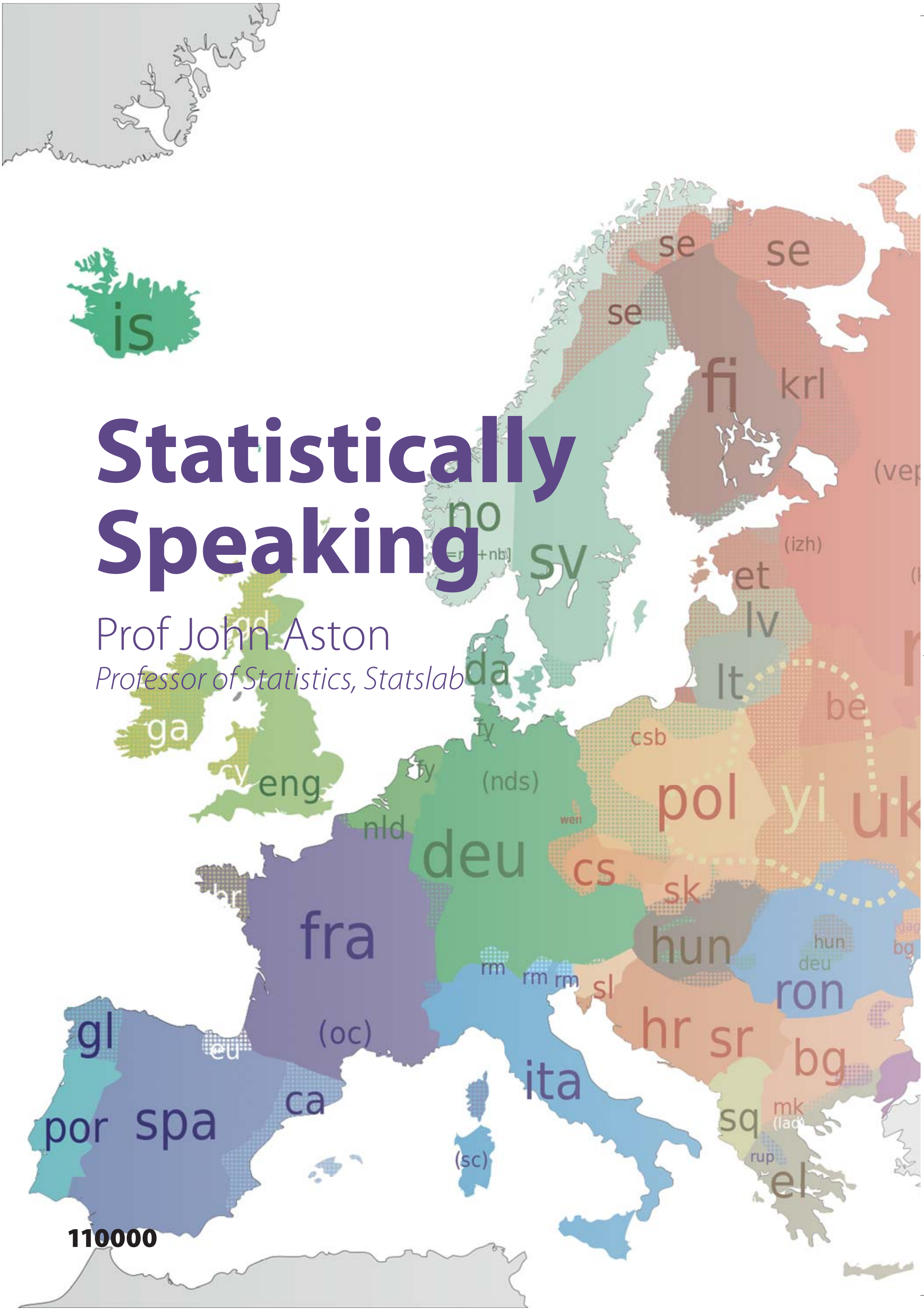
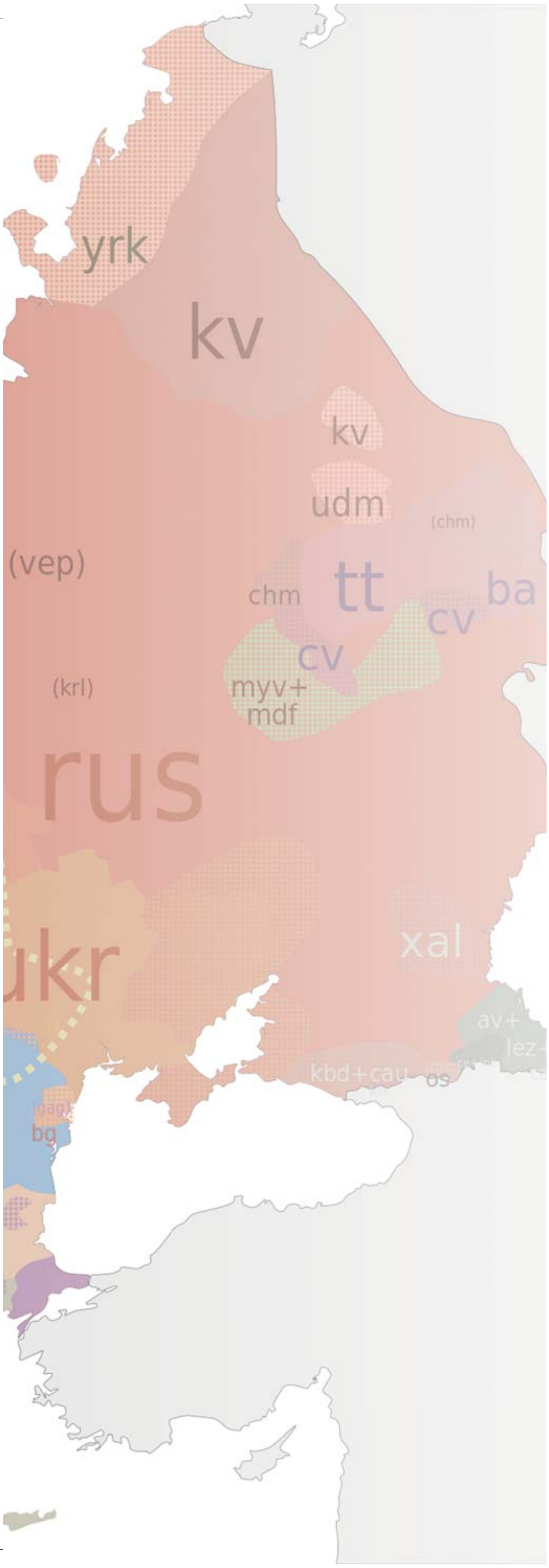
Statistically Speaking

