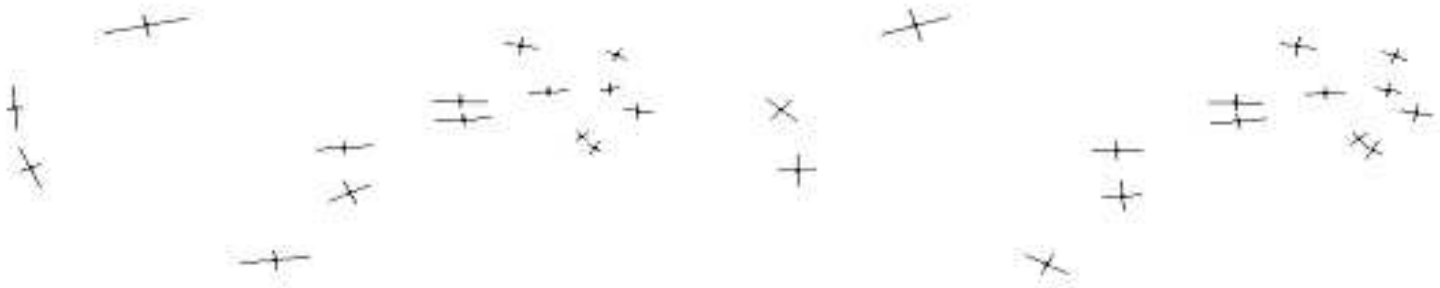
Mouse skull data (MS): error eigenvalues computed using the residuals after Procrustes alignment on the Monte-Carlo data, against the expected ones which were used when generating the simulated data; for 12, 14 and 16 model components; the two dashed lines show the $\pm 2.8\sigma$ range.

Likelihood Based Modeling with Localisation Uncertainties.

We start from the Procrustes model with an assumption of uniform measurement errors.

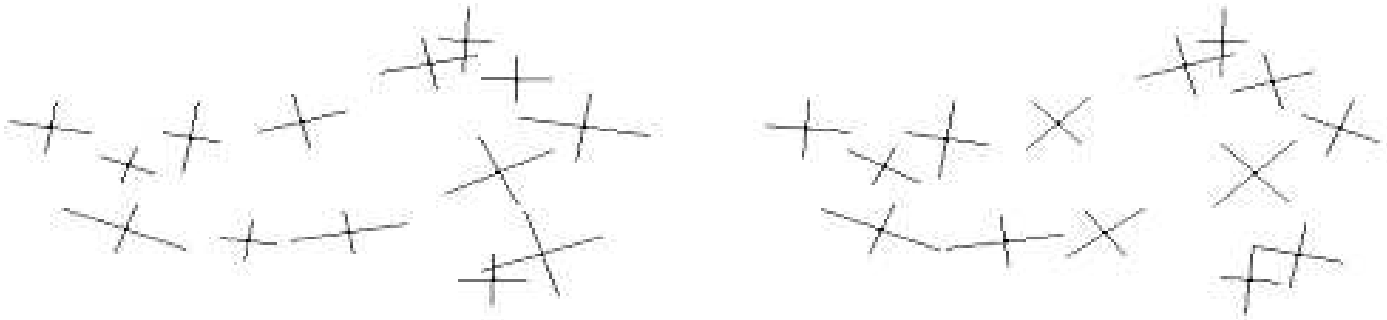
- Re-estimate the point specific localisation kernal (co-variance) based upon fitted residuals.
- **Correct the estimated co-variance for degree of freedom effects.**
- For a specific order of linear model, estimate the position of each shape sample using maximum Likelihood (i.e. not just L.S. alignment to the mean).
- Estimate the principle components of variation of the data in a "whitened" data space (i.e. A covariance weighted Likelihood, not L.S.).
- iterate the above steps until convergence.

Monte-Carlo Evaluation of Estimated Location Uncertainties.



