Generalized Principal Component Analysis

Estimation & Segmentation of Hybrid Models

René Vidal (Johns Hopkins University)
Yi Ma (University of Illinois at Urbana-Champaign)
S. Shankar Sastry (University of California at Berkeley)

January 16, 2006

Copyright ©2004 Reserved
No parts of this draft may be reproduced without written permission from the authors.
This is page v
Printer: Opaque this

To be written (R.V.)

To be written (Y.M.)

To be written (S.S.S.)
Preface

In the past few years, we have encountered a variety of important problems that arise in computer vision, image processing, pattern recognition, and systems identification, that can be abstracted to a common mathematical problem: Given a set of data points sampled from a “mixture” of unknown geometric or statistical models, how to automatically learn or infer these models? The word “mixture” means that the data points are clustered into different groups, each of which belongs to a different model. In the literature, in different contexts, such data sets are sometimes referred to as “mixed,” or “multi-modal,” or “multi-model,” or “heterogeneous,” or “hybrid.” In this book, we have settled on the usage of the expression “mixed data” and the associated model as a “hybrid model.” Thus, a solution to the problem normally requires to segment the data into groups, each belonging to one of the constituent models and to then estimate the parameters of each model. In the case that each of the constituent models are linear, the foregoing problem is reduced to one of fitting multiple low-dimensional linear subspaces to the set of sample points in a high-dimensional space.

The main goal of this book is to introduce a new method to study hybrid models, which we refer to as generalized principal component analysis, with the acronym GPCA.\textsuperscript{1} The general problems that GPCA aims to address represents

\textsuperscript{1}In the literature, the word “generalized” is sometimes used to indicate different extensions to the classical principal component analysis (PCA) [Jolliffe, 1986]. In our opinion these are indeed extensions rather than a more extensive generalization that we propose in this book. Additionally, for the “nonlinear” case when each component is an algebraic variety of higher degree such as a quadratic surface or a more complicated manifold, we may still use the same term GPCA. Other names like “hybrid component analysis” (HCA) have also been suggested and would also be appropriate.
a fairly general class of unsupervised learning problems — many data clustering and modeling methods in machine learning can be viewed as special cases of this method.\(^2\)

A main difficulty associated with estimation of a hybrid model is that, without knowing which subset of sample points belong to which constituent model it is not possible to determine the model that this group belongs to. Thus, there is seemingly a “chicken-and-egg” relationship between data segmentation and model estimation: If the segmentation of the data was known, one could easily fit a single model to each subset of samples using classical model estimation techniques; and conversely, if the models were known, one could easily find the subset of samples that best fit each model. This relationship has been the intuitive and heuristic justification for many iterative techniques for estimating a hybrid (or mixture) model, such as the well-known expectation maximization (EM) algorithm. This book aims to provide a non-iterative and general solution to the problem of simultaneously grouping and model fitting the data, based on a repertoire of new tools drawn from a novel and somewhat unconventional source in the statistics literature: algebraic geometry – mainly polynomial and linear algebra. The immediate reaction of a statistician to this statement is that the methods of classical algebraic geometry are extremely sensitive to noise (or outliers) and modeling uncertainties and are therefore not robust in problems of estimation and model fitting. However, we will show how to combine these new tools with traditional statistical techniques so as to obtain both robust and efficient algorithms for simultaneous data segmentation and model estimation.

There are several reasons why we felt that it was the right time to write a book on estimation of hybrid models. First, the conventional single-model paradigm has shown to have significant limitations in numerous important emerging applications. For example, in image processing, it is well known that to achieve a more economic (compressed) representation of images, different images or different regions of the same image, it is best to represent either the regions of a single image or a class of images adaptively using different sets of linear bases. However, the traditional image processing doctrine advocates the use of a prefixed or prechosen set of bases (for e.g., the discrete cosine transform or a wavelet transform). This approach is largely because segmentation of the image and identification of adaptive bases are difficult problems — they belong to a hybrid model. Yet another example arises from systems theory: classical system identification theory and techniques are of limited use for a system that switches among multiple control systems, called a hybrid linear system. New identification theory and algorithms need to be developed for such systems.

Second, although the classical model estimation techniques, for a single model, have been well studied in statistics (see Appendix C), techniques specifically customized for hybrid models, even for hybrid linear models, has not been com-

---

\(^2\)Classical clustering analysis can be viewed as the specialization of GPCA to the case that each subspace is of dimension one, as we will explain in Section 3.2.
pletely understood and thoroughly developed. Designing a working algorithm, for
many practical problems of this nature is currently a matter of intuition and clever
heuristics: almost a work of art from application from application! Of course, as
a result it is difficult to abstract the lessons from one context and use it in another.

Finally, in our opinion even though attempts have been made in the past to
study and solve many special cases of hybrid models, for instance in statistical
learning, pattern recognition, and system identification, there has never been an
attempt to unify and truly generalize the results in a unified framework that is able
to encompass all aspects of the problem. For instance, both the so called Expecta-
tion Maximization (EM) method and the K-means method have been proposed
to resolve the “chicken-and-egg” difficulty between data segmentation and model
estimation. However, these methods resort primarily to an incremental and iter-
ative scheme that starts from a random initialization. They therefore are prone
to converge to local minima and lack of efficiency. Only recently has GPCA, a
new algebro-geometric approach, been developed that offers a global and non-
iterative solution to this problem. Not only does this new approach lead to simple
and effective algorithms that do not have the same difficulty as the existing meth-
ods, but it also offers new insights to modeling mixed data with a hybrid model.
This book gives an introduction to these new findings, together with their relations
to extent methods.

Thus, the primary goal of this book is to provide a comprehensive introd-
cution of the fundamental statistical, geometric and algebraic concepts associated
with the estimation (and segmentation) of the hybrid models, especially the
hybrid linear models. Traditionally, data modeling and model estimation has
been primarily a topic of study in statistics.3 We contend in this book that the
algebraic and geometric properties are equally (if not more) important for a
complete understanding of hybrid models.4 These properties are largely comple-
mentary and indeed compatible with their statistical counterparts. They together
are responsible for the development of more efficient algorithms and solutions.
Source codes for all the algorithms are available at the official GPCA website:
http://perception.cs1.uiuc.edu/gpca.

The topics covered in this book are of great relevance and importance to both
theoreticians and practitioners in the areas of statistical learning, pattern recog-
nition, computer vision, signal/image processing, and systems identification (in
many fields including control theory, econometrics, and finance). The applica-
tions given in this book highlight our own experiences which are directly relevant
for researchers who work on practical areas such as image representation & seg-
mentation, motion segmentation & estimation from video data, and hybrid system
identification. Source codes for these applications are also available at the official

---

3We will summarize almost all statistical facts used in this book in Appendix C.
4We will develop many of the properties throughout this book, especially those that are di-
rectly useful for algorithm development. Some of the more abstract properties will be summarized
in Appendix A and B.
GPCA website. Nevertheless, even a cursory examination of the literature shows that the number of other application domains is virtually limitless.

