Pattern recognition by dinamic feature analysis based on PCA

Juliana Valencia Aguirre[1]
Andrés Marino Álvarez Mesa[2]
Genaro Daza Santacoloma[3]
Germán Castellanos Domínguez[4]

Resumen

Generalmente, en problemas de reconocimiento de patrones las observaciones son representadas por medio de medidas sobre un conjunto apropiado de variables, estas variables pueden clasificarse en estáticas y dinámicas. La representación estática no es siempre una aproximación precisa de las observaciones. En este sentido, algunos fenómenos son modelados de mejor manera por cambios dinámicos de sus medidas. La ventaja de emplear variables dinámicas radica en el hecho de incluir mayor cantidad de información que permita representar de mejor manera el conjunto de datos. Sin embargo, en etapas de clasificación es más difícil emplear variables dinámicas que estáticas, debido al costo computacional asociado. Con el fin de analizar este tipo de representaciones dinámicas es factible utilizar el Análisis de Componentes Principales (PCA), organizando los datos de manera que se puedan considerar las variaciones introducidas por la dinámica medida en las observaciones.

Por ende, el método que se propone permite evaluar la información dinámica de las observaciones en espacios de características de baja

[1] Estudiante Ingeniería Electrónica IX Semestre, Universidad Nacional de Colombia Sede Manizales, jvalenciaag@unal.edu.co.

[2] Estudiante Ingeniería Electrónica IX Semestre, Universidad Nacional de Colombia Sede Manizales, amalvarezme@unal.edu.co.

[3] Ingeniero Electrónico M. Eng, Universidad Nacional de Colombia Sede Manizales, Estudiante Doctorado en Ingeniería-Automática, gdazasa@unal.edu.co.

[4] Ingeniero en Radiocomunicaciones Ph.D, Universidad Nacional de Colombia Sede Manizales, Profesor Asociado, gcastellanosdo@unal.edu.co.

Fecha de recepción: 25 de marzo de 2009
Fecha de aceptación: 20 de mayo de 2009

dimensión sin deteriorar la precisión del sistema de clasificación. Los algoritmos fueron probados sobre datos reales en el reconocimiento de voces patológicas y normales; PCA se emplea también para seleccionar características dinámicas.

Abstract

Usually, in pattern recognition problems we represent the observations by mean of measures on appropriate variables of data set, these measures can be categorized as Static and Dynamic Features. Static features are not always an accurate representation of data. In these sense, many phenomena are better modeled by dynamic changes on their measures. The advantage of using an extended form (dynamic features) is the inclusion of new information that allows us to get a better representation of the object. Nevertheless, sometimes it is difficult in a classification stage to deal with dynamic features, because the associated computational cost often can be higher than we deal with static features. For analyzing such representations, we use Principal Component Analysis (PCA), arranging dynamic data in such a way we can consider variations related to the intrinsic dynamic of observations.

Therefore, the method made possible to evaluate the dynamic information about of the observations on a lower dimensionality feature space without decreasing the accuracy performance. Algorithms were tested on real data to classify pathological speech from normal voices, and using PCA for dynamic feature selection, as well.

Keywords

Principal component analysis, dynamic features, dynamic feature selection, pattern recognition.

1. Introduction

Generally, pattern recognition tasks deal with representations of observed objects by sets of numeric values (i.e., measurements). Such values are commonly known as features. Once these features are obtained, one can implement functions that assign the observed object to a specific class (set of objects with similar properties). In most of cases, these sets of features are assumed constant with respect to some associated dimension or dimensions (static features). Nevertheless, we can represent the objects to be classified by means of measures that do change over some associated dimension (dynamic features). Thus, instead of representing observations as single vectors whose entries are constant features, we may turn these entries into arrays that properly introduce the dynamic relation. The advantage of using such an extended form is the inclusion of new information that allows us to get a better representation of the object (i.e. the time behavior of a given signal) (Daza, 2006). The main idea in this paper is to extend traditional PCA (normally applied on static settings) to classification using dynamic representations.

