BIENNIAL REPORT

2010 – 2011

Adaptive Informatics Research Centre

Department of Information and Computer Science

Aalto University School of Science

P.O. Box 15400

FI–00076 Aalto, Finland

R Vigário, A. Juslin, L. Koivisto, and E. Oja, editors

Otaniemi, May 2012

Yliopistopaino
Helsinki 2012

Contents

Preface

The **Adaptive Informatics Research Centre (AIRC, adaptiivisen informatiikan tutkimusyksikkö)** was nominated as one of the national Centers of Excellence (CoE) by the Academy of Finland for the period 2006 - 2011. It was financed by the Academy, Tekes, HUT/Aalto University, and Nokia Co.

The present report covers the activities of AIRC during the final two years 2010 and 2011. It concentrates on the research projects, but also lists the degrees and awards given to the staff. The achievements and developments of the previous four years have been reported in the Biennial Reports 2006 - 2007 and 2008 - 2009. The web pages of AIRC, `http://www.cis.hut.fi/research` also contain up-to-date texts.

During 2010 - 2011, the AIRC was operating within the Department of Information and Computer Science (ICS), belonging to the new School of Science of Aalto University. Professor Erkki Oja was the director of AIRC, and Professor Samuel Kaski was the vice-director, with Professors Olli Simula and Juha Karhunen participating in its research projects. In addition, 16 post-doctoral researchers, ca. 30 full-time graduate students, and a number of undergraduate students were working in the AIRC projects.

To briefly list the main numerical outputs of AIRC during the period 2010 - 2011, the Centre produced 5 D.Sc. (Tech.) degrees and 47 M.Sc. (Tech.) degrees. The number of scientific publications appearing during the period was 235, of which 66 were journal papers. Thus the number of papers increased by 33% from the previous two-year period. It can be also seen that the impact of our research is clearly increasing, measured by the citation numbers to our previously published papers and books, as well as the number of users of our public domain software packages.

A large number of talks, some of them plenary and invited, were given by our staff in the major conferences in our research field. We had several foreign visitors participating in our research, and our own researchers made visits to universities and research institutes abroad. In addition to the finances provided by the Academy of Finland, Tekes, and Aalto University, AIRC researchers managed to obtain a large number of external projects (e.g. EU Emime, EU MultilingualWeb, EU PinView, EU ERASysBio, EU NoE T4ME, EU NoE Pascal2, UI-ART, NOVAC, MultiBio, VirtualCoach). Many of these are going on still in 2012. The research staff were active in international organizations, editorial boards of journals, and conference committees, including the conferences MLSP 2010, ICANN 2011, WSOM 2011, ALT 2011, and DS 2011, which all were organized by the AIRC staff in Finland and chaired by AIRC senior faculty. Also, some prices and honours, both national and international, were granted to members of our staff. All these are detailed in this report.

The third and fourth meetings of the Scientific Advisory Board of AIRC were held on Feb. 2 - 3, 2010, and October 11 - 12, 2011, respectively. The final evaluation report written by the members of the Board, Professors Risto Miikkulainen and José C. Príncipe, was quite positive. It begins by stating that *"The Center of Excellence on Adaptive Informatics Research directed by Professor Erkki Oja is on par with the best centers in*

the world in the adaptive informatics and machine learning areas, primarily because of the principled formulation of the research questions, the thematic coherence of its research, quality and quantity of its scientific production and in the novelty of its contributions".

A highlight in summer 2011 was the decision of the Academy of Finland to finance a new Center of Excellence in Computational Inference (COIN) in the 2012 - 2017 CoE programme. The competition was even harder than before, with only 15 CoE's accepted for financing. COIN is partly building on the research agenda of AIRC. E. Oja is still the Director of COIN and S. Kaski the Vice-Director for the first three-year period, after which S. Kaski will move to the position of Director. COIN consists of some of the same groups that comprise AIRC, but with considerable additions from the ICS Department as well as from the University of Helsinki.

Erkki Oja

Professor
Director,
Adaptive Informatics
Research Centre

Samuel Kaski

Professor
Vice-Director,
Adaptive Informatics
Research Centre

Personnel

Professors

Oja, Erkki; D.Sc. (Tech.), Director

Karhunen, Juha; D.Sc. (Tech.), part-time

Kaski, Samuel; D.Sc. (Tech.), Director of HIIT

Kohonen, Teuvo; D.Sc. (Tech.), Emeritus Professor, Academician

Simula, Olli; D.Sc. (Tech.), Dean, Faculty of Information and Natural Sciences, until December 2010, Head of the Department, 2011

Post-doc researchers

Alumäe, Tanel; Ph.D., from April 2009 until March 2011

Creutz, Mathias; D.Sc. (Tech.), part-time, from October to December 2011

Corona, Francesco; Ph.D.

Doddipatla, Rama; Ph.D., from September 2010 until August 2011

Georgii, Elisabeth; Ph.D., until December 2010

Gönen, Mehmet; Ph.D., from September 2010

Hirsimäki, Teemu; D.Sc. (Tech.), until March 2010

Honkela, Antti; D.Sc. (Tech.)

Honkela, Timo; Ph.D., Chief research scientist

Ilin, Alexander; D.Sc. (Tech.)

Klami, Arto; D.Sc. (Tech.)

Koskela, Markus; D.Sc. (Tech.)

Kujala, Jussi; D.Sc. (Tech.), until August 2010

Kurimo, Mikko; D.Sc. (Tech.), Teaching research scientist, Chief research scientist

Laaksonen, Jorma; D.Sc. (Tech.), Teaching research scientist, Chief research scientist from June 2010

Lagus, Krista; D.Sc. (Tech.)

Lahti, Leo; D.Sc. (Tech.), until February 2011

Leen, Gayle; Ph.D., until January 2011

Lendasse, Amaury; Ph.D.

Miche, Yoan; D.Sc. (Tech.)