We have written this book with the idea that it should be friendly to teachers and students. As the reader may found out that hybrid model estimation is a very unique subject which touches upon many fundamental principles and concepts in both statistics and algebra. These principles and concepts can now be learned or taught in the same context so that their strengths and limitations can be better assessed and understood.

Organization of the Book.

Part I of this book develops the fundamental theory and basic algorithms for the identification and estimation of hybrid linear models. The chapters in this part systematically extend classical principal component analysis (PCA) for a single linear subspace, also known as the Karhunen-Loève (KL) expansion, to the case of a subspace arrangement. We begin after an introduction in Chapter 1, with a review of PCA: from a geometric, statistical and robustness standpoint and its extensions in Chapter 2. In Chapter 3, we start to study the problem of modeling data with subspace arrangements. The focus is primarily on existing iterative subspace segmentation methods such as EM and K-Means (or K-Subspaces in our context). These methods are based on either geometric intuition or statistical inference. In Chapter 4 we develop a non-iterative algebraic method for the estimation and segmentation of subspace arrangements, i.e., the Generalized Principal Component Analysis (GPCA) algorithm. Statistical considerations and robustness issues which provide the link between the algebraic techniques and some traditional statistical methods are given in Chapter 5.

Parts II of this book provide a few case studies of real-world problems. The problems are selected from image processing and computer vision. They include image representation & segmentation (Chapter 6), 2-D and 3-D motion segmentation (Chapter 7 and 8), and segmentation of dynamical textures and videos. Some of the case studies are straightforward application of the proposed algorithms, while others require certain more elaborate justification and special domain knowledge. We hope that these case studies will inspire the reader to discover new applications for the general concepts and methods introduced in this book.

Part III extends the theory and application to more complicated hybrid models, namely hybrid dynamical systems and nonlinear models. In particular, Chapter 10 and 11 study the identification of hybrid linear systems and Chapter 12 studies nonlinear extensions of the basic GPCA algorithm to arrangements of quadratic surfaces and other manifolds.

To make the book self-contained, we have summarized relevant notations, concepts and results in mathematical statistics and algebraic geometry in Appendix A, B, and C. They may come by handy for readers who are not so familiar with certain mathematical facts used in the book, especially for the early chapters of the book.
Preface xi

René Vidal, Baltimore, Maryland
Yi Ma, Champaign, Illinois
Shankar Sastry, Berkeley, California
Spring, 2006
Acknowledgments

The seeds of Generalized Principal Component Analysis (GPCA) can be traced back to year 2001 when René and Yi started to work together on a polynomial-based method for solving the motion segmentation problem in computer vision. We soon realized that the important concepts and ideas behind this method can be extended and applied to solving many other problems with a similar nature. That observation has encouraged us to work very actively for the past few years to develop a more complete theoretical and algorithmic paradigm for this method. René has summarized in his PhD thesis at Berkeley much of the work as of May 2003. About one year later, on the day after Yi’s wedding reception, we decided to formalize our findings with a manuscript and sketched an outline of this book at Café Kopi in downtown Champaign.

Following René’s PhD thesis and our earlier papers, many of our graduate students have studied and extended GPCA to many new problems in computer vision, image processing, and system identification. We especially thank Jacopo Piazzi of Johns Hopkins, Kun Huang now at the Biomedical and Informatics Department of Ohio State University, Wei Hong, Yang Yang, Shankar Rao, Andrew Wagner, and John Wright of UIUC. Their research projects have led to many exciting theoretical results, applications, and examples presented in this book. They have also helped develop and maintain the official GPCA website at: http://perception.csl.uiuc.edu/gpca/.

We are greatly indebted to Professor Robert Fossum at the Mathematics Department of the University of Illinois at Urbana-Champaign and Professor Harm Derksen at the Mathematics Department of the University of Michigan. They have painstakingly taught Yi algebraic geometry and have helped giving a rather complete characterization of the algebraic properties of subspace arrangements that
are related to GPCA. Their results have helped providing a rigorous mathematical footing for algorithms developed in this book. Professor Robert Fossum has also helped proofreading the manuscript.

We thank many of our colleagues for valuable collaborations, discussions, suggestions, and moral support. They are Professor Stefano Soatto of the University of California at Los Angeles, Professor Jana Kosecka of George Mason University, Professor Richard Hartley of Australia National University, Professor Jitendra Malik of the University of California at Berkeley, Dr. Harry Shum – the director of Microsoft Research in Asia, and Professor Minh Do, Professor Thomas Huang, Professor Narendra Ahuja, Professor Yizhou Yu of the University of Illinois at Urbana-Champaign.

We also thank Professor Martin Vetterli of Ecole Polytechnique Fédérale de Lausanne and Professor David Donoho of Stanford University for their encouragement for us to apply GPCA to sparse image representation.

We are obviously grateful for all the funding agencies that have supported us. In particular, we would like to thank Dr. Helen Gill and Dr. Haesun Park of the National Science Foundation and Dr. Behzad Kamgar-Parsi of the Office of Naval Research. They have generously supported our research in this direction even when much of the theory and results were still at their infancy. Without their vision and trust, this book would not have been possible.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>vii</td>
<td></td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>xii</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Modeling Data with a Parametric Model</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1</td>
<td>The Choice of a Model Class</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2</td>
<td>Statistical Models versus Geometric Models</td>
<td>3</td>
</tr>
<tr>
<td>1.2</td>
<td>Modeling Mixed Data with a Hybrid Model</td>
<td>5</td>
</tr>
<tr>
<td>1.2.1</td>
<td>Examples of Mixed Data Modeling</td>
<td>6</td>
</tr>
<tr>
<td>1.2.2</td>
<td>Mathematical Representations of Hybrid Models</td>
<td>10</td>
</tr>
<tr>
<td>1.2.3</td>
<td>Noise, Outliers, and Model Selection</td>
<td>14</td>
</tr>
</tbody>
</table>

I Theory, Analysis, and Algorithms 17

| 2 | Data Modeling with a Single Subspace | 19 |
| 2.1 | Principal Component Analysis (PCA) | 19 |
| 2.1.1 | A Geometric Approach to PCA | 20 |
| 2.1.2 | A Statistical View of PCA | 22 |
| 2.1.3 | Determining the Number of Principal Components | 24 |
| 2.2 | Robustness Issues for PCA | 25 |
| 2.2.1 | Outliers | 25 |
| 2.2.2 | Incomplete Data Points | 27 |
2.3 Extensions to PCA
  2.3.1 Nonlinear PCA
  2.3.2 Kernel PCA
  2.4 Bibliographic Notes

3 Iterative Methods for Multiple-Subspace Segmentation
  3.1 Unsupervised Learning Methods for Data Clustering
    3.1.1 K-Means
    3.1.2 Expectation Maximization (EM)
  3.2 Problem Formulation of Subspace Segmentation
    3.2.1 Projectivization of Affine Subspaces
    3.2.2 Subspace Projection and Minimum Representation
  3.3 Subspace-Segmentation Algorithms
    3.3.1 K-Subspaces
    3.3.2 Expectation Maximization for Subspaces
    3.3.3 Relationships between K-Subspaces and EM
  3.4 Bibliographic Notes

