LNCS 3940

Craig Saunders
Marko Grobelnik
Steve Gunn
John Shawe-Taylor (Eds.)

Subspace, Latent Structure and Feature Selection

Statistical and Optimization Perspectives Workshop, SLSFS 2005 Bohinj, Slovenia, February 2005 Revised Selected Papers

Springer

Craig Saunders Marko Grobelnik
Steve Gunn John Shawe-Taylor (Eds.)

Subspace, Latent Structure and Feature Selection

Statistical and Optimization
Perspectives Workshop, SLSFS 2005
Bohinj, Slovenia, February 23-25, 2005
Revised Selected Papers

Springer

Volume Editors

Craig Saunders
Steve Gunn
John Shawe-Taylor
University of Southampton
School of Electronics and Computer Science
ISIS Group
Southampton, SO17 1BJ, UK
E-mail: {cjs,srg,jst}@ecs.soton.ac.uk

Marko Grobelnik
J. Stefan Institute
Department of Intelligent Systems
Jamova 39, 1000 Ljubljana, Slovenia
E-mail: Marko.Grobelnik@ijs.si

Library of Congress Control Number: 2006925463

CR Subject Classification (1998): F.1, F.2, I.2.6-8, I.2.3, I.4-6, F.4.1, H.3

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN 0302-9743
ISBN-10 3-540-34137-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-34137-6 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11752790 06/3142 5 4 3 2 1 0

Preface

The inspiration for this volume was a workshop held under the auspices of the PASCAL Network of Excellence. Details of the event and more information about the Network can be found under the heading 'Workshop Organization.' The aim of this preface is to provide an overview of the contributions to this volume, placing this research in its wider context.

The aim of the workshop was to bring together researchers working on subspace and latent variable techniques in different research communities in order to create bridges and enable cross-fertilization of ideas. For this reason we deliberately sought invited refereed contributions that would survey a broader field of research giving a common notation and entry point into the individual papers.

The five invited contributions are (in alphabetical order of first author) Avrim Blum on *Random Projection, Margins, Kernels and Feature Selection*, Wray Buntine and Aleks Jakulin on *Discrete Principal Components Analysis*, Dunja Mladenić on *Dimensionality Reduction by Feature Selection in Machine Learning*, Roman Rosipal and Nicole Krämer on *Overview and Recent Advances in Partial Least Squares*, and Mike Titterington on *Some Aspects of Latent Structure Analysis.*

Blum considers subspace selection by random projection. The theoretical analysis of this approach provides an important bound on the generalization of large margin algorithms, but it can also be implemented in kernel defined feature spaces through a two-stage process. The paper provides a survey of a number of clean and important theoretical results. Buntine and Jakulin consider method of determining latent structure based on probabilistic generative models of the data. Their paper gives an introduction to these advanced and effective methods presented from within the machine learning community. Titterington's contribution is a closely related paper but comes from the statistics tradition providing a general framework within which discrete and continuous combinations of latent and observed variables can be placed. Mladenić considers the restricted class of axis parallel subspaces that correspond to feature selection. There is a long tradition of this approach within machine learning and the paper provides an overview of a range of techniques for selecting features, discussing their weaknesses and carefully evaluating their performance. Rosipal and Krämer give a detailed introduction to partial least squares, an important method of subspace selection developed within the chemometrics research community. It can be thought of as an adaptation of principal components analysis where the projection directions have been chosen to be well-suited for solving a set of regression tasks. The authors discuss the kernelization of the technique together with other more recent results.

The contributed papers cover a range of application areas and technical approaches. Agakov and Barber develop a probabilistic modelling technique with a novel twist of using encoding models rather than generative ones; Monay et al. again consider computer vision using a probabilistic modelling approach; Navot et al. analyze a simple two Gaussian example to show that feature selection can make significant differences in performance and that techniques such as support vector machines are not able to avoid the difficulties of non-informative features in this case; Bouveyron et al. consider a computer vision application using a probabilistic modelling approach; Gruber and Weiss continue the computer vision theme but introduce prior knowledge to enhance the ability to factorize image data to perform 3D reconstruction; Savu-Krohn and Auer use a clustering approach to reduce feature dimensions for image data; Rogers and Gunn consider random forests as an approach to feature selection; Maurer gives frequentist style generalization bounds on PCA-like subspace methods; and finally Reunanen discusses the biases of using cross-validation to do feature selection and outlines some techniques to prevent the introduction of such a bias.