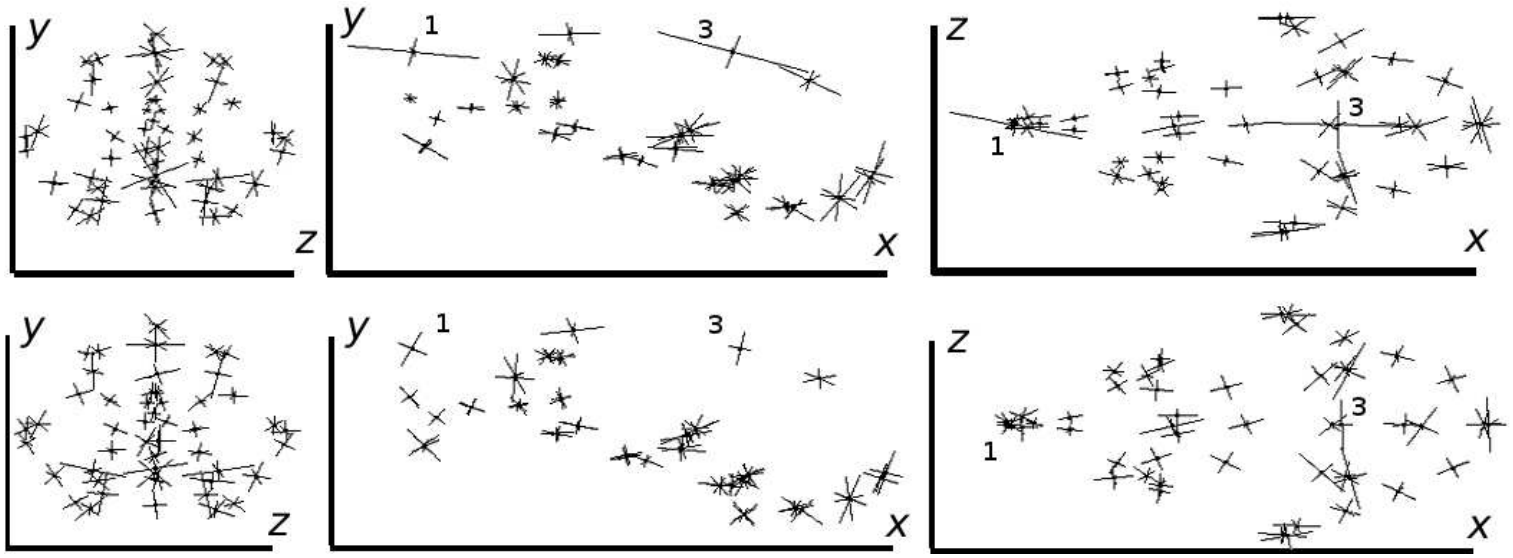
Fly wing data (FL1): error bars (×20) estimated using our method (left), and computed from the residuals left using Procrustes (right); 2-component models.

Monte-Carlo Evaluation of Estimated Location Uncertainties.



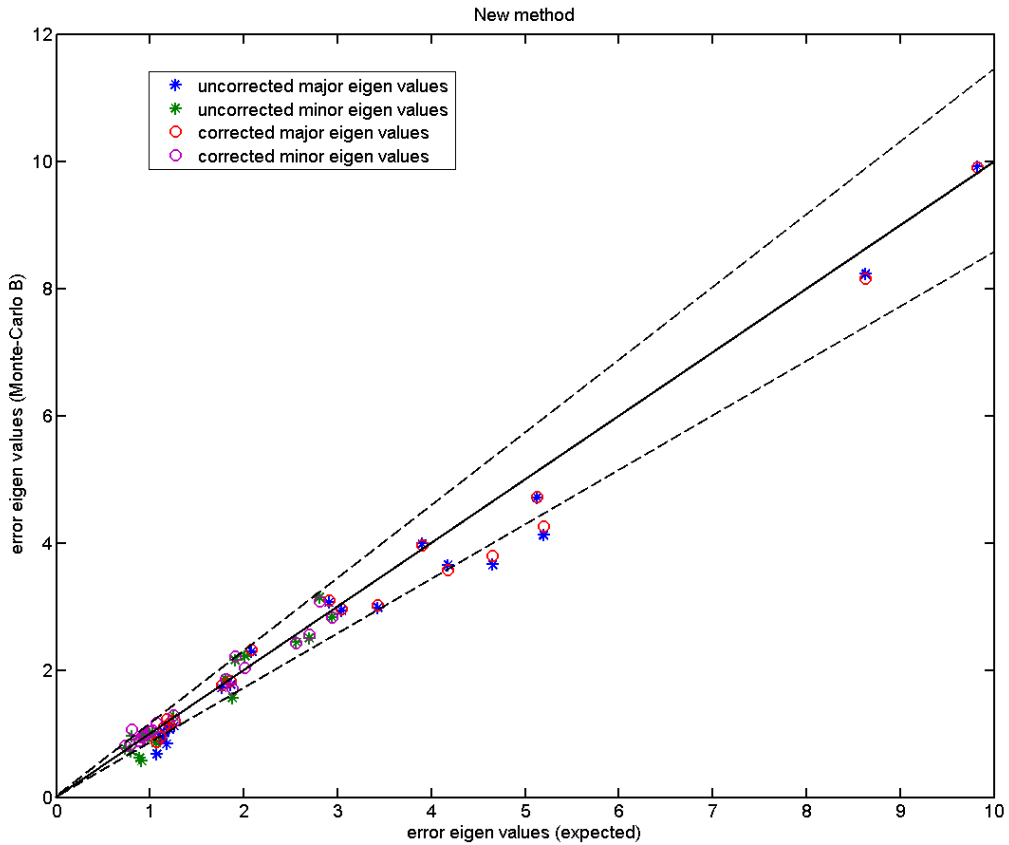
Mouse mandible data (MM1): error bars (×20) estimated using our method (left), and computed from the residuals left using Procrustes (right); 6-component models.