4 Algebraic Methods for Multiple-Subspace Segmentation
  4.1 Introductory Cases of Subspace Segmentation
    4.1.1 Segmenting Points on a Line
    4.1.2 Segmenting Lines on a Plane
    4.1.3 Segmenting Hyperplanes
  4.2 Knowing the Number of Subspaces
    4.2.1 An Introductory Example
    4.2.2 Fitting Polynomials to Subspaces
    4.2.3 Subspaces from Polynomial Differentiation
    4.2.4 Point Selection via Polynomial Division
    4.2.5 The Basic Generalized PCA Algorithm
  4.3 Not Knowing the Number of Subspaces
    4.3.1 Introductory Examples
    4.3.2 Segmenting Subspaces of Equal Dimension
    4.3.3 Segmenting Subspaces of Different Dimensions
    4.3.4 The Recursive GPCA Algorithm
  4.4 Relationships between GPCA, K-Subspaces, and EM
  4.5 Bibliographic Notes
  4.6 Exercises

5 Statistical Techniques and Robustness Issues
  5.1 Estimation of the Vanishing Polynomials
    5.1.1 Sampson Distance
    5.1.2 Generalized Eigenvector Fit
  5.2 Linear Discriminant Analysis
    5.2.1 Fisher Linear Discriminant Analysis (LDA)
    5.2.2 Fisher Discriminant Analysis for Subspaces
5.3 Estimation of the Subspaces via a Voting Scheme .......................... 95
5.3.1 Stacks of Bases and Counters ........................................ 95
5.3.2 Tally the Votes of the Subspaces .................................... 96
5.3.3 Simulation Results ....................................................... 98
5.4 Model Selection for Multiple Subspaces ................................. 99
5.4.1 Minimum Effective Dimension ............................... 100
5.4.2 Hilbert Function for Model Selection ............. 104
5.5 Robust Estimation of Multiple Subspaces .............................. 107
5.5.1 The Sample Influence Function ............................. 108
5.5.2 First Order Approximation of the Sample Influence .... 108
5.5.3 MultivariateTrimming ............................................. 110
5.5.4 Simulation Comparison ......................................... 111
5.6 Bibliographic Notes .................................................. 115

II Applications in Image Processing & Computer Vision 117

6 Image Representation, Segmentation & Classification 119
6.1 Lossy Image Representation ........................................ 119
6.1.1 A Hybrid Linear Model ......................................... 122
6.1.2 Multi-Scale Hybrid Linear Models ........................ 128
6.1.3 Experiments and Comparisons .............................. 132
6.1.4 Limitations .......................................................... 135
6.2 Multi-Scale Hybrid Linear Models in Wavelet Domain ...... 136
6.2.1 Imagery Data Vectors in Wavelet Domain ........ 136
6.2.2 Estimation of Hybrid Linear Models in Wavelet Domain 138
6.2.3 Comparison with Other Lossy Representations ........ 139
6.2.4 Limitations .......................................................... 140
6.3 Image Segmentation .................................................. 141
6.3.1 Hybrid Linear Models for Image Segmentation ........ 141
6.3.2 Dimension and Size Reduction .............................. 144
6.3.3 Experiments .......................................................... 145
6.4 Image Classification .................................................. 145
6.5 Bibliographic Notes .................................................. 147

7 2-D Motion Segmentation from Image Partial Derivatives 148
7.1 An Algebraic Approach to Motion Segmentation .......... 149
7.2 The Multibody Brightness Constancy Constraint .... 150
7.3 Segmentation of 2-D Translational Motion Models ........ 153
7.3.1 The Multibody Optical Flow ................................ 153
7.3.2 Computing the 2-D Translational Model Parameters . 153
7.4 Segmentation of 2-D Affine Motion Models ............... 154
7.4.1 The Multibody Affine Matrix ............................... 155
7.4.2 Computing the Number of 2-D Affine Motion Models 156
7.4.3 Computing the 2-D Affine Motion Model Parameters 158
7.5 Segmentation of Motions Models of Different Type . . . . . . 159
  7.5.1 The Multibody Motion Matrix .......................... 160
  7.5.2 Computing the Number of 2-D Motion Models . . . . . . 160
  7.5.3 Computing the Type of 2-D Motion at Each Pixel . . . . . 162
  7.5.4 Computing the 2-D Motion Model Parameters . . . . . . 163
7.6 Algorithm Summary ............................................ 164
7.7 Experimental Results ........................................... 164
  7.7.1 Simulation Results ........................................ 166
  7.7.2 2-D Translational Motions ................................. 167
  7.7.3 2-D Affine Motions ........................................ 168
  7.7.4 2-D Translational and 2-D Affine Motions ................ 168
7.8 Bibliographic Notes ............................................. 171
7.9 Exercises .................................................... 175

8 3-D Motion Segmentation from Point Correspondences 177
  8.1 The Motion Estimation Problem ............................... 178
    8.1.1 Rigid-Body Motions and Camera Projection Models . 178
    8.1.2 The Fundamental Matrix ................................. 179
    8.1.3 The Homography Matrix .................................. 180
    8.1.4 The Trifocal Tensor ...................................... 181
  8.2 The Motion Segmentation Problem ............................ 182
  8.3 Segmentation of Linear Motion Models ....................... 183
    8.3.1 The Affine Motion Subspaces ............................ 183
    8.3.2 Segmentation of Motion Affine Subspaces ............... 183
  8.4 Segmentation of Bilinear Motion Models ...................... 185
    8.4.1 Segmentation of Fundamental Matrices .................. 185
    8.4.2 Segmentation of Homography Matrices ................... 187
  8.5 Segmentation of Trilinear Motion Models ..................... 191
    8.5.1 The Multibody Trifocal Tensor .......................... 193
    8.5.2 Segmentation of Trifocal Tensors ....................... 195
  8.6 Experimental Results ....................................... 197
    8.6.1 Segmentation of Affine Motion Subspaces ............... 197
    8.6.2 Segmentation of Fundamental Matrices .................. 200
    8.6.3 Segmentation of Homography Matrices ................... 202
  8.7 Bibliographical Notes ...................................... 203
  8.8 Exercises ................................................ 204

9 Dynamical Texture and Video Segmentation 208
  9.1 Modeling Dynamic Scenes with Dynamical Systems ............ 209
    9.1.1 Dynamic Textures Seen by a Static Camera .............. 210
    9.1.2 Dynamic Textures Seen by a Moving Camera ............. 212
  9.2 Spatial Video Segmentation ................................ 213
    9.2.1 Segmenting Dynamic Textures ........................... 213
    9.2.2 Segmenting Moving Dynamic Textures .................... 214
### Contents