Montesino-Pouzols, Federico; Ph.D., until May 2011

Palomäki, Kalle; D.Sc. (Tech.)

Peltonen, Jaakko; D.Sc. (Tech.)

Raiko, Tapani; D.Sc. (Tech.)
Raivio, Kimmo; D.Sc. (Tech.)
Valpola,Harri; D.Sc. (Tech.), part-time, from December 2010
van Leemput, Koen; Ph.D., part-time, from January 2011
Vigário, Ricardo; D.Sc. (Tech.)
Yang, Zhirong; D.Sc. (Tech.)
Yuan, Zhijian; D.Sc. (Tech.), until May 2010

Post-graduate researchers

Ajanki, Antti; M.Sc. (Tech.)
Caldas, Jose; M.Sc., Grant researcher
Chen, Xi; M.Sc. (Tech.)
Cho, Kyunghyun; M.Sc. (Tech.), from May 2011
Dobrinkat, Marcus; M.Sc., until December 2010
Eirola, Otto; M.Sc. (Tech.)
Enarvi, Seppo; M.Sc. (Tech.), from January 2011
Faisal, Ali; M.Sc.
Gonçalves, Nicolau; M.Sc., Grant researcher
van Heeswijk, Mark; M.Sc.
Huopaniemi, Ilkka; M.Sc. (Tech.), until September 2011
Ishikawa, Satoru; M.Sc., grant researcher from August 2010, researcher from August 2011
Ju, Yong Chul; M.Sc., from June 2010 until May 2011
Kallasjoki, Heikki; M.Sc. (Tech.)
Kandemir, Melih; M.Sc.
Karhila, Reima; M.Sc. (Tech.)
Keronen, Sami; M.Sc. (Tech.),
Khan, Suleiman; M.Sc., part-time, from June 2010
Kohonen, Oskar; M.Sc. (Tech.)
Kumlander, Mikaela; M.Sc. (Tech.), part-time, until June 2011
Liitiäinen, Elia; M.Sc. (Tech.), until October 2010
Lindh-Knuutila, Tiina; M.Sc. (Tech.)
Luttinen, Jaakko; M.Sc. (Tech.)
Mansikkaniemi, Andre; M.Sc. (Tech.)
Nevala, Maija; M.Sc. (Tech.); until February 2011
Nieminen, Ilari; M.Sc. (Tech.)
Nybo, Kristian; M.Sc. (Tech.)
Pajarinen, Joni; M.Sc. (Tech.)
Parkkinen, Juuso; M.Sc. (Tech.)
Parviainen, Jukka; M.Sc. (Tech.), University teacher
Paukkeri, Mari-Sanna; M.Sc. (Tech.)
Pylkkönen, Janne; M.Sc. (Tech.)

Pöllä, Matti; M.Sc. (Tech.)
Raitio, Juha; M.Sc. (Tech.), part-time
Remes, Ulpu; M.Sc. (Tech.)
Reyhani, Nima; M.Sc. (Tech.)
Ruokolainen, Teemu; M.Sc. (Tech.)
Sjöberg, Mats; M.Sc. (Tech.)
Sorjamaa, Antti; M.Sc. (Tech.), until December 2010
Smit, Pieter; M.Sc. (Tech.)
Sovilj, Dušan; M.Sc. (Tech.)
Suvitaival, Tommi; M.Sc. (Tech.),
Talonen, Jaakko; M.Sc. (Tech.)
Topa, Hande; M.Sc., from November 2011
Tornio, Matti; M.Sc. (Tech.); from September 2010 until August 2011
Turunen, Ville; M.Sc. (Tech.)
Viinikanoja, Jaakko; M.Sc. (Tech.)
Viitaniemi, Ville; M.Sc. (Tech.), Assistant until December 2010, Researcher from January 2011
Virpioja, Sami; M.Sc. (Tech.)
Virtanen, Seppo; M.Sc. (Tech.)
Väyrynen, Jaakko; M.Sc. (Tech.)
Wagner, Paul; M.Sc. (Tech.)
Yari Saeed Khanloo, Bahman; M.Sc., from September 2010 until August 2011
Ylipaavalniemi, Jarkko; M.Sc. (Tech.)
Yu, Qi; M.Sc. (Tech.)
Zhang, He; M.Sc. (Tech.)

Under-graduate researchers

Alene, Henok; from June 2010 until May 2011
Asikainen, Jukka; from June until August 2011
Augilius, Eimontas; until May 2011
Bahrami Rad, Ali; from October 2011
Calandra, Roberto; from February until July 2011
Cao, Yang; from October 2010
Chao, Wang; from September until November 2010
Gillberg, Leo
Hao, Tele; from January 2011
Hyyrynen, Lasse; from May until December 2010
Izzatdust, Zaur; from June 2011
Kainulainen, Laura; until April 2011
Karp, Elina; until November 2010
Karppa, Juho; from June 2010

Kivimäki, Ilkka; until December 2010
Klapuri, Antti; from June 2010
Kuusela, Mikael
Leino, Katri; from June until August 2011
Leppäaho, Eemeli; from June 2010
Leppänen, Laura; from June until August 2010
Lindell, Rony; from June until August 2010
Lopes Vidal, Alejandro; from June until July 2010
Malmi, Eric
Mannila, Anna; from June until August 2010
Matikainen, Miika; from June until August 2010
Mehmood, Yasir; from June 2010 until June 2011
Mohammadi, Pejman; until May 2010
Noeva, Polina; from June 2011
Osmala, Maria
Pandey, Sanjeev; from October 2011
Peltola, Veli; until August 2011
Pessala, Sini; until February 2010
Pesu, Tommi; from March until November 2011
Puonti, Oula; from May 2010
Ramachandra, Rao; from October 2010 until May 2011
Ramaseshan, Ajay; from September 2010
Ramon Letosa, Jorge; from June until July 2010
Remes, Sami; from June 2011
Saarimäki, Jarno; from February until July 2010
Soppela, Jyri; until July 2010
Vatanen, Tommi
Virtanen, Seppo
Wang, Chiwei; from June 2010
Wu, Jing; until June 2010
Yao, Li; until September 2010
Zhang, Jiefu; from June until August 2010
Zhu, Zhanxing