It is possible to tackle the pattern recognition problem from an information theory point of view (Pentland & Turk, 1991). If we wanted to extract relevant information from a pattern as efficient as possible, then, we should encode it and compare it with a set of previously encoded observations. One approach for extracting the information contained in dynamic features is to capture the variations present in the observations set; and so uses this information to encode and compare with new patterns. It becomes natural to think that the process developed for face recognition well know as Eigenfaces (Pentland & Turk, 1991), can be extended to other types of objects, for instance, observations represented by dynamic features over time.

There are several approaches for deal with dynamic representations. In (Hall, Poskitt & Presnell, 2001) is proposed a method for dimensionality reduction and classification of functional data (observations with dynamic features); particularly, on radar signals. In the same way, (Ferraty & Vieu, 2003) present a non parametric method for functional data discrimination, which is a generalization for the method mentioned in (Hall et. al, 2001). Nevertheless, these techniques do not consider multivariate data, moreover they have an observation alignment constraint, that is, there should be some kind of synchronism among signals in training stage. This restriction is not desirable, especially on biosignals, because we assume that discriminant information is in the changing shape of variables (Silipo, Deco, Vergassola & Bartsch, 1998).

In this article is presented a process for reducing representations given by dynamic characterization, which can be outlined as:

1. Sorting the dynamic features in such a way the variances and covariances can be estimated among all points.
2. Computing PCA on the sorted array of elements representing the observations. The obtained eigenvectors span the basis of a subspace that embraces most of information given by a set of training observations.
3. Since eigenvectors span a subspace from an orthonormal basis, they can be used to project observation vectors. Therefore, making use of the weight vectors from this transformation as features that can be classified by typical algorithms.

All the steps previously mentioned are treated in detail throughout this paper. Finally, we test algorithms on real data to classify pathological speech from normal voices, using PCA for dynamic feature selection after some preliminary experimental results about different parameters (i.e. amount of retained information).

2. Feature Extraction

Feature extraction consists of a transformation of the original data space to a new data set with a reduced number of attributes. That is, feature extraction methods determine an appropriate subspace of dimension *m* (either in a linear or non linear way) in the input space of dimension p (being $m \leq p$). In this case, all available variables are used and the data are transformed to a space of reduced dimension. Thus, the aim is to replace the original variables by a smaller set of underlying characteristics. There are several reasons for feature extraction (Webb, 2002)(Jain, Duin & Mao, 2000)(Wang & Paliwal, 2003): 1) to provide a relevant set of features for a classifier, resulting in an improvement of the performance (especially for very simple classifiers); 2) to reduce the redundancy of the input data; 3) to recover the meaningful underlying variables or features that describe the observations, leading to better understanding of the data generation process; and 4) to produce a low-dimensional representation (ideally in two dimensions) with minimum loss of information (Daza et. al, 2009).

2.1 Dynamic Feature Extraction/Selection Using PCA

As it was previously mentioned in the introduction, dynamic features refer to numeric values that represent measures changing over some associated dimension (usually time or space). ξ_{ij} (t) Let be the j-th dynamic feature belonging to i-th observation; being n the number of observations and the number of features, which change over t =1,2..., T. We can represent each observation X_i by a matrix of size (T x p).

$$\mathbf{X}_i = \begin{bmatrix} \xi_{i1}(1) & \xi_{i2}(1) & \cdots & \xi_{ip}(1) \\ \xi_{i1}(2) & \xi_{i2}(2) & \cdots & \xi_{ip}(2) \\ \vdots & \vdots & & \vdots \\ \xi_{i1}(\mathrm{T}) & \xi_{i2}(\mathrm{T}) & \cdots & \xi_{ip}(\mathrm{T}) \end{bmatrix} \quad (1)$$

Our aim is to perform somehow PCA on the observations set, such that dynamic information can be extracted in the principal components. In order to achieve this, we need to dispose the training set in such a way we take account of all possible covariances among the t instants and dynamic variables. A natural way for it, is to represent each observation by Γ_i, which is a supervector of size (T x l) (see equation (2)).

$$\Gamma_i = \left[\xi_{i1}(1) \quad \xi_{i1}(2) \quad \cdots \quad \xi_{i1}(\mathrm{T}) \quad \xi_{i2}(1) \quad \cdots \quad \xi_{i2}(\mathrm{T}) \quad \cdots \quad \xi_{ip}(1) \quad \cdots \quad \xi_{ip}(\mathrm{T})\right]^T \qquad (2)$$

The mean vector $\overline{\Gamma}$ of the whole observation set is given by the following equation:

$$\overline{\Gamma} = \frac{1}{n}\sum_{i=1}^{n}\Gamma_i \qquad (3)$$

The covariance matrix **S** can be computed after centering each one of the observation supervectors by extracting $\overline{\Gamma}$ from each one of them. $\Phi_i = \Gamma_i - \overline{\Gamma}$ are the centered observations that allow us to calculate **s** from equation (4).