Monte-Carlo Evaluation of Estimated Location Uncertainties.



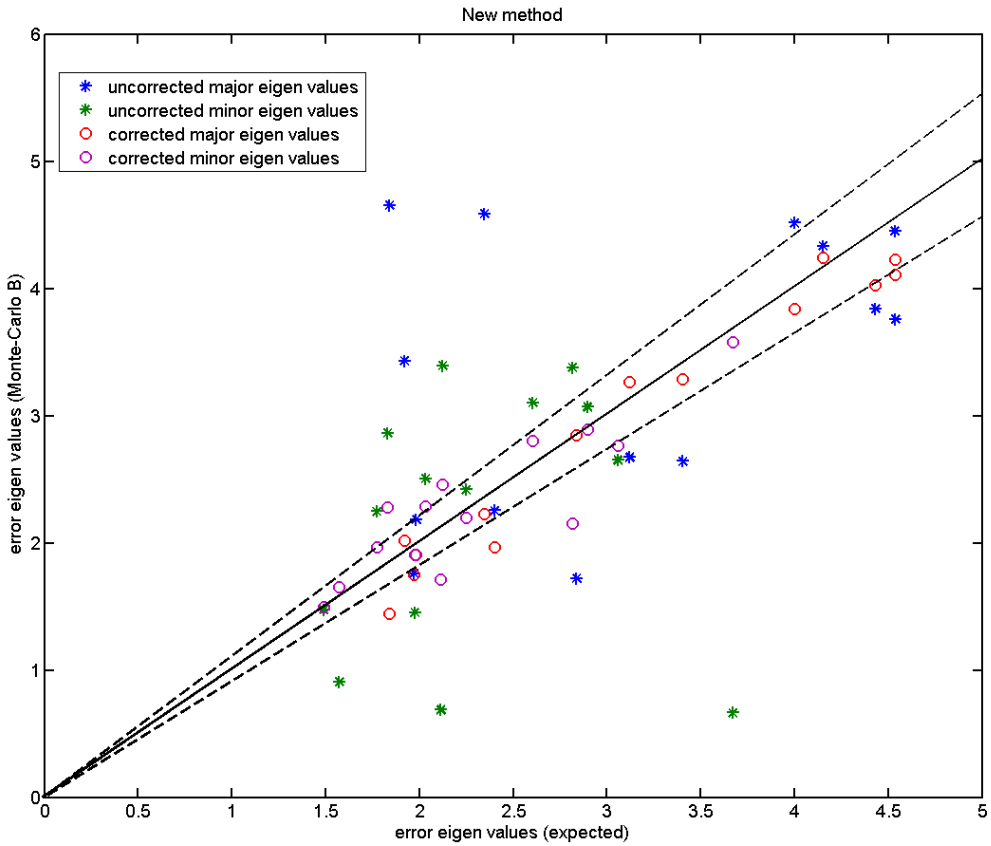
Mouse skull data: error bars (×30) estimated using our covariance-based method (top); and those computed using Procrustes residuals (bottom); 14-component models; projection planes: zy (left), xy (middle) and xz (right).

Monte-Carlo Evaluation of Estimated Location Uncertainties.