We commend the volume to you as a broad introduction to many of the key approaches that have been developed for subspace identification and feature selection. At the same time the contributed talks give insightful examples of applications of the techniques and highlight recent developments in this rapidly expanding research area. We hope that the volume will help bridge the gaps between different disciplines and hence enable creative collaborations that will bring benefit to all involved.

February 2006

Marko Grobelnik
Steve Gunn
Craig Saunders
John Shawe-Taylor

Workshop Organization

Many of the papers in this proceedings volume were presented at the PASCAL Workshop entitled *Subspace, Latent Structure and Feature Selection Techniques: Statistical and Optimization Perspectives* which took place in Bohinj, Slovenia during February, 23–25 2005.

The workshop was part of a Thematic Programme Linking Learning and Statistics with Optimization that ran over the first half of 2005. The PASCAL Network is a European Network of Excellence funded by the European Union under the IST programme. It currently has around 300 researchers at 55 institutions. Its center of gravity is machine learning, but it aims to build links with both optimization and statistics as well as with a range of application areas. It sponsors and co-sponsors a wide range of workshops either organized independently or co-located with international conferences. More information can be found on the website http://www.pascal-network.org.

The Bohinj workshop was hosted by the Institute Josef Stefan, which provided all of the local organization. We are indebted to them for all of the hard work that they put into making the event such a success, although even they could not have planned the magical winter scene that awaited us on our arrival. Particular thanks are due to Tina Anžič, who handled the reservations and hotel bookings as well as many of the travel arrangements.

This work was supported by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

Organizing Committee

Marko Grobelnik	Jožef Stefan Institute, Ljubljana, Slovenia
Steve Gunn	ISIS Group, University of Southampton, UK
Craig Saunders	ISIS Group, University of Southampton, UK
John Shawe-Taylor	ISIS Group, University of Southampton, UK

Lecture Notes in Computer Science

For information about Vols. 1–3893

please contact your bookseller or Springer

Vol. 3994: V.N. Alexandrov, G.D. van Albada, P.M.A. Sloot, J. Dongarra (Eds.), Computational Science – ICCS 2006, Part IV. XXIX, 1094 pages. 2006.

Vol. 3993: V.N. Alexandrov, G.D. van Albada, P.M.A. Sloot, J. Dongarra (Eds.), Computational Science – ICCS 2006, Part III. XXX, 1138 pages. 2006.

Vol. 3992: V.N. Alexandrov, G.D. van Albada, P.M.A. Sloot, J. Dongarra (Eds.), Computational Science – ICCS 2006, Part II. XXIX, 1121 pages. 2006.

Vol. 3991: V.N. Alexandrov, G.D. van Albada, P.M.A. Sloot, J. Dongarra (Eds.), Computational Science – ICCS 2006, Part I. CCXX, 1090 pages. 2006.

Vol. 3987: M. Hazas, J. Krumm, T. Strang (Eds.), Location- and Context-Awareness. X, 289 pages. 2006.

Vol. 3986: K. Stølen, W.H. Winsborough, F. Martinelli, F. Massacci (Eds.), Trust Management. XIV, 474 pages. 2006.

Vol. 3984: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), Computational Science and Its Applications - ICCSA 2006, Part V. XXV, 1045 pages. 2006.

Vol. 3983: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), Computational Science and Its Applications - ICCSA 2006, Part IV. XXVI, 1191 pages. 2006.

Vol. 3982: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), Computational Science and Its Applications - ICCSA 2006, Part III. XXV, 1243 pages. 2006.

Vol. 3981: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), Computational Science and Its Applications - ICCSA 2006, Part II. XXVI, 1255 pages. 2006.

Vol. 3980: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), Computational Science and Its Applications - ICCSA 2006, Part I. LXXV, 1199 pages. 2006.

Vol. 3979: T.S. Huang, N. Sebe, M.S. Lew, V. Pavlović, T. Kölsch, A. Galata, B. Kisačanin (Eds.), Computer Vision in Human-Computer Interaction. XII, 121 pages. 2006.

Vol. 3978: B. Hnich, M. Carlsson, F. Fages, F. Rossi (Eds.), Recent Advances in Constraints. VIII, 179 pages. 2006. (Sublibrary LNAI).

Vol. 3976: F. Boavida, T. Plagemann, B. Stiller, C. Westphal, E. Monteiro (Eds.), Networking 2006. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems. XXVI, 1276 pages. 2006.