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}\Phi_i\Phi_i^T = \mathbf{G}\mathbf{G}^T \qquad (4)$$

where

$$\mathbf{G} = \left[\Phi_1 \quad \Phi_2 \quad \ldots \quad \Phi_n\right]$$

In most of cases, we are way far from computing the eigenvectors and eigenvalues of such a huge matrix. Nevertheless, we can make use of the rank properties of ; in special, the one that states $\mathbf{G}\mathbf{G}^T$ have the same non-null eigenvalues than $\mathbf{G}^T\mathbf{G}$. The relation between these two matrices is given by this simple yet useful trick (Pentland & Turk, 1991):

$$\mathbf{G}^T\mathbf{G}\hat{\mathbf{v}}_i = \lambda\mathbf{v}_i \qquad (5)$$

$\mathbf{v}_i$ are the eigenvectors of $\mathbf{G}^T\mathbf{G}$. Pre-multiplying by **G** on both sides of (5), we have that,

$$\mathbf{G}\mathbf{G}^T\mathbf{G}\hat{\mathbf{v}}_i = \lambda\mathbf{G}\mathbf{v}_i = \mathbf{S}\mathbf{G}\hat{\mathbf{v}}_i$$

Thence the eigenvectors corresponding to non-zero eigenvalues of are $\mathbf{v}_i = \mathbf{G}\hat{\mathbf{v}}_i / \left\|\mathbf{G}\hat{\mathbf{v}}_i\right\|$. We must also consider that **S** is positive semidefinite; and so the largest eigenvalues of (5) are the largest eigenvalues of **S**. The eigenvectors associated with the *k* largest eigenvalues of **S** are selected as Principal Directions, which span a subspace from an orthonormal basis that contains most of the information given by observations.

To construct that subspace containing most of the information from observations, we shall project centralized observations onto the chosen eigenvectors. In fact, by projecting these observations onto the eigenvectors' basis, we try to reproduce the observation in the original space as a linear combination of the principal directions as it can be seen in (6).

$$\hat{\Phi}_i = \sum_{k=1}^{m} \omega_k v_k \tag{6}$$

The reconstruction $\omega_k = v_k \Phi_i$ weights can be though as the new set of features and taking advantage of the orthonormality property of the basis we can classify observations using geometric criteria to partition the subspace off.

Using the magnitudes of the entries of the eigenvectors that span the representation basis, might tell us what variables are to be chosen (Jolliffe, 2002). Let be the vector given by:

$$\boldsymbol{\rho} = \sum_{\alpha=1}^{k} \left| \lambda_a \mathrm{v}_a \right| \tag{7}$$

rearranging in the following manner:

$$\boldsymbol{\rho} = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1T} & \rho_{21} & \cdots & \rho_{2T} & \cdots & \rho_{p1} & \cdots & \rho_{pT} \end{bmatrix}^{\mathrm{T}}$$

$$\Rightarrow \ \mathbf{P} = \begin{bmatrix} \rho_{11} & \rho_{21} & \cdots & \rho_{p1} \\ \rho_{12} & \rho_{22} & \cdots & \rho_{p2} \\ \vdots & \vdots & & \vdots \\ \rho_{1T} & \rho_{2T} & \cdots & \rho_{pT} \end{bmatrix}$$

To obtain the vector , which is the sum of the elements of each column of. Thus:

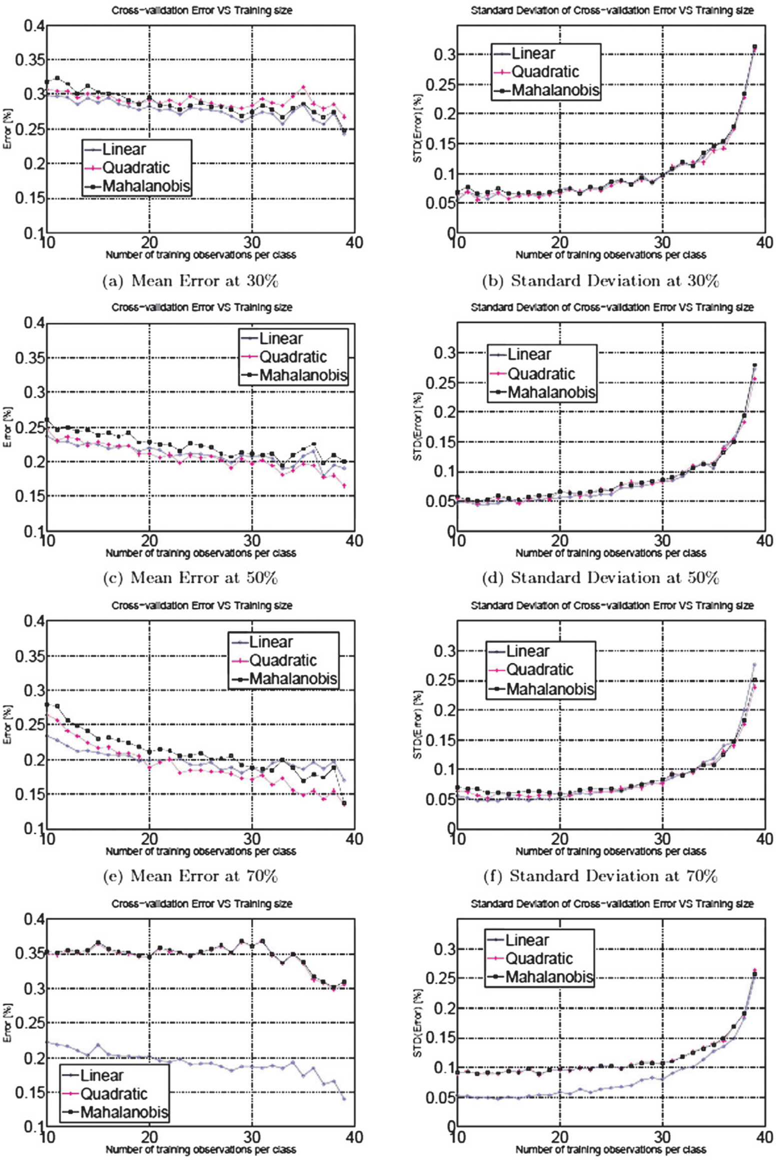
$$\mathbf{r} = \left[\sum_{t=1}^{T} \mathbf{P}_{1t}, \sum_{t=1}^{T} \mathbf{P}_{2t}, \cdots, \sum_{t=1}^{T} \mathbf{P}_{pt}, \right]^{\mathrm{T}} \tag{8}$$

and the assumption is that the largest entries of ***r*** point out the best features, since they present higher overall correlations with principal components.

3. Experimental Setup

Algorithms were tested on pathological speech database. A total of 90 observations (40 pathological and 50 normal voices), each one represented by 39 vectors corresponding to dynamic features over 110 time intervals. These 39 features are: 12 Mel Frequency Cepstral Coefficients (MFCC) and an Energy Coefficient, besides its first and second order derivatives are calculated.

The training was carried out using balanced groups, that is, the same number of observations per class. The accuracy of the classification is given in terms of cross-validation. Several training sets along with different amounts of retained variance were tested; Figure 1 shows the validation mean error and its variance vs. the size of training sets for 30%, 50%, 70%, and 90% of cumulative variance. Another experiment consists on determining whether a dynamic feature is relevant or not. For this experiment we choose a fixed number for training and retained variance (30 training samples per class and 70% of cumulative variance). The main idea is to perform PCA over all dynamic variables and create a queue by sorting the values of in descending order (see Figure 2). Then, PCA and classification are carried out on an incremental set of dynamic features. Figure 3 depicts the validation error obtained from incremental sets.



(a) Mean Error at 30%

(b) Standard Deviation at 30%

(c) Mean Error at 50%

(d) Standard Deviation at 50%

(e) Mean Error at 70%

(f) Standard Deviation at 70%

(g) Mean Error at 90%

(h) Standard Deviation at 90%

Figure 1: Cross-validation for several training sets along with different amounts of retained variance.