Support staff

Koivisto, Leila; Secretary until September 2010
Pihamaa, Tarja; Secretary
Ranta, Markku; B.Eng., Works engineer
Sirola, Miki; D.Sc. (Tech.), Laboratory engineer

Awards and activities

Prizes and academic awards received by personnel of the unit

Name: Professor Erkki Oja

- INNS Hebb Award, 2010
- Honorary Doctor, University of Eastern Finland

Name: Professor Olli Simula

- 2011 EspooAmbassador prize

Gayle Lee, Jaakko Peltonen, and Sami Kaski

- Best paper award, "Focused Multi-task Learning Using Gaussian Processes",in ECML PKDD 2011.

Hannu Pulakka, Ulpu Remes, Santeri Yrttiaho, Kalle Palomäki, Mikko Kurimo, and Paavo Alku

- Best paper award, "Low-Frequency Bandwidth Extension of Telephone Speech Using Sinusoidal Synthesis and Gaussian Mixture Model", ISCA Award for the best student paper of Interspeech 2011.

Jaakko Viinikanoja

- Master's Thesis Award of Pattern Recognition Society of Finland

Important international positions of trust held by personnel of the unit

Professor Juha Karhunen:

- Program Committee Member:
  The 5th Int. Workshop on Hybrid Artificial Intelligence Systems (HAIS2010), San Sebastian, Spain, June 2010
  Analysis and Signal Separation (LVA/ICA 2010), Saint-Malo, France, September 2010
  The European Symposium on Artificial Neural Networks (ESANN 2010), Bruges, Belgium, April 2010
  The 11th Int. Conf. on Intelligent Data Engineering and Automated Learning (IDEAL 2010), Paisley, Scotland, UK, September 2010
  Int. Conf. on Neural Computation (ICNC 2010), Valencia, Spain, October 2010
  The 4th Int. Conf. on Computational Intelligence in Security for Information Systems (CISIS'11)
  The European Symposium on Artificial Neural Networks (ESANN 2011)
  The 6th Int. Conf. on Hybrid Artificial Intelligence Systems (HAIS11)
  Int. Conf. on Neural Computation Theory and Applications (NCTA 2011)
  The 2011 IEEE Machine Learning for Signal Processing Workshop
  Int. Conf. on Intelligent Data Engineering and Automated Learning (IDEAL 2011)

- Editorial Board Member, Neural Processing Letters, The Netherlands.

Professor Samuel Kaski:

- Program Committee Member:
  ICML2010 Workshop on Reinforcement learning and search in very large spaces, Haifa, Israel, 21.-24.6.2010.
  2010 IEEE/WIC/ACM International Conference on Web Intelligence (WI'10), Toronto, Canada, Aug. 31 - Sept 3, 2010.
  ECML PKDD 2010, Barcelona, Spain, Sept. 20-24, 2010.
  European Symposium on Artificial European Symposium on Artificial Neural Networks ESANN'2010, April 2010, Bruges, Belgium.
  International Conference on Machine Learning, ICML 2010, June 21 - June 25, 2010, Haifa, Israel.
  Asia-Pacific Bioinformatics Conference, APBC2010, Jan 18-21, 2010, Bangalore, India.
  ICPR 2010, the 20th International Conference on Pattern Recognition, 23-26 August, 2010, Istanbul, Turkey.
  Mining and Learning with Graphs, MLG 2010, Washington, D.C., USA, July 24-25.
  KDD-2010, the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, July 25-28, Washington, D.C., USA.

- Session Chairman:
  ICML2010 Workshop on Reinforcement learning and search in very large spaces, Haifa, Israel, 2010

- Member of Steering Committee, EU NoE PASCAL2, UK.

- Associate Editor, International Journal of Knowledge Discovery in Bioinformatics, USA, 2010.
- Editorial Board Member:
  Intelligent Data Analysis, The Netherlands.
  International Journal of Neural Systems, Singapore.
  Cognitive Neurodynamics, Germany.
- Opponent at the doctoral dissertation of Jüri Reimand, University of Tartu, Estonia, 2010.

Professor Erkki Oja:

- Conference Chairman:
  IEEE International Workshop on Machine Learning for Signal Processing (MLSP) 2010.
  International Conference on Artificial Neural Networks (ICANN) 2011.
- Member of the committee, College of Fellows, International Neural Network Society, INNS, USA, 2011.
- Member of the Award Committee IEEE Computational Intelligence Society, USA, 2011.
- Member of the Fellowship Committee, IEEE Computational Intelligence Society, USA, 2011.
- Member of the EU FET Flagship -program External Advisory Group 2011.
- Editorial Board Member:
  Natural Computing - An International Journal, The Netherlands.
  Neural Computation, USA.
  International Journal of Pattern Recognition and Artificial Intelligence, Singapore.
- Invated talk in Sino-foreign-interchange Workshop on Intelligence Science and Intelligent Data Engineering (IScIDE), Harbin, China, May 31, 2010.
- International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain 2010.

Professor Olli Simula:

- Scientific Council Member, Institute Eurecom, France, 2011.

Dr. Francesco Corona:

- Pre-examiner of doctoral thesis:
  Gines Rubio, Universidad de Granada, Spain.

Dr. Antti Honkela:

- Program Committee Member, 5th IAPR International Conference on Pattern Recognition in Bioinformatics, PRIB 2010, Nijmegen, The Netherlands, 22.-24.9.2010.
- Session Chairman:
  EPSRC Symposium Workshop on Learning and inference in computational systems biology (LICSB), University of Warwick, U.K., 30.-31.3.2010.