9.3 Temporal Video Segmentation .................................. 215  
  9.3.1 Modeling Videos with Hybrid Linear Systems .... 215  
  9.3.2 Video Segmentation via Hybrid System Identification 216  
  9.3.3 Experimental Results ................................. 217  
9.4 Spatial-Temporal Video Segmentation ...................... 217  
9.5 Bibliographical Notes ................................. 217  
9.6 Exercises ........................................ 218  

### III Applications in System Biology 219

### IV Extensions to Arrangements of Dynamical Systems  
and Nonlinear Varieties 221

10 Switched ARX Systems ..................................... 223  
  10.1 Problem Statement ..................................... 225  
  10.2 Identification of a Single ARX System ............ 226  
  10.3 Identification of Hybrid ARX Systems ............ 229  
     10.3.1 The Hybrid Decoupling Polynomial ............. 230  
     10.3.2 Identifying the Hybrid Decoupling Polynomial 231  
     10.3.3 Identifying System Parameters and Discrete States 234  
     10.3.4 The Basic Algorithm and its Extensions .... 236  
  10.4 Simulations and Experiments .......................... 237  
     10.4.1 Error in the Estimation of the Model Parameters 238  
     10.4.2 Error as a Function of the Model Orders .... 238  
     10.4.3 Error as a Function of Noise ............... 239  
     10.4.4 Experimental Results on Test Datasets ...... 240  
  10.5 Bibliographic Notes ................................. 241  

11 Switched ARMA Models ......................................... 243

12 Extensions to Arrangements of Nonlinear Surfaces .... 244  
  12.1 Arrangements of Quadratic Surfaces .............. 244  
     12.1.1 Problem Formulation .......................... 245  
     12.1.2 Properties of the Fitting Polynomials .... 246  
     12.1.3 Generalized Principal Surface Analysis (GPSA) 249  
     12.1.4 Variations to the Basic GPSA Algorithm .... 252  
     12.1.5 Experiments on Synthetic Data .............. 254  
  12.2 Other Nonlinear Extensions ............................ 255  
  12.3 Bibliographic Notes ................................. 255  

### V Appendices 257  

A Basic Facts from Algebraic Geometry 259
### Contents

A.1 Polynomial Ring .................................................. 259  
A.2 Ideals and Algebraic Sets ........................................ 261  
A.3 Algebra and Geometry: Hilbert’s Nullstellensatz .............. 263  
A.4 Algebraic Sampling Theory ...................................... 264  
A.5 Decomposition of Ideals and Algebraic Sets ................. 267  
A.6 Hilbert Function, Polynomial, and Series .................... 268  

B Algebraic Properties of Subspace Arrangements .......... 270  
B.1 Ideals of Subspace Arrangements ............................... 271  
B.2 Subspace Embedding and PL-Generated Ideals ................ 272  
B.3 Hilbert Function of Subspace Arrangements ................. 275  
  B.3.1 Hilbert Function and GPCA ................................ 275  
  B.3.2 Special Cases of Hilbert Functions ....................... 278  
  B.3.3 Computation of Hilbert Function ......................... 280  
B.4 Bibliographic Notes ............................................. 282  

C Basic Facts from Mathematical Statistics ................ 284  
C.1 Estimation of Parametric Models ............................... 284  
  C.1.1 Uniformly Minimum Variance Unbiased Estimates ......... 286  
  C.1.2 Maximum Likelihood Estimates ............................ 287  
  C.1.3 Estimates from a Large Number of Samples ............. 288  
C.2 Expectation Maximization ..................................... 290  
C.3 Estimation of Mixture Models ................................ 293  
  C.3.1 Maximum-Likelihood Estimates ............................ 293  
  C.3.2 Minimax Estimates .......................................... 293  
C.4 Model Selection Criteria ...................................... 294  
  C.4.1 Akaike and Bayesian Information Criteria .............. 295  
  C.4.2 Model Selection Criteria for PCA ......................... 297  
C.5 Robust Statistical Methods .................................... 298  
  C.5.1 Sample Influence ............................................ 299  
  C.5.2 Robust Covariance Estimation ............................ 301  
  C.5.3 Random Sampling Techniques .............................. 302  

References ......................................................... 306
Chapter 1
Introduction

The primary goal of this book is to study how to model a data set that consists of multiple subsets with each drawn from a different primitive model. In different contexts, such a data set is sometimes referred to as “mixed,” “multi-modal,” “multi-model,” “piecewise,” “heterogeneous,” or “hybrid.” To unify the terminology, in this book, we will refer to such data as “mixed data” and the model used to fit the data as a “hybrid model.” Thus, a hybrid model typically consists of multiple constituent (primitive) models. Modeling mixed data with a hybrid model implies grouping the data into multiple (disjoint) subsets and fitting each subset with one of the constituent models. In the literature, the words “group,” “cluster,” “partition,” “decompose,” or “segment” are often used interchangeably. However, in this book, we will use the words “group,” “cluster,” or “partition” primarily for the data points,¹ and use the words “decompose” or “segment” for the associated models.²

An ever growing number of problems that arise today in computer vision, image processing, pattern recognition, system identification or system biology requires us to model mixed data. The techniques in this book are most evolved from studying the specific case of modeling data with an arrangement of subspaces.³

¹For instance, we may say “group (or cluster or partition) the data into multiple subsets,” or “a group (or a cluster) of sample points.”
²For instance, we may say “decompose (or segment) a hybrid model into its constituent primitive models.”
³In this book, we will use interchangeably “a mixture,” “a collection,” “a union,” or “an arrangement” of subspaces or models. But be aware that, in the case of subspaces, the formal terminology in Algebraic Geometry is “an arrangement of subspaces.”
also called hybrid (piecewise) linear models, with algorithms that we refer to as
*generalized principal component analysis* (GPCA). We also discuss the development
of techniques for hybrid quadratic models and more general hybrid models.
We also discuss extensions to hierarchical and recursive approaches to finding hy-
brid (sub)models inside an initial grouping of data according to an initial hybrid
model. Taken together these techniques provide a powerful set of general tech-
niques which are computationally efficient and statistically robust. In this chapter,
we give a brief introduction to some basic concepts associated with data model-
ing along with some motivating examples for modeling mixed data and a brief
account of a few related approaches.

1.1 Modeling Data with a Parametric Model

In scientific studies or engineering practice, one is frequently called upon to infer
(or learn) a quantitative model $M$ of a given set of sample points, denoted as
$X = \{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^D$. For instance, Figure 1.1 shows a simple example
in which one is given a set of four sample points on a two dimensional plane.
Obviously, these points can be fitted perfectly by a (one-dimensional) straight
line. The line can then be called a “model” for the given points. The reason for

![Figure 1.1. Four sample points on a plane are fitted by a straight line. However, they can also be fitted by many other smooth curves, for example the one indicated by the dashed curve.](image)

inferring such a model is because it serves many useful purposes: It can reveal
the information encoded in the data or the underlying mechanisms from which
the data were generated; Or it may simplify significantly the representation of the
given data set or it help to predict effectively future samples.