Vol. 3970: T. Braun, G. Carle, S. Fahmy, Y. Koucheryavy (Eds.), Wired/Wireless Internet Communications. XIV, 350 pages. 2006.

Vol. 3968: K.P. Fishkin, B. Schiele, P. Nixon, A. Quigley (Eds.), Pervasive Computing. XV, 402 pages. 2006.

Vol. 3967: D. Grigoriev, J. Harrison, E.A. Hirsch (Eds.), Computer Science – Theory and Applications. XVI, 684 pages. 2006.

Vol. 3966: Q. Wang, D. Pfahl, D.M. Raffo, P. Wernick (Eds.), Software Process Change. XIV, 356 pages. 2006.

Vol. 3964: M. Ü. Uyar, A.Y. Duale, M.A. Fecko (Eds.), Testing of Communicating Systems. XI, 373 pages. 2006.

Vol. 3962: W. IJsselsteijn, Y. de Kort, C. Midden, B. Eggen, E. van den Hoven (Eds.), Persuasive Technology. XII, 216 pages. 2006.

Vol. 3960: R. Vieira, P. Quaresma, M.d.G.V. Nunes, N.J. Mamede, C. Oliveira, M.C. Dias (Eds.), Computational Processing of the Portuguese Language. XII, 274 pages. 2006. (Sublibrary LNAI).

Vol. 3959: J.-Y. Cai, S. B. Cooper, A. Li (Eds.), Theory and Applications of Models of Computation. XV, 794 pages. 2006.

Vol. 3958: M. Yung, Y. Dodis, A. Kiayias, T. Malkin (Eds.), Public Key Cryptography - PKC 2006. XIV, 543 pages. 2006.

Vol. 3956: G. Barthe, B. Gregoire, M. Huisman, J.-L. Lanet (Eds.), Construction and Analysis of Safe, Secure, and Interoperable Smart Devices. IX, 175 pages. 2006.

Vol. 3955: G. Antoniou, G. Potamias, C. Spyropoulos, D. Plexousakis (Eds.), Advances in Artificial Intelligence. XVII, 611 pages. 2006. (Sublibrary LNAI).

Vol. 3954: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision – ECCV 2006, Part IV. XVII, 613 pages. 2006.

Vol. 3953: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision – ECCV 2006, Part III. XVII, 649 pages. 2006.

Vol. 3952: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision – ECCV 2006, Part II. XVII, 661 pages. 2006.

Vol. 3951: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision – ECCV 2006, Part I. XXXV, 639 pages. 2006.

Vol. 3950: J.P. Müller, F. Zambonelli (Eds.), Agent-Oriented Software Engineering VI. XVI, 249 pages. 2006.

Vol. 3947: Y.-C. Chung, J.E. Moreira (Eds.), Advances in Grid and Pervasive Computing. XXI, 667 pages. 2006.

Vol. 3946: T.R. Roth-Berghofer, S. Schulz, D.B. Leake (Eds.), Modeling and Retrieval of Context. XI, 149 pages. 2006. (Sublibrary LNAI).

Vol. 3945: M. Hagiya, P. Wadler (Eds.), Functional and Logic Programming. X, 295 pages. 2006.

Vol. 3944: J. Quiñonero-Candela, I. Dagan, B. Magnini, F. d'Alché-Buc (Eds.), Machine Learning Challenges. XIII, 462 pages. 2006. (Sublibrary LNAI).

Vol. 3943: N. Guelfi, A. Savidis (Eds.), Rapid Integration of Software Engineering Techniques. X, 289 pages. 2006.

Vol. 3942: Z. Pan, R. Aylett, H. Diener, X. Jin, S. Göbel, L. Li (Eds.), Technologies for E-Learning and Digital Entertainment. XXV, 1396 pages. 2006.

Vol. 3940: C. Saunders, M. Grobelnik, S. Gunn, J. Shawe-Taylor (Eds.), Subspace, Latent Structure and Feature Selection. X, 209 pages. 2006.

Vol. 3939: C. Priami, L. Cardelli, S. Emmott (Eds.), Transactions on Computational Systems Biology IV. VII, 141 pages. 2006. (Sublibrary LNBI).

Vol. 3936: M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, A. Yavlinsky (Eds.), Advances in Information Retrieval. XIX, 584 pages. 2006.

Vol. 3935: D. Won, S. Kim (Eds.), Information Security and Cryptology - ICISC 2005. XIV, 458 pages. 2006.