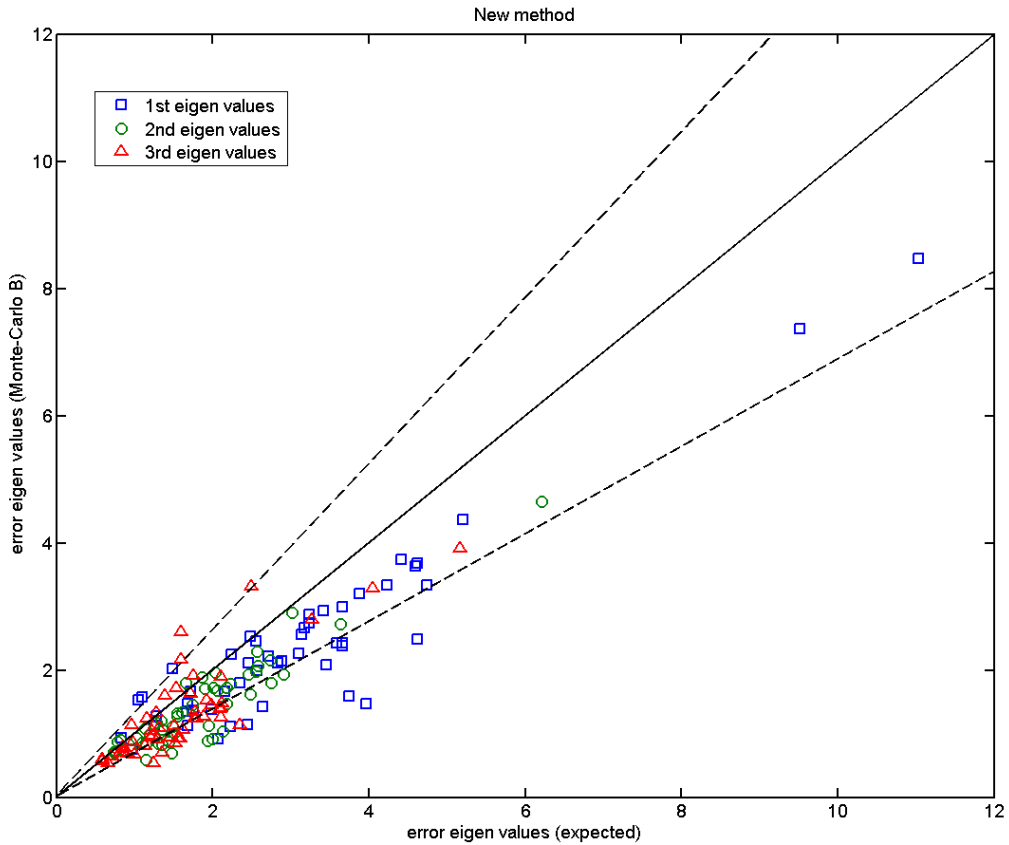
Fly wing data (P-FL1): error eigenvalues estimated using the Monte-Carlo data against the expected ones (estimated using the original data) which were used when generating the simulated data; using 2 model components; the two dashed lines show the $\pm 2.8\sigma$ range.

Monte-Carlo Evaluation of Estimated Location Uncertainties.



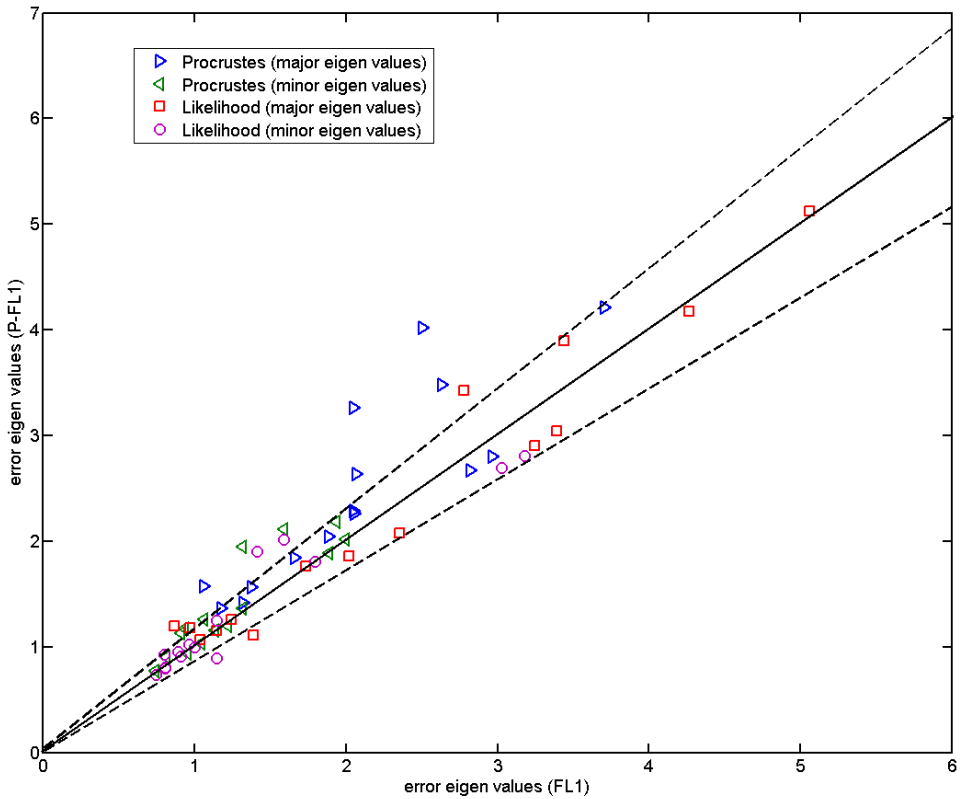
Mouse mandible data (MM1): error eigenvalues estimated using the Monte-Carlo data against the expected ones used when generating simulated data; independent 6-component models; there is considerable error in the uncorrected estimates; dashed lines show the $\pm 2.8\sigma$ range.