Dr. Timo Honkela:

- Program Committee Member:
  ICANN 2010 Intl Conference on Artificial Neural Networks, Thessaloniki, Greece, 15.-18.9.2010.
- Chair, International Federation on Information Processing (IFIP), WG12.1 (Knowledge Representation and Reasoning), Austria.
- Representative of Finland, International Federation on Information Processing (IFIP), TC12 (Artificial Intelligence), Austria.
- Member of Executive Board, ENNS, European Neural Networks Society, The Netherlands.
- Editorial Board Member, Constructivist Foundations, Austria.
- Pre-examiner of a doctoral theses, Alicia Perez Ramarez, The University of the Basque Country, Spain 2010.
- Opponent at the doctoral dissertation:
  Michael Kai Petersen, Technical University of Denmark, Denmark, 2010.
  Bartlomiej Wilkowski, Technical University of Denmark, Denmark, 2011.

Dr. Arto Klami:

- Program Committee Member:
  PASCAL2/ICANN Challenge on MEG Mind Reading, 2011.

Dr. Markus Koskela:

- Program Committee Member:
  The 2010 IEEE Pacific-Rim Conference on Multimedia (IEEE-PCM2010), Shanghai, China, 21.-24.9.2010
- Pre-examiner of a doctoral theses, Pablo Toharia Rabasco, Universidad Politecnica de Madrid, Spain, 2010.

Dr. Mikko Kurimo:

- Program Committee Member:
  The Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11.-16.7.2010.
  Interspeech 2010, Makuhari, Japan, 26.-30.9.2010.
  Interspeech 2011 Firenze, Italy.
  Eusipco 2011 Barcelona, Spain.
  ACL-HLT Portland kesäkuu 2011.

- Session Chairman:
  Interspeech 2011 Firenze, Italy.

- Editorial board member, ACM Transactions on Speech and Language Processing.

- Editor, Multilingual Information Access Evaluation I - Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Part I. Lecture Notes in Computer Science.

- Evaluator of funding applications, Swiss National Science Foundation 2011.

Dr. Jorma Laaksonen:

- Program Committee Member:
  The International Conference on Image Analysis and Recognition (ICIAR 2010), Povoa de Varzim, Portugal, June 21-23, 2010.

- Associate editor and member of editorial board, Pattern Recognition Letters, The Netherlands, 2011.

Dr. Yoan, Miche:

- Programme committee member:
  European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 28-30th April 2010

Dr. Jaakko Peltonen:

- Program Committee Member:
  Seventh International Symposium on Neural Networks (ISNN 2010), Shanghai, China, June 6-9, 2010 ISNN 2011, 8th International Symposium on Neural Networks, Guilin, Kiina, 29.5.2011-1.6.2011

- Chairman of the session:
  ECML PKDD 2011, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Ateena, Kreikka, 5.9.2011-9.9.2011.

- Editor of scientific journals:
  Neurocomputing (special issue on Machine Learning for Signal Processing; managing guest editor).

- Editorial Board Member, Neural Processing Letters, The Netherlands.

Dr. Tapani Raiko:

- Program Committee Member:
  The 27th International Conference on Machine Learning (ICML 2010), Haifa, Israel, 21.-24.6. 2010

- Editor of scientific journals:
  Neurocomputing (special issue on Machine Learning for Signal Processing

Dr. Kimmo Raivio:

- ICT Domain expert, COST (European Cooperation in Science and Technology) 2010, Belgium.

Dr. Miki Sirola:

- Program Committee Member:
  International Conference on Networked Digital Technologies (NDT 2010), Prague, Czech Republic, 7 - 9 July 2010 International Conference on Networked Digital Technologies (NDT 2011). Macau, China, 18 - 21 July 2011
  IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2011)
  International Conference on Computational Intelligence and Bioinformatics (CIB 2011).

- Session Chairman:
  IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2011).

Dr. Ricardo Vigário:

- Program Committee Member:

  International Steering Committee of the 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'2010), St. Malo, France, September 27-30, 2010.
  Int. Symposium on Intelligent Data Analysis (IDA'2011), Porto, Portugal.

Dr. Zhirong Yang:

- IEEE / WIC / ACM International Conferences on Web Intelligence, Toronto, Canada, 31.8.- 3.9.2010.

- The 28th International Conference on Machine Learning, Bellevue, Washington, 28/6/2011-2/7/2011.

- IEEE/WIC/ACM International Conference on Web intelligence, Lyon, France, 22-27 August, 2011.

M.Sc. Leo Lahti:

- Member of the COST/EUGESMA working group, U.K., 2009.

Important domestic positions of trust held by personnel of the unit

Professor Juha Karhunen:

- Program Committee Member:
  The 20th IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010), Kittilä, Finland, August-September 2010
  Int. Conf. on Artificial Neural Networks (ICANN 2011)

Professor Samuel Kaski:

- Program Committee Member:
  Machine Learning for Signal Processing, MLSP 2010, Kittilä, Finland, 29.8.-1.89.2010.
  International Conference on Artificial Neural Networks (ICANN 2011), Espoo, Finland.

- Session Chairman:
  Machine Learning for Signal Processing, MLSP 2010, Kittilä, Finland, 2010.

- Invited talk: Methods for analyzing multiple metabolomics data sources, Metabomeeting 2011, Finland.

Professor Erkki Oja:

- Programme committee member:
  International Conference of Artificial Neural Networks (ICANN) 2011.

- Chairman, Research council for natural sciences and engineering, Academy of Finland
- Member, Governing Board of the Academy of Finland
- Opponent at the doctoral dissertation of Ville Heikkinen, University of Eastern Finland, 2011.