Vol. 3934: J.A. Clark, R.F. Paige, F.A. C. Polack, P.J. Brooke (Eds.), Security in Pervasive Computing. X, 243 pages. 2006.

Vol. 3933: F. Bonchi, J.-F. Boulicaut (Eds.), Knowledge Discovery in Inductive Databases. VIII, 251 pages. 2006.

Vol. 3931: B. Apolloni, M. Marinaro, G. Nicosia, R. Tagliaferri (Eds.), Neural Nets. XIII, 370 pages. 2006.

Vol. 3930: D.S. Yeung, Z.-Q. Liu, X.-Z. Wang, H. Yan (Eds.), Advances in Machine Learning and Cybernetics. XXI, 1110 pages. 2006. (Sublibrary LNAI).

Vol. 3929: W. MacCaull, M. Winter, I. Düntsch (Eds.), Relational Methods in Computer Science. VIII, 263 pages. 2006.

Vol. 3928: J. Domingo-Ferrer, J. Posegga, D. Schreckling (Eds.), Smart Card Research and Advanced Applications. XI, 359 pages. 2006.

Vol. 3927: J. Hespanha, A. Tiwari (Eds.), Hybrid Systems: Computation and Control. XII, 584 pages. 2006.

Vol. 3925: A. Valmari (Ed.), Model Checking Software. X, 307 pages. 2006.

Vol. 3924: P. Sestoft (Ed.), Programming Languages and Systems. XII, 343 pages. 2006.

Vol. 3923: A. Mycroft, A. Zeller (Eds.), Compiler Construction. XIII, 277 pages. 2006.

Vol. 3922: L. Baresi, R. Heckel (Eds.), Fundamental Approaches to Software Engineering. XIII, 427 pages. 2006.

Vol. 3921: L. Aceto, A. Ingólfsdóttir (Eds.), Foundations of Software Science and Computation Structures. XV, 447 pages. 2006.

Vol. 3920: H. Hermanns, J. Palsberg (Eds.), Tools and Algorithms for the Construction and Analysis of Systems. XIV, 506 pages. 2006.

Vol. 3918: W.K. Ng, M. Kitsuregawa, J. Li, K. Chang (Eds.), Advances in Knowledge Discovery and Data Mining. XXIV, 879 pages. 2006. (Sublibrary LNAI).

Vol. 3917: H. Chen, F.Y. Wang, C.C. Yang, D. Zeng, M. Chau, K. Chang (Eds.), Intelligence and Security Informatics. XII, 186 pages. 2006.

Vol. 3916: J. Li, Q. Yang, A.-H. Tan (Eds.), Data Mining for Biomedical Applications. VIII, 155 pages. 2006. (Sublibrary LNBI).

Vol. 3915: R. Nayak, M.J. Zaki (Eds.), Knowledge Discovery from XML Documents. VIII, 105 pages. 2006.

Vol. 3914: A. Garcia, R. Choren, C. Lucena, P. Giorgini, T. Holvoet, A. Romanovsky (Eds.), Software Engineering for Multi-Agent Systems IV. XIV, 255 pages. 2006.

Vol. 3910: S.A. Brueckner, G.D.M. Serugendo, D. Hales, F. Zambonelli (Eds.), Engineering Self-Organising Systems. XII, 245 pages. 2006. (Sublibrary LNAI).

Vol. 3909: A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 612 pages. 2006. (Sublibrary LNBI).

Vol. 3908: A. Bui, M. Bui, T. Böhme, H. Unger (Eds.), Innovative Internet Community Systems. VIII, 207 pages. 2006.

Vol. 3907: F. Rothlauf, J. Branke, S. Cagnoni, E. Costa, C. Cotta, R. Drechsler, E. Lutton, P. Machado, J.H. Moore, J. Romero, G.D. Smith, G. Squillero, H. Takagi (Eds.), Applications of Evolutionary Computing. XXIV, 813 pages. 2006.

Vol. 3906: J. Gottlieb, G.R. Raidl (Eds.), Evolutionary Computation in Combinatorial Optimization. XI, 293 pages. 2006.

Vol. 3905: P. Collet, M. Tomassini, M. Ebner, S. Gustafson, A. Ekárt (Eds.), Genetic Programming. XI, 361 pages. 2006.

Vol. 3904: M. Baldoni, U. Endriss, A. Omicini, P. Torroni (Eds.), Declarative Agent Languages and Technologies III. XII, 245 pages. 2006. (Sublibrary LNAI).