4. Discussion

Figure 1 shows the classifiers performance (linear, quadratic, and Mahalanobis). The mean error asymptotically decreases while the number of training samples per class is augmented. Moreover, standard deviation increases in the same way. The best classifier for a 30% cumulative variance was the linear classifier, according Figure 1(a). For a 50% cumulative variance the best classifier was the quadratic one. From Figure 1(e), for training sets greater than 15 samples per class, we can see how the classifier performance using 70% cumulative variance is better than 30% and 50% cumulative variance results. In this case, the quadratic classifier is the most accurate. For 90% cumulative variance (Figure 1(g)), linear classifier has a mean error similar to 70% cumulative variance results. The performance of quadratic and Mahalanobis classifiers get worse. When more than 30 training samples per class are used the standard deviation of the error grows strongly, which means that pattern recognition system is overtrained. In Table 1, the mean error results for 30 training samples per class are presented. The lowest average error for the three classifiers is obtained using 70% cumulative variance, and the best classifier is the quadratic.

Table 1: Cross Validation error and standard deviation (30 training samples per class)

Cumulative variance	Classifier	Mean error ± Standard deviation
30%	Linear Quadratic Mahalanobis	0.27 ± 0.095 0.28 ± 0.095 0.275 ± 0.095
50%	Linear Quadratic Mahalanobis	0.21 ± 0.08 0.196 ± 0.078 0.215 ± 0.085
70%	Linear Quadratic Mahalanobis	0.19 ± 0.076 0.172 ± 0.077 0.19 ± 0.078
90%	Linear Quadratic Mahalanobis	0.18 ± 0.078 0.365 ± 0.011 0.365 ± 0.011

From Figures 2, and 3 it can be seen how PCA not only gives us a way for representing dynamic information, it can be applied for dynamic feature selection, as well. These performance curves show that using around 9 of the 39 features (the ones selected by PCA) are enough for classify with almost the same accuracy. Mostly of selected features correspond to MFCC; it has been shown in literature that these coefficients are suitable for good representations of speech (Goldwasser et al, 1998) (Westwood, 1999, besides features that correspond to first and second derivatives are not relevant in accordance with the weights calculated.

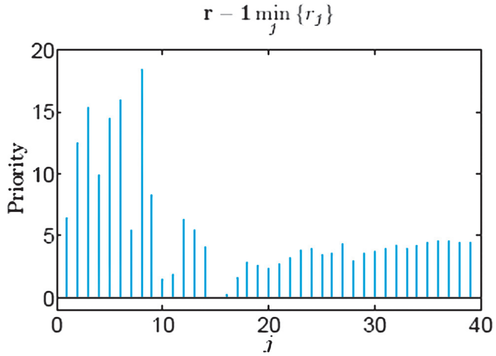
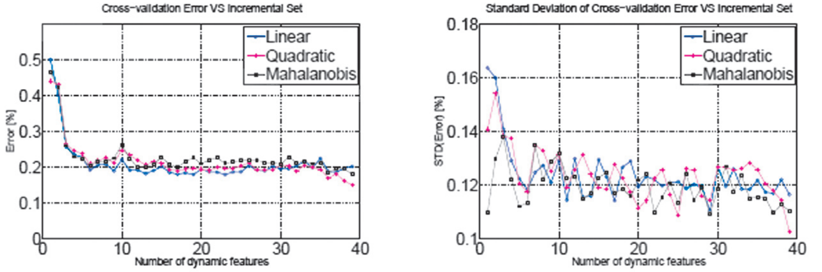


FIGURE 2: INCREMENTAL ORDER GIVEN BY r - 1 $\min_j \{r_j\}$.
WEIGHT OF THE *jth* DYNAMIC FEATURE.



(a) Incremental Performance mean Error

(b) Incremental Performance Standard Deviation of Error

FIGURE 3: INCREMENTAL PERFORMANCE (MEAN VALUE AND STANDARD DEVIATION).

Table 2 presents the mean error results increasing the number of original features. These features are sorted in descend order according to weights calculated from equation (8). The linear classifier exhibits the best performance.