Monte-Carlo Evaluation of Estimated Location Uncertainties.



Mouse skull data (MS): error eigenvalues estimated using the Monte-Carlo data against the expected ones used when generating the simulated data; independent 14-component models; the two dashed lines show the $\pm 2.8\sigma$ range. (Bias due to eigen-vector orientation estimation)

Monte-Carlo Evaluation of Estimated Location Uncertainties.



Fly wing data: error eigenvalues estimated using the likelihood method and using the residuals after Procrustes alignment, when methods are applied to P-FL1 (4 semi-landmarks (16-19) added to FL1) against those when applied to FL1; the plot is for the 15 common landmarks (1-15). The dashed lines are the $\pm 2.8\sigma$ range.

Efficiency of Information Extraction.

Fisher Information $= \sum_i \frac{1}{\sigma_i^2}$

FI value	Procrustes	Likelihood
3D Mouse skull data (14-component model)	23.62	111.88
Mouse mandible data (6-component model)	6.60	19.55
Fly wing data (2-component model) 15 (+4) points	13.81 (+0.88)	25.46 (+2.47)

Here we show the Fisher Information (FI) value estimated from residual variances for the two methods and the three data sets studied.

Again, for Procrustes, variances used to compute the FI value are obtained from the residuals left after alignment between the data and the simulated linear model.

Our Likelihood method gives FI values roughly between two and four times those obtained using Procrustes.

The extra information gained by adding 4 semi-landmarks (25 % more data) is 10 %, which is about what would be expected for 4 new 1D measurements.

Conclusions.

Main conclusions;

- It is possible to extend Procrustes analysis to incorporate an-isotropic measurement uncertainty.
- A linear model can be constructed which estimates linear correlations in shape along with measurement error.
- The estimated models are statistically self-consistent.
- The models are more efficient (up to 4 x more data).

Contradicted Myths;

- Geometry (e.g. tangent spaces) are the theoretical basis for Procrustes.
- It is impossible to simultaneously estimate measurement error and linear model parameters.
- Semi-landmarks must be excluded from analysis.

Acknowledgements

Max Planck Institute for Evolutionary Biology, Ploen. Funding.

Hossein Ragheb, Software.

Paul Bromiley, Infrastructure.

Diethard Tautz, Anya Schunke, Mouse Data.

Chris Klingenberg, Fly Data.

www.tina-vision.net (Tina memo 2013-003)

The full paper was published in;

H. Ragheb, N. A. Thacker, P.A. Bromiley, A. C. Schunke, and D. Tautz. Quantitative Shape Analysis with Weighted Covariance Estimates for Increased Statistical Efficiency, Frontiers in Zoology, 10(16), April 2013.

Questions

Q - As the method involves a whitening re-projection of the data does this mean it is no longer modelling biology?

A - This cuts to the core of the method. This is done to correctly weight the landmarks during model construction in accordance with the application of Likelihood. The model can be projected back into the original space to model landmark positions afterwards.

The problems of lack of statistical self-consistency, inability to deal with semi-landmarks and poor efficiency of information extraction (as illustrated here), all stem from the decision to use a homgenous noise model. They are all solved by using the appropriate Likelihood construction.

What I didn't say - those who think that Procrustes somehow extracts biologically meaningful eigen-vectors, as a consequence of homogenous weighting should ask themselves why they are orthonormal. This property seems to contradict any biological argument. In the end all we can expect to do is to represent the orginal data in a more convenient form and lower dimensionality, in preparation for statistical hypothesis tests.

Q - All of the mathematics appears to be conventional linear algebra, there is no new mathematical idea?

A - Correct. The difficult thing here is not so much the mathematics as understanding how statistical principles justify this as the correct approach. It extracts the best (most informative) model of data.

Q - The paper you cite is five years old, the field seems to have overlooked it.

A - Yes. I do not work in the area of genetics/morphometrics, and beyond the journal publication and putting the software on our web site, I have made no efforts to publicise the work. In any case it will be difficult to distinguish this paper from any of the hundreds of other shape analysis papers, unless one has a good understanding of statistics.