1.1.1 The Choice of a Model Class

However, inferring the “correct” model of a given data set is an elusive, if not
impossible, task. A fundamental difficulty is that, if we are not specific about
what we mean by a “correct” model, there could easily be many different models
that fit the given data set “equally well.” For instance, for the example shown in
Figure 1.1, any smooth curve that passes through the sample points would seem
to be an as valid model as the straight line. Furthermore, if there were noise in
the given sample points, then any curve, including the line, passing through the points exactly would unlikely be the “ground truth.”

The question now is: in what sense then can we say a model is correct for a given data set? Firstly, to make the model-inference a well-posed problem, we need to impose additional assumptions or restrictions on the class of models considered. This is to say we should not be looking for any model that can describe the data. Instead, we seek a model $M^*$ that is the best among a restricted class of models $\mathcal{M} = \{M\}$. In fact, the well-known No Free Lunch Theorem in statistical learning stipulates that in the absence of prior information or assumptions about the models, there is no reason to prefer one learning algorithm over another [Duda et al., 2000]. Secondly, we need to specify how restricted the class of models need to be. A common strategy is to try to get away with the simplest possible class of models that is just necessary to describe the data or solve the problem at hand – known as the principle of Occam’s Razor. More precisely, the model class should be rich enough that it contains at least one model that can fit the data to a desired accuracy and yet be simple enough so as to make the inference of the best model for the given data tractable.

In engineering practice, a popular strategy is to start from the simplest class of models, and only increase the complexity of the models when the simpler models become inadequate. For instance, to fit a set of sample points, one may try first the simplest class of models, namely linear models, then hybrid (piecewise) linear models, followed by (piecewise) quadratic models, and eventually by general topological manifolds. One of the goals of this book is to demonstrate that between them piecewise linear and quadratic models can already achieve an excellent balance between expressiveness and simplicity for many important practical problems.

1.1.2 Statistical Models versus Geometric Models

Roughly speaking, without any training samples whose group membership are known a priori, the problem of modeling mixed data falls into the category of unsupervised learning. In the literature, almost all unsupervised-learning methods fall into one of two categories. The first category of methods model the data as random samples from a probabilistic distribution and try to learn the distribution from the data. We call such models as statistical models. The second category of methods model the overall geometric shape of the data set as smooth manifolds or topological spaces. We call such models as geometric models.

---

4 Or equivalently, we may impose a non-uniform prior distribution over all models.

5 Roughly speaking, a smooth manifold is a special topological space that is locally smooth – Euclidean space-like, and has the same dimension everywhere. A general topological space may have singularities and consist of components of different dimensions.
Statistical Learning.

In the statistical paradigm, one assumes that the points \( x_i \) in the data set \( X \) are drawn independently from a common probability distribution \( p(x) \), then the task of learning a model from the data becomes one of inferring the most likely probability distribution within a family of distributions of interest (for example the Gaussian distributions). Normally the family of distributions is parameterized and denoted as \( \mathcal{M} = \{ p(x|\theta) \theta \in \Theta \} \). Consequently, the optimal model \( p(x|\theta^*) \) is given by the maximum likelihood (ML) estimate\(^6\)

\[
\theta^*_{ML} = \arg \max_{\theta \in \Theta} \prod_{i=1}^{N} p(x_i|\theta).
\] (1.1)

If a prior distribution (density) \( p(\theta) \) of the parameter \( \theta \) is also given, then, following the Bayesian rule, the optimal model is given by the maximum a posterior (MAP) estimate

\[
\theta^*_{MAP} = \arg \max_{\theta \in \Theta} \prod_{i=1}^{N} p(x_i|\theta)p(\theta).
\] (1.2)

Many effective methods and algorithms have been developed in the statistics and machine learning literature to infer the optimal distribution \( p(x|\theta^*) \) or a good approximation of it if the exact solution is computationally prohibitive. A brief review is given in Appendix C. The estimated probability distribution gives a generative description of the samples and can be used to generate new samples or predict the outcomes of new observations.

Geometric Modeling.

However, in many practical scenarios, it is rather difficult to know a priori the statistical origins of the data, since we frequently begin with only a few raw sample points, which are insufficient to determine a unique optimal distribution within a large (unrestricted) functional space. Very often, the data points are subject to certain hard geometric constraints, and can only be represented as a singular distribution.\(^7\) It is very ineffective to learn such a singular distribution via statistical means [Vapnik, 1995]. Thus, an alternative data-modeling paradigm is to directly learn the overall geometric shape of the given data set \( X \). Typical methods include fitting one or more geometric primitives such as points\(^8\), lines, subspaces, surfaces, and manifolds to the data set. For instance, the approach of classical principal component analysis (PCA) is to fit a lower-dimensional subspace, say \( S = \text{span}\{u_1, u_2, \ldots, u_d\} \), to a data set in a high-dimensional space,

---

\(^6\)If the true distribution from which the data are drawn is \( q(x) \), then the maximum likelihood estimate \( p(x|\theta^*) \) minimizes the Kullback-Leibler (KL) divergence: \( d(p, q) = \int p \log \frac{p}{q} \, dx \) among the given class of distributions.

\(^7\)Mathematically, singular distributions are represented as generalized functions in the Sobolev space. The delta function \( \delta(\cdot) \) is one such example.

\(^8\)As the means of clusters.
say $X = \{x_i\} \subset \mathbb{R}^D$. That is,

$$x_i = y_{i1}u_1 + y_{i2}u_2 + \cdots + y_{id}u_d + \epsilon_i, \quad \forall x_i \in X,$$

(1.3)

where $d < D$, $y_{ij} \in \mathbb{R}$, and $u_1, u_2, \ldots, u_d \in \mathbb{R}^D$ are unknown parameters and need to be determined—playing the role of the parameters $\theta$ in the foregoing statistical model. The line model in Figure 1.1 is an example of PCA for the four points on the plane. In the above equation, the terms $\epsilon_i \in \mathbb{R}^D$ denote possible errors between the samples and the model. PCA minimizes the error $\sum_i \|\epsilon_i\|^2$ for the optimal subspace (see Chapter 2 for details).\(^9\) In general, a geometric model gives an intuitive description of the samples, and it is often preferred to a statistical one as a “first-cut” description of the given data set. Its main purpose is to capture global geometric, topological, or algebraic characteristics of the data set, such as the number of clusters and their dimensions. A geometric model always gives a more compact representation of the original data set, which makes it useful for data compression and dimension reduction.

As two competing data-modeling paradigms, the statistical modeling techniques in general are more effective in the high-noise (or high-entropy) regime when the generating distribution is (piecewise) non-singular; and the geometric techniques are more effective in the low-noise (or low-entropy) regime when the underlying geometric space is (piecewise) smooth, at least locally. The two paradigms thus complement each other in many ways. Once the overall geometric shape, the clusters and their dimensions, of the data set are obtained from geometric modeling, one can choose the class of probabilistic distributions more properly for further statistical inference. Since samples normally have noise and sometimes contain outliers or incomplete data, in order to robustly estimate the optimal geometric model, one often resorts to statistical techniques. Therefore, while this book puts more emphasis on geometric and algebraic modeling techniques, we will also thoroughly investigate their connection to and combination with various statistical techniques (see Chapter 5).