Table 2: Incremental Performance. Mean Error and Standard Deviation.

Number of dynamic features	Mean error ± standard deviation		
	Linear	Quadratic	Mahalanobis
1	0,5 ± 0,163	0,44 ± 0,14	0,48 ± 0,11
2	0,4 ± 0,16	0,43 ± 0,153	0,42 ± 0,131
3	0,26 ± 0,141	0,27 ± 0,138	0,26 ± 0,138
4	0,23 ± 0,13	0,25 ± 0,137	0,22 ± 0,122
5	0,22 ± 0,122	0,24 ± 0,12	0,22 ± 0,116
6	0,19 ± 0,119	0,21 ± 0,118	0,21 ± 0,116
7	0,21 ± 0,126	0,22 ± 0,135	0,22 ± 0,135
8	0,21 ± 0,128	0,24 ± 0,133	0,22 ± 0,122
9	0,19 ± 0,121	0,22 ± 0,127	0,24 ± 0,129

5. Conclusion and Future Work

It was shown how PCA can be applied for analysis of dynamic features if observations are disposed, such that we can account for all covariances representing dynamic information in data. There is a compromise between classification accuracy, and the quantity of information that should be retained. On tests, best results were obtained for retaining 70% of cumulative variance and training sets around 30 samples per class. There are sundry techniques for dealing with dynamic features, but in the most of the cases these need to set a lot of free parameters, for example hidden Markov Models, Neural Networks, etc, without obtaining a significant improvement.

As future work, we would like to extend this approach by using other eigenrepresentations and perhaps the use of kernel methods for non-linear representation that would take into account higher order information instead of only using covariance.

References

Daza, 2006. Daza, G., (2006). *Metodología de reducción de dimensión para sistemas de reconocimiento automático de patrones sobre bioseñales*. Master's thesis, Universidad Nacional de Colombia.

Daza et. al, 2009. Daza, G., Arias, J., Godino, J., Sáenz, N., Osma, V., & Castellanos, G. (2009). Dynamic feature extraction: An application to voice pathology detection. *Intelligent automation and softcomputing*, vol. 15, No. 4, 665-680.

Duda, 2001. Duda, R., O., Hart, P., & Stork, D. (2001). *Pattern Classification.* United States of America: John Wiley and Sons.

Ferraty & Vieu, 2003. Ferraty, F., & Vieu, P. (2003). *Curves discrimination: A nonparametric functional approach.* Computational Statistics & Data Analysis.

Goldwasser et al, 1998. Goldwasser, L., Junqua, J. C., Kuhn, R., & Nguyen, P. (1998). Eigenfaces and eigenvoices dimensionality reduction for specialized pattern recognition. *In Proceedings of IEEE 1998 Workshop on Multimedia Signal Processing*.

Hall, Poskitt & Presnell. Hall, P., Poskitt, D., & Presnell, B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics*, 43(1):1–9.

Jain, Duin & Mao, 2000. Jain, A., Duin, R., & Mao, J. (2000) Statistical pattern recognition: A Review. *IEEE Tran*. PAMI, 22(1), pp. 4-37.

Jolliffe, 2002. Jolliffe, I. T. (2002). *Principal component analysis.* New York: Springer series.

Pentland & Turk, 1991. Pentland, A. P., & Turk, M. A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience,* 3(1).

Pentland & Turk, 1991. Pentland, A. P., & Turk, M. A. (1991). Face recognition using eigenfaces. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,* 586-591.

Silipo, Deco, Vergassola & Bartsch, 1998. Silipo, R., Deco, G., Vergassola, R., & Bartsch, H. (1998). Dynamics extraction in multivariate biomedical time series. *Biological Cybernetics*, 79:15–27.

Theodoridis, 2003. Theodoridis, S., & Konstantin's, K. (2003). *Pattern Recognition*. Athens: Academic Press.

Wang & Paliwal, 2003. Wang, X & Paliwal, K. (2003). *Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition*. Pattern recognition.

Webb, 2002. Webb, A. (2002). *Statistical pattern recognition*. Second ed., John Wiley & Sons Ltd.

Westwood, 1999. Westwood, R. (1999). *Speaker adaptation using eigenvoices*. Master's thesis, Cambridge University.