Prof John Aston

Professor of Statistics, Statslab





It might seem somewhat implausible that functional analysis, non-Euclidean geometry and statistical shape analysis have much to tell us about the spread of European languages. Historical linguistics has traditionally been something of a qualitative discipline, but recently there has been considerable interest in taking a more quantitative approach to the subject, through textual analysis but also through the analysis of acoustic recordings. It is this latter data that has allowed some more unusual links between mathematics and phonetics to be made.

Acoustic recordings yield considerable quantities of data which to all intents and purposes can be seen as continuous over time. For example, in Figure 1 below, a two dimensional surface (spectrogram) can be seen, where the first axis represents time while the second represents the frequency of the sound wave being recorded. This spectrogram not only conveys all the time and frequency information contained in the word being said, but can be treated (when suitably normalised) as a random element, say X, $X \in L^2$.

Functional Data Analysis

The relatively new field of functional data analysis (FDA) (see [2], [4]) is something of a cross between functional analysis and classical statistics. Unlike the usual univariate or multivariate analysis undertaken in most statistics, FDA is the branch of statistics that concerns data where the fundamental unit of that data is a function in some suitable (often infinite dimensional Hilbert) space. The main idea is to use the properties of smoothness and regularity in the functions to allow statistical analysis to be carried out, even though the functions are only ever discretely observed with noise.

