

# 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

© BALÁZS KÉGL, 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 committee:

\_\_\_\_\_ 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

<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Unsupervised Learning . . . . .	1
1.1.1 The Formal Model . . . . .	4
1.1.2 Areas Of Applications . . . . .	4
1.1.3 The Simplest Case . . . . .	5
1.1.4 A More Realistic Model . . . . .	8
1.2 Principal Curves . . . . .	9
1.3 Outline of the Thesis . . . . .	10
<b>2 Vector Quantization and Principal Component Analysis</b>	<b>12</b>
2.1 Vector Quantization . . . . .	12
2.1.1 Optimal Vector Quantizer . . . . .	13
2.1.2 Consistency and Rate Of Convergence . . . . .	14
2.1.3 Locally Optimal Vector Quantizer . . . . .	15
2.1.4 Generalized Lloyd Algorithm . . . . .	16
2.2 Principal Component Analysis . . . . .	17
2.2.1 One-Dimensional Curves . . . . .	18
2.2.2 Principal Component Analysis . . . . .	22
2.2.3 Properties of the First Principal Component Line . . . . .	25
2.2.4 A Fast PCA Algorithm for Data Sets . . . . .	26
<b>3 Principal Curves and Related Areas</b>	<b>28</b>
3.1 Principal Curves with Self-Consistency Property . . . . .	28
3.1.1 The HS Definition . . . . .	28

3.1.2	The HS Algorithm for Data Sets . . . . .	31
3.1.3	The Bias of the HS Algorithm . . . . .	33
3.2	Alternative Definitions and Related Concepts . . . . .	36
3.2.1	Alternative Definitions of Principal Curves . . . . .	36
3.2.2	The Self-Organizing Map . . . . .	37
3.2.3	Nonlinear Principal Component Analysis . . . . .	42
<b>4</b>	<b>Learning Principal Curves with a Length Constraint</b>	<b>44</b>
4.1	Principal Curves with a Length Constraint . . . . .	44
4.2	Learning Principal Curves . . . . .	47
<b>5</b>	<b>The Polygonal Line Algorithm</b>	<b>56</b>
5.1	The Polygonal Line Algorithm . . . . .	56
5.1.1	Stopping Condition . . . . .	58
5.1.2	The Curvature Penalty . . . . .	59
5.1.3	The Penalty Factor . . . . .	60
5.1.4	The Projection Step . . . . .	61
5.1.5	The Vertex Optimization Step . . . . .	61
5.1.6	Convergence of the Inner Loop . . . . .	63
5.1.7	Adding a New Vertex . . . . .	64
5.1.8	Computational Complexity . . . . .	65
5.1.9	Remarks . . . . .	67
5.2	Experimental Results . . . . .	68
5.2.1	Comparative Experiments . . . . .	69
5.2.2	Quantitative Analysis . . . . .	71
5.2.3	Failure Modes . . . . .	77
<b>6</b>	<b>Application of Principal Curves to Hand-Written Character Skeletonization</b>	<b>82</b>
6.1	Related Work . . . . .	83
6.1.1	Applications and Extensions of the HS Algorithm . . . . .	83
6.1.2	Piecewise Linear Approach to Skeletonization . . . . .	85
6.2	The Principal Graph Algorithm . . . . .	85
6.2.1	Principal Graphs . . . . .	86
6.2.2	The Initialization Step . . . . .	92
6.2.3	The Restructuring Step . . . . .	94
6.3	Experimental Results . . . . .	100
6.3.1	Skeletonizing Isolated Digits . . . . .	100

6.3.2	Skeletonizing and Compressing Continuous Handwriting . . . . .	105
<b>7</b>	<b>Conclusion</b>	<b>109</b>
	<b>Bibliography</b>	<b>111</b>



# List of Figures

1	An ill-defined unsupervised learning problem . . . . .	2
2	Projecting points to a curve . . . . .	19
3	Geometrical properties of curves . . . . .	20
4	Distance of a point and a line segment . . . . .	22
5	The first principal component line . . . . .	25
6	Self-consistency . . . . .	29
7	Computing projection points . . . . .	32
8	The two sources of bias of the HS algorithm . . . . .	35
9	The flow chart of the polygonal line algorithm . . . . .	57
10	The evolution of the polygonal principal curve . . . . .	58
11	A nearest neighbor partition of induced by the vertices and segments of $\mathbf{f}$ . . . . .	62
12	The flow chart of the optimization step . . . . .	64
13	$\Delta'_n(\mathbf{f})$ may be less than $\Delta_n(\mathbf{f})$ . . . . .	64
14	The circle example . . . . .	70
15	The half circle example . . . . .	71
16	Transformed data sets . . . . .	72
17	Small noise variance . . . . .	73
18	Large sample size . . . . .	74
19	Sample runs for the quantitative analysis . . . . .	75
20	The average distance of the generating curve and the polygonal principal curve. . .	76
21	The relative difference between the standard deviation of the noise and the measured <i>RMSE</i> . . . . .	76
22	Failure modes 1: zig-zagging curves . . . . .	77
23	Correction 1: decrease the penalty parameter . . . . .	78
24	Failure modes 2: complex generating curves . . . . .	80
25	Correction 2: “smart” initialization . . . . .	81
26	Representing a binary image by the integer coordinates of its black pixels . . . . .	86

27	Results on characters not containing loops or crossings . . . . .	87
28	The flow chart of the extended polygonal line algorithm . . . . .	88
29	Evolution of the skeleton graph . . . . .	89
30	Roles of vertices of different types . . . . .	90
31	Examples of transforming the skeleton into an initial skeleton graph . . . . .	95
32	Paths, loops, simple paths, branches, and deletion . . . . .	96
33	The role of the angle in deleting short branches . . . . .	97
34	Deleting short branches . . . . .	98
35	Removing short loops . . . . .	99
36	Removing a path in merging star3-vertices . . . . .	100
37	Merging star3-vertices . . . . .	101
38	Removing a line-vertex in the filtering operation . . . . .	101
39	Filtering vertices . . . . .	102
40	Skeleton graphs of isolated 0's . . . . .	102
41	Skeleton graphs of isolated 1's . . . . .	102
42	Skeleton graphs of isolated 2's . . . . .	103
43	Skeleton graphs of isolated 3's . . . . .	103
44	Skeleton graphs of isolated 4's . . . . .	103
45	Skeleton graphs of isolated 5's . . . . .	103
46	Skeleton graphs of isolated 6's . . . . .	104
47	Skeleton graphs of isolated 7's . . . . .	104
48	Skeleton graphs of isolated 8's . . . . .	104
49	Skeleton graphs of isolated 9's . . . . .	104
50	Original images of continuous handwritings . . . . .	105
51	Skeleton graphs of continuous handwritings . . . . .	106

# List of Tables

1	The relationship between four unsupervised learning algorithms . . . . .	68
2	The average radius and <i>RMSE</i> values . . . . .	75
3	Vertex types and their attributes . . . . .	92
4	Vertex degradation rules . . . . .	93
5	Length thresholds in experiments with isolated digits . . . . .	100
6	Length thresholds in experiments with continuous handwriting . . . . .	105
7	Compression of Alice's handwriting . . . . .	107
8	Compression of Bob's handwriting . . . . .	108