### 1.2 Modeling Mixed Data with a Hybrid Model

However, in practice, many data sets $X$ cannot be modeled well by any single model $M$ in a pre-chosen or preferred model class $\mathcal{M}$. Nevertheless, it is often the case that if we group such a data set $X$ into multiple subsets:

$$X = X_1 \cup X_2 \cup \cdots \cup X_n, \quad \text{(with } X_l \cap X_m = \emptyset, \text{ for } l \neq m, \text{)}$$

(1.4)

then each subset $X_j$ can be modeled sufficiently well by a primitive model in the chosen model class:

$$M^*_j = \arg \min_{M \in \mathcal{M}} \left[ \text{Error}(X_j, M) \right], \quad j = 1, 2, \ldots, n.$$

(1.5)

---

\(^9\)When the data points $x_i$ are independent samples drawn from a Gaussian distribution, the geometric formulation of PCA coincides with the classical statistical formulation [Jolliffe, 1986].
Precisely in this sense, we call the data set $X$ *mixed* (with respect to the chosen model class $\mathcal{M}$) and call the collection of models $\{M^*_j\}_{j=1}^n$ together a *hybrid model for $X$*. For instance, suppose we are given a set of sample points shown in Figure 1.2. These points obviously cannot be fitted well by any single line, plane or smooth surface in $\mathbb{R}^3$; but once they are grouped into three subsets, each subset can be fitted well by a line or a plane (as a standard PCA problem). Note that in this example the topology of the data is indeed “hybrid” – two of the subsets are of dimension one and the other is of dimension two.

### 1.2.1 Examples of Mixed Data Modeling

In fact, the aforementioned example of mixed data is quite representative of many real data sets that one often encounters in practice. To motivate further the importance of modeling mixed data, we give below a few real-world problems that arise in computer vision and image processing. Most of these problems will be revisited later in this book with more detailed and principled solutions given.

**Example One: Vanishing Points in a Perspective Image**

The first example is the problem of *vanishing point detection* in computer vision. It is known in computer vision that the perspective images of a group of parallel lines in space all pass through a common point on the image plane which is the so-called vanishing point – a fact already well-known to and extensively exploited by Renaissance artists. Detecting vanishing points is very important for many practical applications such as estimating camera orientation and reconstructing scene structure, especially for man-made environments. A line on the image plane is described as the set of points $(x, y)$ described by an equation $ax + by + c = 0$. For each of the lines passing through the same vanishing point, its coefficient vector $x = [a, b, c]^T \in \mathbb{R}^3$ must lie on a 2-D subspace, whose normal vector is exactly the vanishing point $v = [v_x, v_y, 1]^T$, i.e., $v^T x = 0$. The vanishing point
is the point in the plane \((v_x, v_y) \in \mathbb{R}^2\). The extra entry 1 in the vector \(v \in \mathbb{R}^3\) may be thought of as representing membership in the plane.

For a scene that consists of multiple sets of parallel lines, as is usually the case for man-made objects and environments, the problem of detecting all the vanishing points from the set of all edge segments is then mathematically equivalent to clustering points into multiple 2-D subspaces in \(\mathbb{R}^3\). As we will see later that this is a special case of the subspace-segmentation problem addressed by GPCA (see Chapter 4). Figure 1.3 shows the application of the GPCA algorithm to one such example, in which edge segments are correctly grouped to three vanishing points.

**Example Two: Motion Segmentation from Two Images**

The second example is the so-called motion segmentation problem that arises also in the field of computer vision: given a sequence of images of multiple moving objects in a scene, how does one segment the images so that each segment corresponds to only one moving object? This is a very important problem in applications such as motion capture, vision-based navigation, target tracking, and surveillance. If we study the image sequence two images at a time, as it has been known in computer vision, feature points that belong to the same moving object are subject to either linear or quadratic constraints (see Chapter 8), depending on the type of motions and camera models. Therefore, mathematically, the problem of motion segmentation is equivalent to segmentation of points to different linear subspaces and quadratic surfaces. Figure 1.4 shows two images of a moving checker board and cube. The image on the left shows the starting positions of the board and the cube and their directions of motion; and the image on the right shows the final positions. The image on the right also shows the segmentation results obtained using the GPCA algorithm applied to their motion flow for points on the cube and the board. We will describe in detail the motion segmentation method used to achieve the above result in Chapter 7 and 8.

**Example Three: Image Representation and Segmentation**

A third example arises in the context of image representation and segmentation. It is commonplace that, in an image, pixels at different regions have significantly
different local color/texture profiles (normally an $N \times N$ window around a pixel). Conventional image representation/compression schemes (JPEG, JPEG2000) often ignore such differences and apply the same linear filters or bases (for example, the Fourier transform, discrete cosine transform, wavelets, or curvelets) to the entire set of local profiles. Nevertheless, modeling the set of local profiles as a mixed data set allows us to segment the image into different regions and represent them differently. Each region consists of only those pixels whose local profiles span the same low-dimensional linear subspace.\footnote{Unlike the previous two examples, there is no rigorous mathematical justification that local profiles from a region of similar texture must span a low-dimensional linear subspace. However, there is strong empirical evidence that a linear subspace is normally a very good approximation.} The basis of the subspace can be viewed as a bank of adaptive filters for the associated image region and the subspaces (and hence the segmentation of the image) can be effectively estimated via the GPCA algorithm. Figure 1.5 shows regions of an image segmented by such a hybrid representation. The so-obtained subspaces (and their bases) normally pro-

![Figure 1.4. Clustering feature points according to different 3-D motions.](image)

![Figure 1.5. Image segmentation based on fitting different linear subspaces (and bases) to regions of different textures. The three segments (or subspaces) correspond to the ground, the cloud, and the sky.](image)
1.2. Modeling Mixed Data with a Hybrid Model

Figure 1.6. Classifying a subset of the Yale Face Database B consisting of 64 frontal views under varying lighting conditions for subjects 2, 5 and 8. Left: Image data projected onto the first three principal components. Right: Classification of the images given by GPCA.

Example Four: Object Classification

The fourth example arises in the context of image-based object classification. Given a collection of unlabeled images $\{I_i\}_{i=1}^{n}$ of several different faces taken under varying illumination, we would like to classify the images corresponding to the face of the same person. For a Lambertian object, it has been shown that the set of all images taken under all lighting conditions forms a cone in the image space, which can be well approximated by a low-dimensional subspace [Ho et al., 2003]. Therefore, we can classify the collection of images by estimating a basis for each one of those subspaces, because the images of different faces will live in different subspaces. This is obviously another subspace-segmentation problem. In the example shown in Figure 1.6, we use a subset of the Yale Face Database B consisting of $n = 64 \times 3$ frontal views of three faces (subjects 5, 8 and 10) under 64 varying lighting conditions. For computational efficiency, we first down-sample each image to a size of 30 x 40 pixels. We then project the data onto their first three principal components using PCA, as shown in Figure 1.6 (left). We apply GPCA to the projected data in $\mathbb{R}^3$ and obtain three affine subspaces of dimension 2, 1, and 1, respectively. Despite the series of down-sampling and projection, the subspaces lead to a perfect classification of the images, as shown in Figure 1.6 (right).