Vol. 3903: K. Chen, R. Deng, X. Lai, J. Zhou (Eds.), Information Security Practice and Experience. XIV, 392 pages. 2006.

Vol. 3902: R. Kronland-Martinet, T. Voinier, S. Ystad (Eds.), Computer Music Modeling and Retrieval. XI, 275 pages. 2006.

Vol. 3901: P.M. Hill (Ed.), Logic Based Program Synthesis and Transformation. X, 179 pages. 2006.

Vol. 3900: F. Toni, P. Torroni (Eds.), Computational Logic in Multi-Agent Systems. XVII, 427 pages. 2006. (Sublibrary LNAI).

Vol. 3899: S. Frintrop, VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. XIV, 216 pages. 2006. (Sublibrary LNAI).

Vol. 3898: K. Tuyls, P.J. 't Hoen, K. Verbeeck, S. Sen (Eds.), Learning and Adaption in Multi-Agent Systems. X, 217 pages. 2006. (Sublibrary LNAI).

Vol. 3897: B. Preneel, S. Tavares (Eds.), Selected Areas in Cryptography. XI, 371 pages. 2006.

Vol. 3896: Y. Ioannidis, M.H. Scholl, J.W. Schmidt, F. Matthes, M. Hatzopoulos, K. Boehm, A. Kemper, T. Grust, C. Boehm (Eds.), Advances in Database Technology - EDBT 2006. XIV, 1208 pages. 2006.

Vol. 3895: O. Goldreich, A.L. Rosenberg, A.L. Selman (Eds.), Theoretical Computer Science. XII, 399 pages. 2006.

Vol. 3894: W. Grass, B. Sick, K. Waldschmidt (Eds.), Architecture of Computing Systems - ARCS 2006. XII, 496 pages. 2006.

¥359.05元

Table of Contents

Invited Contributions

Contributed Papers

Discrete Component Analysis

Wray Buntine[1] and Aleks Jakulin[2]

[1] Helsinki Institute for Information Technology (HIIT),
Dept. of Computer Science, PL 68,
00014, University of Helsinki, Finland
Wray.Buntine@hiit.fi
[2] Department of Knowledge Technologies,
Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
jakulin@acm.org

Abstract. This article presents a unified theory for analysis of components in discrete data, and compares the methods with techniques such as independent component analysis, non-negative matrix factorisation and latent Dirichlet allocation. The main families of algorithms discussed are a variational approximation, Gibbs sampling, and Rao-Blackwellised Gibbs sampling. Applications are presented for voting records from the United States Senate for 2003, and for the Reuters-21578 newswire collection.

1 Introduction

Principal component analysis (PCA) [MKB79] is a key method in the statistical engineering toolbox. It is well over a century old, and is used in many different ways. PCA is also known as the Karhünen-Loève transform or Hotelling transform in image analysis, and a variation is latent semantic analysis (LSA) in text analysis [DDL$^+$90]. It is a kind of eigen-analysis since it manipulates the eigenspectrum of the data matrix. It is usually applied to measurements and real valued data, and used for feature extraction or data summarization. LSA might not perform the centering step (subtracting the mean from each data vector prior to eigen-analysis) on the word counts for a document to preserve matrix sparseness, or might convert the word counts to real-valued `tf*idf` [BYRN99]. The general approach here is *data reduction.*

Independent component analysis (ICA, see [HKO01]) is in some ways an extension of this general approach, however it also involves the estimation of so-called latent, unobservable variables. This kind of estimation follows the major statistical methodology that deals with general unsupervised methods such as clustering and factor analysis. The general approach is called *latent structure analysis* [Tit], which is more recent, perhaps half a century old. The data is modelled in a way that admits unobservable variables, that influence the observable variables. Statistical inference is used to "reconstruct" the unobservable variables from the data jointly with general characteristics of the unobservable variables themselves. This is a theory with particular assumptions (i.e., a "model"), so the method may arrive at poor results.

C. Saunders et al. (Eds.): SLSFS 2005, LNCS 3940, pp. 1–33, 2006.

Relatively recently the statistical computing and machine learning community has become aware of seemingly similar approaches for discrete observed data that appears under many names. The best known of these in this community are probabilistic latent semantic indexing (PLSI) [Hof99], non-negative matrix factorisation (NMF) [LS99] and latent Dirichlet allocation (LDA) [BNJ03]. Other variations are discussed later in Section 5. We refer to these methods jointly as *Discrete Component Analysis* (DCA), and this article provides a unifying model for them.