One of the most important quantities in FDA is

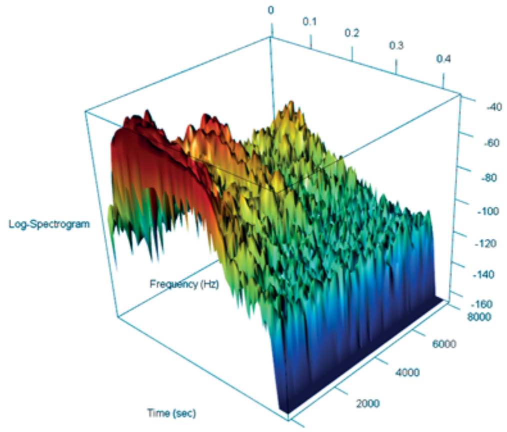


Figure 1 An example of a raw spectrogram (in logarithmic scale) as obtained by taking a windowed discrete fourier transform of a 22kHz sound sample of a single syllable (the word one ("un") in French). The fourier transform was computed every 10 ms to yield the discretised version of the function.

the covariance operator. For a random square integrable function X, with $\mathbb{E}(X)=0$, the operator $C(y)=\mathbb{E}(\langle X,y\rangle X), y \in L^2$ is defined as the covariance operator, where $\langle \cdot,\cdot \rangle$ is the usual inner product in L^2 (see [2] for more details). It is, by definition, non-negative definite, and in many data analysis situations is assumed to be a trace class operator, i.e. $\Sigma_j \lambda_j < \infty$, where λ_j are the eigenvalues of the spectral decomposition of the operator. In many situations, FDA proceeds by using one of the fundamental theorems of Stochastic Processes, the Karhunen-Loeve decomposition of the operator, to provide a basis for expansion of the data. This allows a possible dimension reduction on the data to be performed, something that has been common in multivariate statistics since the early 20th century, in the finite dimensional setting. However, in the case of examining differences between languages, it is the operator itself which will be of interest.

Assume that we are interested in understanding the relationship between languages through their acoustic properties. Given a set of recordings for a particular language, spectrograms can be produced and an estimate of the associated covariance operators obtained. Languages, of course, have many characteristics, but it has been shown that one characteristic of interest is the variational patterns that are present in the sounds. These differences are captured exactly by the covariance operators. Therefore by comparing covariance operators we can provide one particular comparison of the languages themselves.

Statistics in Non-Euclidean Spaces

However, covariance operators are not the usual type of data that statistical analysis is designed for. They are non-negative definite trace class operators, so do not lie in a standard "Euclidean" space. The usual Euclidean metrics used in statistical analysis, extended to FDA, are not valid given the restricted space. This requires a new type of metric to be used, one with its roots in statistical shape analysis (see [1]), where non-Euclidean geometry is commonplace.

Let us start by considering a closely related finite dimensional problem, defining a distance between two positive definite matrices. Possibly, the simplest approach to take would be to take the matrix logarithm and compute the usual Frobenius norm between the matrix logarithms. This is indeed a Riemannian distance on the space of positive definite matrices, and as such allows statistical analysis to be developed. However, even if our covariance operators were positive definite, their trace class nature implies that their eigenvalues tend to zero, and hence the equivalent of the matrix logarithm is unbounded. However, this is not the case for all transformations. For example, the square-root transformation is well defined and the resulting operator, while not guaranteed to be trace-class, is still a Hilbert-Schmidt operator, and as such the distances are still well defined.

The square-root of a matrix, or operator, is, however, not uniquely defined. It would be somewhat more elegant if the distance between two languages was independent of the choice of square-root. This is a well studied problem in statistical shape analysis, where the equivalent problem is that of how to match shapes that are subject to rotation and translation. The shape of dog is still a dog, whether it is standing with its head to the left or to the right. Equivalently the uniqueness of the square-root is defined up to its rotation, and as such by quotienting out the rotation group we obtain a unique distance. These ideas yield the following metric to measure the distances between our languages. For two covariances C_1 and C_2, the Procrustes metric is defined as

$$\begin{aligned} d_P(C_1,C_2)^2 &= \inf_{R\in O\{L^2(\Omega)\}} \| L_1 - L_2R \|_{\mathrm{HS}}^2 \\ &= \inf_{R\in O\{L^2(\Omega)\}} \operatorname{tr}\{(L_1 - L_2R)^*(L_1 - L_2R)\}, \end{aligned}$$