Example Five: Video Segmentation and Event Detection

The last example arises in the context of detecting events from video sequences. A typical video sequence contains multiple activities or events separated in time. For instance, Figure 1.7 left shows a news sequence where the host is interviewing a guest and the camera is switching between the host, the guest and both of them. Let us assume that all the consecutive frames of the same scene can be modeled as

\footnote{The legitimacy of the projection process will be addressed in Chapter 3.}
the output from a linear dynamical system and that different scenes correspond to different dynamical systems. Since the image data live in a very high-dimensional space (~ $10^5$, the number of pixels), we first project the image data onto a low-dimensional subspace (~ 10) using principal component analysis (PCA) and then apply GPCA to the projected data to identify the different dynamical systems (see Chapter 9). Figure 1.7 shows the segmentation results for two video sequences. In both cases, a perfect segmentation is obtained.

![Figure 1.7] (Clustering frames of a news video sequence into groups of scenes by modeling each group with a linear dynamical system. Left: 30 frames of a video sequence clustered into 3 groups: host, guest, and both of them. Right: 60 frames of a news video sequence from Iraq clustered into 3 groups: rear of a car with a burning wheel, a burnt car with people, and a burning car.)

As we see from the above examples, in some cases, one can rigorously show that a given data set belongs to a collection of linear and quadratic surfaces of the same dimension (example one) or of possibly different dimensions (example two). In many other cases, one can use piecewise linear structures to approximate the data set and obtain a more compact and meaningful geometric representation of the data, including segments, dimensions, and bases (examples three, four, and five). Subspace (or surface) segmentation is a natural abstraction of all these problems and thus merits systematic investigation. From a practical standpoint, the study will lead to many general and powerful modeling tools that are applicable also to many types of data, such as feature points, images, videos, audios, dynamical data, genomic data, proteomic data, and other bio-informatic data sets.

### 1.2.2 Mathematical Representations of Hybrid Models

Obviously, whether the model associated with a given data set is hybrid or not depends on the class of primitive models considered. In this book, the primitives are normally chosen to be simple classes of smooth manifolds or non-singular distributions. For instance, one may choose the primitive models to be linear subspaces. Then one can use an arrangement of linear subspaces $\{S_i\}_{i=1}^n \subset \mathbb{R}^D$,

$$Z = S_1 \cup S_2 \cup \cdots \cup S_n,$$

(1.6)
also called a hybrid linear model, to approximate many nonlinear manifolds or piecewise smooth topological spaces. This is the typical model studied by generalized principal component analysis (GPCA). Or as its statistical counterpart, one can assume that the samples points are drawn independently from a mixture of Gaussian distributions \( \{ p_i(\mathbf{x}), \mathbf{x} \in \mathbb{R}^D \} \): 
\[
q(\mathbf{x}) = \pi_1 p_1(\mathbf{x}) + \pi_2 p_2(\mathbf{x}) + \cdots + \pi_n p_n(\mathbf{x}),
\]
with \( \pi_i > 0 \) and \( \pi_1 + \pi_2 + \cdots + \pi_n = 1 \). This is the typical model studied in mixtures of probabilistic principal component analysis (PPCA) [Tipping and Bishop, 1999a]. A classical way of estimating such a mixture model is the so-called expectation maximization (EM) method, which infers the membership of each sample as a hidden random variable (see Appendix C for a review). In this book, we will study and clarify the similarities and differences between these geometric models and statistical models (see Chapter 3 and 4).

**Difficulties with Conventional Data-Modeling Methods.**

One may have been wondering why not simply to enlarge the class of primitive models to include such hybrid models so that we can deal with them by the conventional single-model paradigms for learning distribution- or manifold-like models? If this were the case, then there would be no need of developing a theory for hybrid models and thus no need of this book! However, the most compelling reason that we do need hybrid models is that smooth manifolds and non-singular distributions are *not rich enough to describe the structure of many commonly observed data*, as we have seen in the examples in the previous section. On one hand, the underlying topological space of a mixed data set may contains multiple manifolds of different dimensions which may probably intersect with each other, as the case with a collection of multiple subspaces. Conventional estimation techniques for manifold-like models do not apply well to this class of spaces. On the other hand, if one represents a hybrid model as a probabilistic distribution, then the distribution will typically not be a (piecewise) smooth function but is singular. Conventional statistical-learning techniques become rather ineffective in inferring such singular distributions [Vapnik, 1995].

An alternative approach to model mixed data is first to segment the data set into coherent subsets and then to model each subset using the classical single-model methods. This is a popular approach adopted by many practitioners in industry. The fundamental difficulty with this approach is that, without knowing which subset of sample points belongs to which constituent model, there is seemingly a “chicken-and-egg” relationship between data segmentation and model estimation: If the segmentation of the data was known, one could fit a model to each subset of samples using classical model estimation techniques; and conversely, if the constituent models were known, one could easily find the subset of samples that best fit each model. This relationship has been the rationale that supports all the *iterative* modeling techniques for mixed data, such as the well-known expectation maximization (EM) algorithm and the K-means method (see Appendix C). These iterative methods share several drawbacks:
• The iteration needs to start with a good initial guess of the solution; otherwise the iteration is likely to converge to an irrelevant local minimum.

• Without knowing \textit{a priori} the number of models and the dimension of each model, the algorithm may diverge if it starts with a wrong guess on these key parameters.

• There are cases in which it is difficult to solve the grouping problem correctly, yet it is possible to obtain a good estimate of the models. In such cases a direct estimation of the models without grouping seems more appropriate than that based on incorrectly segmented data.