All the above approaches assume that the data is formed from individual observations (documents, individuals, images), where each observation is described through a number of variables (words, genes, pixels). All these approaches attempt to summarize or explain the similarities between observations and the correlations between variables by inferring latent variables for each observation, and associating latent variables with observed variables.

These methods are applied in the social sciences, demographics and medical informatics, genotype inference, text and image analysis, and information retrieval. By far the largest body of applied work in this area (using citation indexes) is in genotype inference due to the Structure program [PSD00]. A growing body of work is in text classification and topic modelling (see [GS04, BPT04]), and language modelling in information retrieval (see [AGvR03, BJ04, Can04]). As a guide, argued in the next section, the methods apply when PCA or ICA might be used, but the data is discrete.

Here we present in Section 3 a unified theory for analysis of components in discrete data, and compare the methods with related techniques in Section 5. The main families of algorithms discussed in Section 7 are a variational approximation, Gibbs sampling, and Rao-Blackwellised Gibbs sampling. Applications are presented in Section 8 for voting records from the United States Senate for 2003, and the use of components in subsequent classification.

2 Views of DCA

One interpretation of the DCA methods is that they are a way of approximating large sparse discrete matrices. Suppose we have a $500,000$ documents made up of $1,500,000$ different words. A document such as a page out of Dr. Seuss's *The Cat in The Hat*, is first given as a *sequence of words*.

> So, as fast as I could, I went after my net. And I said, "With my net I can bet them I bet, I bet, with my net, I can get those Things yet!"

It can be put in the *bag of words* representation, where word order is lost. This yields a list of words and their counts in brackets:

> after(1) and(1) as(2) bet(3) can(2) could(1) fast(1) get(1) I(7) my(3) net(3) said(1) so(1) them(1) things(1) those(1) went(1) with(2) yet(1) .

Although the word 'you' never appears in the original, we do not include 'you (0)' in the representation since zeros are suppressed. This sparse vector can be represented as a vector in full word space with $1,499,981$ zeroes and the counts above making the non-zero entries in the appropriate places. Given a matrix made up of rows of such vectors of non-negative integers dominated by zeros, it is called here a *large sparse discrete matrix.*

Bag of words is a basic representation in information retrieval [BYRN99]. The alternative is a sequence of words. In DCA, either representation can be used and the models act the same, up to any word order effects introduced by incremental algorithms. This detail is made precise in subsequent sections.

In this section, we argue from various perspectives that large sparse discrete data is not well suited to standard PCA or ICA methods.

2.1 Issues with PCA

PCA has been normally applied to numerical data, where individual instances are vectors of real numbers. However, many practical datasets are based on vectors of integers, non-negative counts or binary values. For example, a particular word cannot have a negative number of appearances in a document. The vote of a senator can only take three values: Yea, Nay or Not Voting. We can transform all these variables into real numbers using `tf*idf`, but this is a linear weighting that does not affect the shape of a distribution.

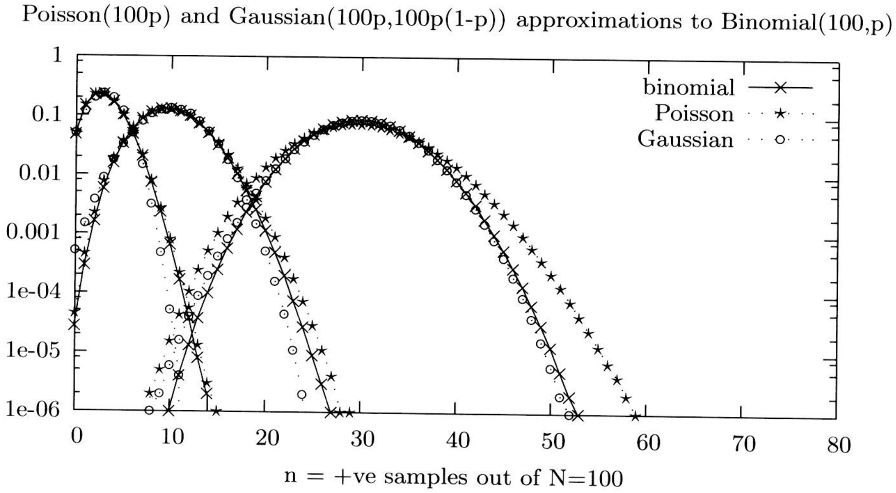
With respect to modelling count data in linguistic applications, Dunning makes the following warning [Dun94]:

> Statistics based on the assumption of normal distribution are invalid in most cases of statistical text analysis unless either enormous corpora are used, or the analysis is restricted to only the very most common words (that is, the ones least likely to be of interest). This fact is typically ignored in much of the work in this field. Using such invalid methods may seriously overestimate the significance of relatively rare events. Parametric statistical analysis based on the binomial or multinomial distribution extends the applicability of statistical methods to much smaller texts than models using normal distributions and shows good promise in early applications of the method.