Professor Olli Simula:

- Chairman, IEEE Computer Chapter, Finland.
- Program Committee Member: 8th International Conference on Self-Organizing Maps (WSOM 2011).
  Twenty-second International Conference on Algorithmic Learning Theory (ALT 11) and Fourteenth International Conference on Discovery Science (DS 2011), jointly ALT/DS 2011.

Dr. Francesco Corona:

- Programme and organising committee member:
  Machine Learning for Signal Processing, MLSP 2010, Kittilä, Finland, 29.8.-1.89.2010.
  ICANN 2011: International Conference on Artificial Neural Networks, Espoo (Finland), June 14-17th, 2011.
  WSOM 2011: 8th International Workshop on Self-Organizing Maps, Espoo (Finland), June 13rd-15th, 2011.

- Pre-examiner of a doctoral thesis:
  Harri Niska, University of Eastern Finland, 2011.

Dr. Antti Honkela:

- Program Committee Member:
  Machine Learning for Signal Processing, MLSP 2010, Kittilä, Finland, 29.8.-1.89.2010.

Dr. Timo Honkela:

- Program Committee Member:
  Contexts of Language: How to analyze context?, Helsinki 10.-11.12.2010.
  International Conference of Artificial Neural Networks (ICANN) 2011.

- Editorial Board Member, Puhe ja kieli, Finland.
- Editorial Board Member, Tieteessä tapahtuu, Finland.
- Member of Steering Committee, Langnet - Finnish Graduate School in Language Studies, Finland, 2009.
- Board Member, Hecse - Helsinki Graduate School in Computer Science and Engineering, Finland, 2009.
- Opponent at the doctoral dissertation:
  Antti Airola, University of Turku.

Dr. Markus Koskela:

- Vice President, Suomen hahmontunnistuksen seura ry, Pattern Recognition Society of Finland 2010-2011.
- Pre-examiner of the doctoral thesis:
  Ville Heikkinen, University of Eastern Finland, 2011.
  Sami Huttunen, University of Oulu, 2011.

- Opponent at the doctoral dissertation of Sami Huttunen, University of Oulu, 2011.

Dr. Mikko Kurimo:

- Program Committee Member: International Conference of Artificial Neural Networks (ICANN) 2011.
  Morpho Challenge Workshop 2010, Espoo, Finland, 2.-3-9.2010
  IEEE Workshop on Machine Learning for Signal Processing, Kittilä, Finland, 29.8. -1.9.2010.

- Session chairman, IEEE Workshop on Machine Learning for Signal Processing, Kittilä, Finland, 29.8. -1.9.2010.

- Editor, Proceedings of the Morpho Challenge 2010 workshop.

- Opponent at the doctoral dissertation of Rahim Saeidi, University of Eastern Finland, 2011.

Dr. Jorma Laaksonen:

- Programme committee: 8th WORKSHOP ON SELF-ORGANIZING MAPS, WSOM 2011 Otaniemi, 13-15.6.2011.
  2010 IEEE International Workshop on Machine Learning for Signal Processing, August 29 - September 1, 2010 Kittilä, Finland.

- Pre-examiner of a doctoral theses: Alexey Andriyashin, Universtiy of Eastern Finland, 2011. Teemu Kinnunen, Lappeenranta University of Technology, 2011.

- Opponent at the doctoral dissertations:
  Vili Kellokumpu, University of Oulu, 2011.
  Teemu Kinnunen, Lappeenranta University of Technology, 2011.

- Programme committee member:
  ICANN 2011: International Conference on Artificial Neural Networks, June 14-17th, 2011, Espoo, Finland.

Dr. Jaakko Peltonen:

- Programme committee: The Twentieth IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010), Kittilä, Finland
  ICANN 2011, 8th Intl Workshop on Self-Organizing Maps, 14.-17.6.2011
  WSOM 2011, 8th Intl Workshop on Self-Organizing Maps, 13.-15.6.2011

- Advisor and board member,Helsinki Graduate School on Computational Science and Engineering.

Dr. Tapani Raiko:

- Programme and organising committee, 2010 IEEE International Workshop on Machine Learning for Signal Processing, Kittilä, Finland, 29.8.-1.9.2010.

- Vice Chairman, Finnish Artificial Intelligence Society, Suomen Tekoälyseura, Finland, 2010.

Dr. Kimmo Raivio:

- Pre-examiner of a doctoral thesis:
  Gabor, Fekete, University of Jyväskylä, 2010.

Dr. Ricardo Vigário:

- Membership of the organising committee, Int. Conf. on Artificial Neural Networks (ICANN'2011), Espoo.

- Pre-examiner of a doctoral thesis:
  Igor Kalyakin, University of Jyväskylä.

M.Sc. Leo Lahti:

- Vice Chairman, Finnish Society for Bioinformatics, 2010.

M.Sc. Mari-Sanna Paukkeri:

- Program Committee Member, International Conference of Artificial Neural Networks (ICANN).

Research visits abroad by personnel of the unit

- Lasse Hyyrynen, University of Tallin, Estonia, 4 months, 2010
- Tommi Vatanen, CERN, Geneve, Switzerland, 1 month, 2010.
- Dr.(Tech.) Kalle Palomäki, University of Sheffield, UK, 4 months, 2010.
- Dr.(Tech.) Antti Honkela, University of Manchester, UK, 7 months, 2010.
- Mikael Kuusela, CERN, Geneve, Switzerland, 3 months, 2010.
- Dr.(Tech.) Tapani Raiko, New York University, 1 month, USA, 2010.
- Dr.(Tech) Leo Lahti, European Bioinformatics Institute EBI, UK, 1 month, 2010.
- Dr.(Tech.) Mikko Kurimo, Nagoya Institute of Technology, Japan, 2 months, 2010.
- PhD Francesco Corona, University of Cagliari, Italy, 2 months, 2010.
- M.Sc. Ulpu Remes, Nagoya Institute of Technology, Japan, 4 months, 2010.
- M.Sc. Dusan Sovilj, Leibniz Institute for Baltic Sea Research, Warnemünde, Germanu, 2 months, 2010.
- Dr.(Tech.) Jaakko Peltonen, University of Sheffield, Sheffield Institute for Translational Neuroscience, UK,2 months, 2010.
- PhD Francesco Corona, University of Cagliari, Italy, 3 months, 2011.
- Dr.(Tech.) Jaakko Peltonen, University of Sheffield, Sheffield Institute for Translational Neuroscience, UK,8 months, 2011.
- Dr.(Tech.) Arto Klami, University of California, Berkeley, USA, 2 weeks, 2011.
- B.Sc. Mikael Kuusela, Fermilab Country, USA, 2 months, 2011.
- Dr.(Tech.) Kalle Palomäki, University of Sheffield, UK, 3 weeks, 2011.
- M.Sc. Ulpu Remes, Nagoya Institute of Technology, Japan, 4 months, 2011.