where L_i are such that $C_i = L_i L_i^*$, for i = 1, 2, and $O\{L^2(\Omega)\}$ is the space of unitary operators on L^2. Procrustes was the Greek innkeeper of myth who fitted everyone to his iron bed by either stretching or chopping them to size, and as such this metric equivalently gives a distance that disregards the orientations of the initial estimates of the covariance operator. This distance, although somewhat complex, has a simple closed form solution, where, for any choice of L_i satisfying the above,

$$d_P(C_1, C_2)^2 = \|L_1\|_{\text{HS}}^2 + \|L_2\|_{\text{HS}}^2 - 2\sum_{k=1}^{\infty} \sigma_k .$$

where σ_k are the singular values of the compact operator $L_2^* L_1$. It can be shown that, even when there are only finite amounts of discretised data present, the estimates of the distance converge asymptotically.

Investigating the Relationships in Romance Languages

The statistical analysis of non-Euclidean and functional data are of interest in and of themselves, and are some of the fastest growing areas of modern statistics. However, this is in many ways because of their ability to be used to give insights into other areas such as historical linguistics. In a recent study, recordings of the pronunciation of the numbers one to ten were taken from four different romance languages (French, Italian, Spanish and Portuguese) with one language having two different dialects present (Iberian Spanish and American Spanish). 219 spectrograms (there were several repetitions of each word in each language) were generated from the sound samples, and preprocessed to form aligned functions from which covariances were formed. The distances between these covariances were then examined.

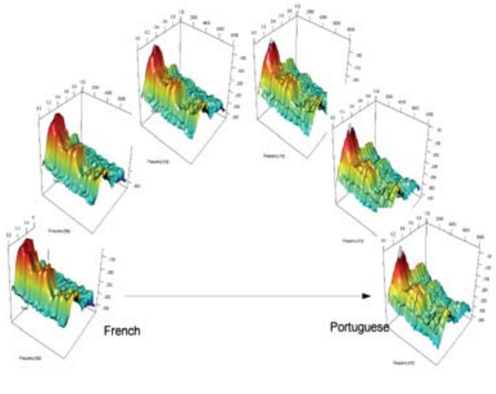


Figure 2 Representation of the geodesic taking a speaker saying the French word "un" and turning it into the Portuguese word "um". The geodesic is based on the Procrustes metric in the space of covariance functions.

It is possible to use the Procrustes metric to not only define distances between covariances but also by extension to define geodesics within the space of covariance functions (see [3]). These can then be used to define covariances for languages "between" any two of the observed languages or even to predict how one speaker might sound when speaking another language. Figure 2 shows one such predicted path. Here a speaker saying the word "un" (one in French) is mutated along a geodesic path into saying the word "um" (one in Portuguese). The speaker characteristics are retained but the variations attributed to the languages are captured via the geodesic path. These spectrograms can then be transformed back into audio to hear the results. This opens up a world of possibilities of discovering how one language might be related to another, or how historical language groups might have evolved into modern day languages.

The integration of concepts from geometry, analysis and other areas of mathematics into data analysis through statistics has a long history. However, modern data sources are constantly raising new challenges and areas such as non-Euclidean FDA are being developed for applications as diverse as brain imaging to those seen here in linguistics.

Acknowledgements

Considerable thanks are due to my collaborators John Coleman (Oxford), Ian Dryden (Nottingham), Pantelis Hadjipantelis (Warwick), Davide Pigoli (Cambridge) and Piercesare Secchi (Milan), for all their work upon which this article is based.

References

[1] Ian L. Dryden, Kanti V. Mardia; 1998; *Statistical Shape Analysis*; Wiley; Chichester.
[2] Horvath L, Kokoszka P; 2012; *Inference for Functional Data with Applications*; Springer, New York.
[3] Davide Pigoli, John A.D. Aston, Ian L. Dryden, Piercesare Secchi; 2014; *Distances and inference for covariance operators*; Biometrika; 101:409-422.
[4] James O Ramsay, Bernard W. Silverman; 2005; *Functional Data Analysis (2nd Ed)*; Springer, New York.