While PCA is not always considered a method based on Gaussians, it can be justified using Gaussian distributions [Row98, TB99]. Moreover, PCA is justified using a least squares distance measure, and most of the properties of Gaussians follow from the distance measure alone. Rare events correspond to points far away under an L_2 norm.

Fundamentally, there are two different kinds of large sample approximating distributions that dominate discrete statistics: the Poisson and the Gaussian. For instance, a large sample binomial is approximated as a Poisson[1] when the probability is small and as a Gaussian otherwise [Ros89]. Figure 2.1 illustrates this by showing the Gaussian and Poisson approximations to a binomial with

[1] This is a distribution on integers where a rate is given for events to occur, and the distribution is over the total number of events counted.



sample size $N = 100$ for different proportions ($p = 0.03, 0.01, 0.03$). Plots are done with probability in log scale so the errors for low probability values are highlighted. One can clearly see the problem here: the Gaussian provides a reasonable approximate for medium values of the proportion p but for small values it severely underestimates low probabilities. When these low probability events occur, as they always will, the model becomes distorted.

Thus in image analysis based on analogue to digital converters, where data is counts, Gaussian errors can sometimes be assumed, but the Poisson should be used if counts are small. DCA then avoids Gaussian modelling of the data, using a Poisson or multinomial directly.

Another critique of the general style of PCA comes from the psychology literature, this time it is used as a justification for DCA [GS02]. Griffiths and Steyvers argue against the least squares distance of PCA:

> While the methods behind LSA were novel in scale and subject, the suggestion that similarity relates to distance in psychological space has a long history (Shepard, 1957). Critics have argued that human similarity judgments do not satisfy the properties of Euclidean distances, such as symmetry or the triangle inequality. Tversky and Hutchinson (1986) pointed out that Euclidean geometry places strong constraints on the number of points to which a particular point can be the nearest neighbor, and that many sets of stimuli violate these constraints.

They also considered power law arguments which PCA violates for associated words.

2.2 Component Analysis as Approximation

In the data reduction approach for PCA, one seeks to reduce each J-dimensional data vector to a smaller K-dimensional vector. This can be done by approximating the full data matrix as a product of smaller matrices, one representing the

$$\underset{\text{data matrix}}{\text{I documents}\left\{\overbrace{\begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,J} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ w_{I,1} & w_{I,2} & \cdots & w_{I,J} \end{pmatrix}}^{J \text{ words}}\right.} \cong \underset{\text{score matrix}}{\overbrace{\begin{pmatrix} l_{1,1} & \cdots & l_{1,K} \\ l_{2,1} & \cdots & l_{2,K} \\ \vdots & \ddots & \vdots \\ l_{I,1} & \cdots & l_{I,K} \end{pmatrix}}^{K \text{ components}}} * \underset{\text{loading matrix}^T}{\left.\overbrace{\begin{pmatrix} \theta_{1,1} & \theta_{2,1} & \cdots & \theta_{J,1} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1,K} & \theta_{2,K} & \cdots & \theta_{J,K} \end{pmatrix}}^{J \text{ words}}\right\}\text{K components}}$$

Fig. 1. The matrix approximation view

reduced vectors called the component/factor *score matrix*, and one representing a data independent part called the component/factor *loading matrix*, as shown in Figure 1. In PCA according to least squares theory, this approximation is made by eliminating the lower-order eigenvectors, the least contributing components [MKB79].

If there are I documents, J words and K components, then the matrix on the left has $I * J$ entries and the two matrices on the right have $(I + J) * K$ entries. This represents a simplification when $K \ll I, J$. We can view DCA methods as seeking the same goal in the case where the matrices are sparse and discrete.