\textbf{Hybrid Models as Algebraic Sets.}

In this book, instead of manifolds or distributions, we will represent hybrid models mainly as algebraic sets.\footnote{Roughly speaking, an algebraic set is the common zero-level set of a family of algebraic equations, see Appendix A. For instance, most constraints among multiple images of the same scene are given in the form of algebraic equations [Ma et al., 2003].} To see the merit of such a representation, let us suppose that data that belong to the \textit{i}-th constituent model can be described as the zero-level set of some polynomials in a prime ideal \( p_i \), i.e., an (irreducible) algebraic variety:\footnote{A prime ideal is an ideal that cannot be decomposed further as the intersection of two other ideals. Geometrically, its zero-level set corresponds to an algebraic set that cannot be reduced to multiple algebraic sets. An irreducible algebraic set is called an algebraic variety. A subspace is one such example.}

\begin{equation}
Z_i \doteq \{ \mathbf{x} : p(\mathbf{x}) = 0, \ p \in p_i \} \subset \mathbb{R}^D, \quad i = 1, 2, \ldots, n. \tag{1.8}
\end{equation}

The (mixed) data from a union of \( n \) such models then belong to an algebraic set:\footnote{Notice that the “union” of algebraic varieties corresponds to the “multiplication” of the polynomials associated with the varieties.}

\begin{equation}
Z \doteq Z_1 \cup Z_2 \cup \cdots \cup Z_n = \{ \mathbf{x} : p_1(\mathbf{x})p_2(\mathbf{x}) \cdots p_n(\mathbf{x}) = 0, \ \forall p_i \in p_i, \ i = 1, 2, \ldots, n \}. \tag{1.9}
\end{equation}

From a number of (random) sample points on the algebraic set \( X = \{ \mathbf{x} \in Z \} \), one can determine the (radical) ideal of polynomials that vanish on the set \( Z \):\footnote{According to Hilbert’s Nullstellensatz, there is a one-to-one correspondence between algebraic sets and radical ideals [Eisenbud, 1996].}

\begin{equation}
X \rightarrow q(Z) \doteq \{ q(\mathbf{x}) = 0, \ \forall \mathbf{x} \in Z \}. \tag{1.10}
\end{equation}

Obviously, the ideal \( q \) is no longer a prime ideal. Thus, once the ideal \( q \) is obtained, the constituent models \( p_i \) (or \( Z_i \)) can be subsequently retrieved by \textit{decomposing} the ideal \( q \) into irreducible prime ideals:\footnote{For the special case in which the ideal is generated by a single polynomial, the decomposition is equivalent to factoring the polynomial into irreducible factors.}

\begin{equation}
q \rightarrow q = p_1 \cap p_2 \cap \cdots \cap p_n. \tag{1.11}
\end{equation}
Clearly, this representation establishes a natural correspondence between common terminologies used in algebraic geometry with the heuristic languages developed in (mixed) data modeling.

**Modeling Hybrid Topologies and Degenerate Distributions.**

Despite its pure algebraic nature, the above representation is closely related to, as well as complements, the aforementioned two paradigms of data modeling. From the geometric viewpoint, unlike a smooth manifold $M$ which sometimes can be implicitly represented as the level set of a single function, an algebraic set $Z$ is the zero-level set of a family of polynomials. Because of that, algebraic sets $Z$ may have components with different dimensions and singular topologies. From the statistical viewpoint, one can also view the irreducible components $\{Z_i\}$ of $Z$ as the “means” of a collection of probabilistic distributions $\{p_i(\cdot)\}$ and the overall set $Z$ as the “skeleton” of their mixture $q(\cdot)$. For instance, a piecewise linear structure can be viewed as the skeleton of a mixture of Gaussian distributions (see Figure 1.8). Therefore, hybrid models represented by algebraic sets can be interpreted as a special class of generative models such that the random variables have very low entropy outside the algebraic sets but high entropy inside.

![Figure 1.8](image.png)

Figure 1.8. Comparison of three representations of the same data set: 1. a (nonlinear) manifold, 2. a (mixed Gaussian) distribution, or 3. a (piecewise linear) algebraic set.

As we will show in this book, if the primitive varieties are simple models such as linear subspaces or quadratic surfaces, then in principle, the problem of segmenting mixed data and estimating a hybrid model can be solved *non-iteratively* (see Chapter 4). As it turns out, the correct number of models and their dimensions can also be correctly determined via purely algebraic means, at least in the noise-free case (see Chapter 4). The algebraic theory of GPCA will be thoroughly developed from Chapter 2 through Chapter 4. The algorithms developed in these chapters are algebraic in their methods. The most common concern that one hears is that algebraic methods are extremely sensitive to noise. But an amazing fact is that even the basic algebraic algorithm works extremely well with moderate noise in the data; and it is very efficient even for high-dimensional data as long as the data in fact is clustered on fairly low-dimensional structures.
1.2.3 Noise, Outliers, and Model Selection

However, in many real-world applications, the given data samples may be corrupted with noise or contaminated with outliers. Figure 1.9 shows one such example. Unlike the noiseless or low-noise scenario, the problem of finding the “correct” model becomes much more challenging in presence of a significant amount of noise or outliers. Proper statistical and robust statistical techniques therefore need to be developed for the estimation and segmentation of algebraic sets such as subspace arrangements. These issues will be carefully treated in Chapter 5.

Another important observation is that, in the presence of noise and outliers, a hybrid linear model is not necessarily the best if it has the highest fidelity to the data. This is especially the case when the number of subspaces and their dimensions are not known a priori. In fact, for every point in the data set, one can fit a separate line to it, which results in no modeling error at all. But such a model is not so appealing since it has exactly the same complexity as the original data.

In general, the higher is the model complexity, the smaller the error is.\(^\text{17}\) A good model should strike a balance between the complexity of a model \(M\) and its fidelity to the data \(X\).\(^\text{18}\) Many general model selection criteria have been proposed in the statistics or machine learning literature, including the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the minimum description length (MDL), and the minimum message length (MML). See Appendix C for a brief review. Despite some small differences, these criteria all tradeoff

\(^{17}\)For example, any function can be approximated arbitrarily close by a piecewise linear function with a sufficient number of pieces.

\(^{18}\)For instance, the complexity of a model can be measured as the minimum number of bits needed to fully describe the model and the data fidelity can be measured by the distance from the sample points to the model.
1.2. Modeling Mixed Data with a Hybrid Model

Modeling error for model complexity and minimize an objective of the following form:

$$\min_{M \in \mathcal{M}} J(M) = [\alpha \cdot \text{Complexity}(M) + \beta \cdot \text{Error}(X, M)].$$

In this book, we will introduce a model complexity measure that is specially designed for an arrangement of linear subspaces of arbitrary dimensions, namely the effective dimension (see Chapter 5).

There is yet another fundamental tradeoff that is often exploited for model selection. When the model complexity is too high, the model tends to over-fit the given data, including the noise in it. Such a model does not generalize well in the sense that it would not predict well the outcome of new samples; when the model complexity is too low, the model under-fits the data and, again, would result in a large error in the prediction. Therefore, a good model should minimize the prediction error. The relationships between modeling error and prediction error as a function of model complexity is plotted in Figure 1.10. Unfortunately, the “optimal” models obtained from trading off modeling error and prediction error can be different, as illustrated in the figure. In such a case, a choice between the two objectives has to be made. In the unsupervised learning setting, it is often difficult to obtain the prediction error curve;\footnote{Unless one does cross-validation within the given data set itself.} and for purposes such as data compression, the prediction error is of less concern than the modeling error. In these cases, we often choose the tradeoff between the modeling error and the model complexity (see Chapter 5).

In the end, all these statistical variations of the algebraic GPCA algorithms can be easily combined with the EM and K-means algorithms (for instance for initial-
izing them) so as to improve the generality, efficiency, robustness, and optimality of the existing methods for modeling mixed data.