Research visits by foreign researchers to the unit

- PhD Tanel Alumäe, Tallinn University of Technology, Estonia, 3 months, 2010
- PhD Rama Sanand Doddipadla, IIT Kanpur, India, 4 months, 2010.
- M.Sc. Kairit Sirts, Tallinn University of Technology, Estonia, 1 month, 2010.
- Prof. Morgan Nelson, ICSI Berkeley, USA, 1 day, 2010.
- PhD Federico Montesino Pouzols, University of Sevilla, Spain, 12 months, 2010.
- PhD Alberto Guillen, University of Granada, Spain, 1 month, 2010.
- M.Sc. Benoit Frenay, M.Sc., Universite Catholique de Louvain, France, 2 months, 2010.
- Prof. Antoni Neme, National Autonomous University of Mexico, Meksiko, 4 months, 2010.
- PhD Koen van Leemput, Massachusetts General Hospital; Harvard Medical School, USA, 12 months, 2010.
- Assist. Prof. Milen Chechev, Sofia University, Bulgaria, 4 months, 2010.
- PhD Roman Naumov, Institute of Higher Nervous Activity and Neurophysiology, Moscow, Russia Federation, 1 month, 2010.
- M.Sc. Satoru Ishikawa, University of Aizu, Japan, 5 months, 2010.
- PhD Bjoern Menze, MIT, USA, 1 month, 2010.
- PhD Stan Smits, EIT ICT Labs, The Netherlands, Hollanti, 2 days, 2010.
- PhD Marie Sjölinder, SICS, Sweden, 3 days, 2010.
- Jorge Ramon Letosa, University of Zaragoza, Spain, 3 months, 2010.
- Alejandro Lopez Vidal, University Carlos III de Madrid, Spain, 3 months, 2010.
- PhD Mauricio Kugler, Nagoya Institute of Technology, Japan, 2 days, 2011.
- PhD Mathew Migimai Doss, IDIAP, Switzerland, 2 days, 2011.
- Prof. Eric Severin, University of Lille, France, 2 weeks, 2011.
- Prof. Asoke Nandi, University of Liverpool, UK, 1 day, 2011.
- Dean Jiang Xiuwen, Dahlian University, China, 1 day, 2011.
- Prof. Jose Principe, University of Florida, 2 days, 2011.
- Prof. Risto Miikkulainen, Univesrtiy of Texas at Austin, USA, 2 days, 2011.
- Prof. Thomas Griffiths, University of California, Berkeley, USA, 4 days, 2011.
- Prof. Geoffrey Hinton, University of Toronto, Canada, 4 days, 2011.
- Prof. John Shawe-Taylor, Univesrity College London, UK, 4 days, 2011.

- Prof. Joshua Tenenbaum, Massachusetts Institute of Technology, USA, 4 days, 2011.
- Prof. Barbara Hammer, Bielefeld University, Germany, 3 days, 2011.
- Hiroshi Mamitsuka, Tokyo University, Japan, 1 week, 2011.

Doctoral dissertations

Mutual Dependency-Based Modeling of Relevance in Co-Occurrence Data

Eerika Savia

Dissertation for the degree of Doctor of Science in Technology on 18 June 2010.

External examiners:
Tapio Salakoski (University of Turku, Finland)
David R. Hardoon (University College London, UK)
Opponent:
Michal Rosen-Zvi (IBM Reasearch Lab, Haifa, Israel)

Abstract:
In the analysis of large data sets it is increasingly important to distinguish the relevant information from the irrelevant. This thesis outlines how to find what is relevant in so-called co-occurrence data, where there are two or more representations for each data sample.

The modeling task sets the limits to what we are interested in, and in its part defines the relevance. In this work, the problem of finding what is relevant in data is formalized via dependence, that is, the variation that is found in both (or all) co-occurring data sets was deemed to be more relevant than variation that is present in only one (or some) of the data sets. In other words, relevance is defined through dependencies between the data sets.

The method development contributions of this thesis are related to latent topic models and methods of dependency exploration. The dependency-seeking models were extended to nonparametric models, and computational algorithms were developed for the models. The methods are applicable to mutual dependency modeling and co-occurrence data in general, without restriction to the applications presented in the publications of this work. The application areas of the publications included modeling of user interest, relevance prediction of text based on eye movements, analysis of brain imaging with fMRI and modeling of gene regulation in bioinformatics. Additionally, frameworks for different application areas were suggested.

Until recently it has been a prevalent convention to assume the data to be normally distributed when modeling dependencies between different data sets. Here, a distribution-free nonparametric extension of Canonical Correlation Analysis (CCA) was suggested, together with a computationally more efficient semi-parametric variant. Furthermore, an alternative view to CCA was derived which allows a new kind of interpretation of the results and using CCA in feature selection that regards dependency as the criterion of relevance.

