PRINCIPAL CURVES: LEARNING, DESIGN, AND APPLICATIONS

BALÁZS KÉGL

A THESIS
IN
THE DEPARTMENT
OF
COMPUTER SCIENCE

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
CONCORDIA UNIVERSITY
MONTRÉAL, QUÉBEC, CANADA

DECEMBER 1999

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mr. Balázs Kégl**

Entitled: **Principal Curves: Learning, Design, and Applications**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

_____ Chair

_____ External Examiner

_____ Examiner

_____ Examiner

_____ Examiner

_____ Supervisor

Approved _____

Chair of Department or Graduate Program Director

_____ 19 _____ _____

Dr. Nabil Esmail, Dean

Faculty of Engineering and Computer Science

To my parents and Agnès

Abstract

Principal Curves: Learning, Design, and Applications

Balázs Kégl, Ph.D.
Concordia University, 2005

The subjects of this thesis are unsupervised learning in general, and principal curves in particular. Principal curves were originally defined by Hastie [Has84] and Hastie and Stuetzle [HS89] (hereafter HS) to formally capture the notion of a smooth curve passing through the "middle" of a d-dimensional probability distribution or data cloud. Based on the definition, HS also developed an algorithm for constructing principal curves of distributions and data sets.

The field has been very active since Hastie and Stuetzle's groundbreaking work. Numerous alternative definitions and methods for estimating principal curves have been proposed, and principal curves were further analyzed and compared with other unsupervised learning techniques. Several applications in various areas including image analysis, feature extraction, and speech processing demonstrated that principal curves are not only of theoretical interest, but they also have a legitimate place in the family of practical unsupervised learning techniques.

Although the concept of principal curves as considered by HS has several appealing characteristics, complete theoretical analysis of the model seems to be rather hard. This motivated us to redefine principal curves in a manner that allowed us to carry out extensive theoretical analysis while preserving the informal notion of principal curves. Our first contribution to the area is, hence, a new *theoretical model* that is analyzed by using tools of statistical learning theory. Our main result here is the first known consistency proof of a principal curve estimation scheme.

The theoretical model proved to be too restrictive to be practical. However, it inspired the design of a new *practical algorithm* to estimate principal curves based on data. The polygonal line algorithm, which compares favorably with previous methods both in terms of performance and computational complexity, is our second contribution to the area of principal curves. To complete the picture, in the last part of the thesis we consider an *application* of the polygonal line algorithm to hand-written character skeletonization.

Acknowledgments

I would like to express my deep gratitude to my advisor, Adam Krzyżak, for his help, trust and invaluable professional support. He suggested the problem, and guided me through the stages of this research. My great appreciation goes to Tamás Linder for leading me through the initial phases of this project, for the fruitful discussions on both the theoretical and the algorithmic issues, and for his constant support in pursuing my ideas. I would also like to thank Tony Kasvand for showing me the initial directions in the skeletonization project.

Contents

List of Figures

List of Tables