When applying PCA to large sparse discrete matrices, or LSA using word count data interpretation of the components, if it is desired, becomes difficult (it was not a goal of the original method [DDL$^+$90]). Negative values appear in the component matrices, so they cannot be interpreted as "typical documents" in any usual sense. This applies to many other kinds of sparse discrete data: low intensity images (such as astronomical images) and verb-noun data used in language models introduced by [PTL93], for instance.

The cost function being minimized then plays an important role. DCA places constraints on the approximating score matrix and loading matrix in Figure 1 so that they are also non-negative. It also uses an entropy distance instead of a least squares distance.

2.3 Independent Components

Independent component analysis (ICA) was also developed as an alternative to PCA. Hyvänen and Oja [HO00] argue that PCA methods merely find uncorrelated components. ICA then was developed as a way of representing multivariate data with truly *independent* components. In theory, PCA approximates this also if the data is Gaussian [TB99], but in practice it rarely is.

The basic formulation is that a K-dimensional data vector $\boldsymbol{w}$ is a linear invertible function of K independent components represented as a K-dimensional latent vector $\boldsymbol{l}$, $\boldsymbol{w} = \boldsymbol{\Theta l}$ for a square invertible matrix $\boldsymbol{\Theta}$. Note the ICA assumes $J = K$ in our notation. $\boldsymbol{\Theta}$ plays the same role as the loading matrix above. For some univariate density model U, the independent components are distributed as

$p(\boldsymbol{l}\,|\,U) = \prod_k p(l_k\,|\,U)$, thus one can get a likelihood formula $p(\boldsymbol{w}\,|\,\boldsymbol{\Theta}, U)$ using the above equality[2].

The Fast ICA algorithm [HO00] can be interpreted as a maximum likelihood approach based on this model and likelihood formula. In the sparse discrete case, however, this formulation breaks down for the simple reason that $\boldsymbol{w}$ is mostly zeros: the equation can only hold if $\boldsymbol{l}$ and $\boldsymbol{\Theta}$ are discrete as well and thus the gradient-based algorithms for ICA cannot be justified. To get around this in practice, when applying ICA to documents [BKG03], word counts are sometimes first turned into `tf*idf` scores [BYRN99].

To arrive at a formulation more suited to discrete data, we can relax the equality in ICA (i.e., $\boldsymbol{w} = \boldsymbol{\Theta}\boldsymbol{l}$) to be an expectation:

$$\mathbb{E}_{\boldsymbol{w}\sim p(\boldsymbol{w}|\boldsymbol{l},U)}\left[\boldsymbol{w}\right] \;=\; \boldsymbol{\Theta}\boldsymbol{l}\;.$$

We still have independent components, but a more robust relationship between the data and the score vector. Correspondence between ICA and DCA has been noted in [BJ04, Can04]. With this expectation relationship, the dimension of $\boldsymbol{l}$ can now be less than the dimension of $\boldsymbol{w}$, $K < J$, and thus $\boldsymbol{\Theta}$ would be a rectangular matrix.

3 The Basic Model

A good introduction to these models from a number of viewpoints is by [BNJ03, Can04, BJ04]. Here we present a general model. The notation of words, bags and documents will be used throughout, even though other kinds of data representations also apply. In statistical terminology, a word is an observed variable, and a document is a data vector (a list of observed variables) representing an instance. In machine learning terminology, a word is a feature, a bag is a data vector, and a document is an instance. Notice that the bag collects the words in the document and loses their ordering. The bag is represented as a data vector $\boldsymbol{w}$. It is now J-dimensional. The latent, hidden or unobserved vector $\boldsymbol{l}$ called the component *scores* is K-dimensional. The term *component* is used here instead of topic, factor or cluster. The parameter matrix is the previously mentioned component loading matrix $\boldsymbol{\Theta}$, and is $J \times K$.

At this point, it is also convenient to introduce the symbology used throughout the paper. The symbols summarised in Table 1 will be introduced as we go.

3.1 Bags or Sequences of Words?

For a document $\boldsymbol{x}$ represented as a sequence of words, if $\boldsymbol{w} = \text{bag}(\boldsymbol{x})$ is its bagged form, the bag of words, represented as a vector of counts. In the simplest

[2] By a change of coordinates

$$p(\boldsymbol{w}\,|\,\boldsymbol{\Theta}, U) \;=\; \frac{1}{\det(\boldsymbol{\Theta})}\prod_k p\left(\left(\boldsymbol{\Theta}^{-1}\boldsymbol{w}\right)_k\,|\,U\right)$$