Traditionally, latent topic models are one-way clustering models, that is, one of the variables is clustered by the latent variable. We proposed a latent topic model that generalizes in two ways and showed that when only a small amount of data has been gathered, two-way generalization becomes necessary.

In the field of brain imaging, natural stimuli in fMRI studies imitate real-life situations and challenge the analysis methods used. A novel two-step framework was proposed for analyzing brain imaging measurements from fMRI. This framework seems promising for the analysis of brain signal data measured under natural stimulation, once such measurements are more widely available.

Advances in the Theory of Nearest Neighbor Distributions

Elia Liitiäinen

Dissertation for the degree of Doctor of Science in Technology on 22 October 2010.

External examiners:
Mathew D. Penrose (University of Bath, UK)
Dafydd Evans (Cardiff University, UK)
Opponents:
Luc Devroye (McGill University, Montreal, Canada)

Abstract:
A large part of non-parametric statistical techniques are in one way or another related to the geometric properties of random point sets. This connection is present both in the design of estimators and theoretical convergence studies. One such relation between geometry and probability occurs in the application of non-parametric techniques for computing information theoretic entropies: it has been shown that the moments of the nearest neighbor distance distributions for a set of independent identically distributed random variables are asymptotically characterized by the R´enyi entropies of the underlying probability density. As entropy estimation is a problem of major importance, this connection motivates an extensive study of nearest neighbor distances and distributions.

In this thesis, new results in the theory of nearest neighbor distributions are derived using both geometric and probabilistic proof techniques. The emphasis is on results that are useful for finite samples and not only in the asymptotic limit of an infinite sample.

Previously, in the literature it has been shown that after imposing sufficient regularity assumptions, the moments of the nearest neighbor distances can be approximated by invoking a Taylor series argument providing the connection to the R´enyi entropies. However, the theoretical results provide limited understanding to the nature of the error in the approximation. As a central result of the thesis, it is shown that if the random points take values in a compact set (e.g. according to the uniform distribution), then under sufficient regularity, a higher order moment expansion is possible. Asymptotically, the result completely characterizes the error for the original low order approximation.

Instead of striving for exact computation of the moments through a Taylor series expansion, in some cases inequalities are more useful. In the thesis, it is shown that concrete upper and lower bounds can be established under general assumptions. In fact, the upper bounds rely only on a geometric analysis.

The thesis also contains applications to two problems in nonparametric statistics, residual variance and R´enyi entropy estimation. A well-established nearest neighbor entropy estimator is analyzed and it is shown that by taking the boundary effect into account, estimation bias can be significantly reduced. Secondly, the convergence properties of a recent residual variance estimator are analyzed.

Probabilistic Analysis of the Human Transcriptome with Side Information

Leo Lahti

Dissertation for the degree of Doctor of Science in Technology on 17 December 2010.

External examiners:
Juho Rousu (University of Helsinki, Finland)
Simon Rogers (University of Glasgow, UK)
Opponent:
Volker Roth (Universität Basel, Switzerland)

Abstract:
Recent advances in high-throughput measurement technologies and efficient sharing of biomedical data through community databases have made it possible to investigate the complete collection of genetic material, the genome, which encodes the heritable genetic program of an organism. This has opened up new views to the study of living organisms with a profound impact on biological research.

Functional genomics is a subdiscipline of molecular biology that investigates the functional organization of genetic information. This thesis develops computational strategies to investigate a key functional layer of the genome, the transcriptome. The time- and context-specific transcriptional activity of the genes regulates the function of living cells through protein synthesis. Efficient computational techniques are needed in order to extract useful information from high-dimensional genomic observations that are associated with high levels of complex variation. Statistical learning and probabilistic models provide the theoretical framework for combining statistical evidence across multiple observations and the wealth of background information in genomic data repositories.

This thesis addresses three key challenges in transcriptome analysis. First, new preprocessing techniques that utilize side information in genomic sequence databases and microarray collections are developed to improve the accuracy of high-throughput microarray measurements. Second, a novel exploratory approach is proposed in order to construct a global view of cell-biological network activation patterns and functional relatedness between tissues across normal human body. Information in genomic interaction databases is used to derive constraints that help to focus the modeling in those parts of the data that are supported by known or potential interactions between the genes, and to scale up the analysis. The third contribution is to develop novel approaches to model dependency between co-occurring measurement sources. The methods are used to study cancer mechanisms and transcriptome evolution; integrative analysis of the human transcriptome and other layers of genomic information allows the identification of functional mechanisms and interactions that could not be detected based on the individual measurement sources. Open source implementations of the key methodological contributions have been released to facilitate their further adoption by the research community.

Algorithms for approximate Bayesian inference with applications to astronomical data analysis

Yoan Miche

Dissertation for the degree of Doctor of Science in Technology on 2 November 2010.

External examiners:
Thomas Villmann (Hochschule Mittweida, University of Applied Sciences, Germany)
Andrew Ker (University of Oxford, UK)
Opponent:
Tapio Seppänen (University of Oulu, Finland)

Abstract:
In the history of human communication, the concept and need for secrecy between the parties has always been present. One way of achieving it is to modify the message so that it is readable only by the receiver, as in cryptography for example. Hiding the message in an innocuous medium is another, called steganography. And the counterpart to steganography, that is, discovering whether a message is hidden in a specific medium, is called steganalysis. Other concerns also fall within the broad scope of the term steganalysis, such as estimating the message length for example (which is quantitative steganalysis).

In this dissertation, the emphasis is put on classical steganalysis of images first – the mere detection of a modified image –, for which a practical benchmark is proposed: the evaluation of a sufficient amount of samples to perform the steganalysis in a statistically significant manner, followed by feature selection for dimensionality reduction and interpretability. The fact that most of the features used in the classical steganalysis task have a physical meaning, regarding the image, lends itself to an introspection and analysis of the selected features for understanding the functioning and weaknesses of steganographic schemes.

This approach is computationally demanding, both because of the feature selection and the size of the data in steganalysis problems. To address this issue, a fast and efficient machine learning model is proposed, the Optimally-Pruned Extreme Learning Machine (OP-ELM). It uses random projections in the framework of an Artificial Neural Network (precisely, a Single Layer Feedforward Network) along with a neuron selection strategy, to obtain robustness regarding irrelevant features, and achieves state of the art performances.

The OP-ELM is also used in a novel approach at quantitative steganalysis (message length estimation). The re-embedding concept is proposed, which embeds a new known message in a suspicious image. By repeating this operation multiple times for varying sizes of the newly embedded message, it is possible to estimate the original message size used by the sender, along with a confidence interval on this value. An intrinsic property of the image, the inner difficulty, is also revealed thanks to the confidence interval width; this gives an important information about the reliability of the estimation on the original message size.

Methodologies for Time Series Prediction and Missing Value Imputation

Antti Sorjamaa

Dissertation for the degree of Doctor of Science in Technology on 19 November 2010.

External examiners:
Madalina Olteanu (Universtiy of Paris 1, France)
Vincent Wertz(Catholic University of Louvain, Belgium)
Opponent:
Guilherme Barreto (Federal University of Ceará, Brazil)

Abstract:
The amount of collected data is increasing all the time in the world. More sophisticated measuring instruments and increase in the computer processing power produce more and more data, which requires more capacity from the collection, transmission and storage. Even though computers are faster, large databases need also good and accurate methodologies for them to be useful in practice. Some techniques are not feasible to be applied to very large databases or are not able to provide the necessary accuracy.

As the title proclaims, this thesis focuses on two aspects encountered with databases, time series prediction and missing value imputation. The first one is a function approximation and regression problem, but can, in some cases, be formulated also as a classification task. Accurate prediction of future values is heavily dependent not only on a good model, which is well trained and validated, but also preprocessing, input variable selection or projection and output approximation strategy selection. The importance of all these choices made in the approximation process increases when the prediction horizon is extended further into the future.

The second focus area deals with missing values in a database. The missing values can be a nuisance, but can be also be a prohibiting factor in the use of certain methodologies and degrade the performance of others. Hence, missing value imputation is a very necessary part of the preprocessing of a database. This imputation has to be done carefully in order to retain the integrity of the database and not to insert any unwanted artifacts to aggravate the job of the final data analysis methodology. Furthermore, even though the accuracy is always the main requisite for a good methodology, computational time has to be considered alongside the precision.

In this thesis, a large variety of different strategies for output approximation and variable processing for time series prediction are presented. There is also a detailed presentation of new methodologies and tools for solving the problem of missing values. The strategies and methodologies are compared against the state-of-the-art ones and shown to be accurate and useful in practice.

Theses

Master of Science in Technology

2010

Dubrovin, Tero
Tilastollinen tekstianalyysi kaupunkisuunnittelun apuvälineen osallistavassa paikkatietojärjestelmässä

Maaranen, Jyrki
Taloudellisten aikasarjojen ennustaminen ja analysointi ohjatut asiantuntijat -neuroverkolla

Mohammadi, Pejman
Bayesian Integrative Modelling of Metabolic and Transcriptional Data Using Pathway Information

Nevala, Maija
Discovering Functional Gene-MicroRNA Modules with Probabilistic Methods

Nieminen, Ilari
Tag Recommendation in Folksonomies

Oksman, Pekko
Classification of Precipitation Patterns in Weather Radar Images

Roivainen, Tuomo
Classification of taxi trips with self-organizing maps (Taksimatkojen luokittelu neuroverkko menetelmällä)

Räty, Jani
Number Plate Recognition

Saarimäki, Jarno
Performance Metrics for the Atmospheric Model ECHAM5

Tapiovaara, Tero
Normalized Compression Distance in Automatic Evaluation of Machine Translations

Tornio, Matti
Natural Gradient for Variational Bayesian Learning

Wagner, Paul
On the stability of reinforcement learning under partial observability and generalizing representations

Viinikanoja, Jaakko
Locally linear robust Bayesian dependency modeling of co-occurence data

Virtanen, Seppo
Bayesian Exponential Family Projections

Wu, Jing
Online Face Recognition with Application to Proactive Augmented Reality

Yao, Li
Anomaly Detection and Location with an Application to an Energy Management System

2011

Alene, Henok
Graph Based Clustering for Anomaly Detection in IP Networks

Calandra, Roberto
An Exploration of Deep Belief Networks toward Adaptive Learning

Cho, KyungHyun
Improved Learning Algorithms for Restricted Bolzmann Machines

Gillberg, Jussi
Targeted Learning by Imposing Asymmetric Sparsity

Lopez Vidal, Alejandro
Traffic flow simulation and optimization using evolutionary strategies

Mehmood, Yasir
Comparing Talks, Realities and Concerns on Climate Change: An analysis of Textual, Numerical and Categorical Data

Osmala, Maria
Regularized Modelling of Dependencies between Gene Expression and Metabolomics Data in Studying Metabolic Regulation

Smit, Pieter
Stacked transformations for foreign accented speech recognition

Trovo, Francesco
Regret Estimation for Multi Slot Incentive Compatible Multi Armed Bandit

Wang, Chiwei
Latent variable models for a probabilistic timeline browser

Wang, Chao
Software Development of Quadratic Nonnegative Matrix Factorization

Zhu, Zhanxing
Supervised Distance Preserving Projections for Dimensionality Reduction

Research Projects

Chapter 1

Introduction

The Centre of Excellence called the Adaptive Informatics Research Centre (AIRC) started in January 2006 in the Laboratory of Computer and Information Science at Helsinki University of Technology. It followed the tradition of the Neural Networks Research Centre (NNRC), operative for two six-year periods from 1994 to 2005, also under the national Centre of Excellence status. AIRC finished in December 2011 at the Department of Information and Computer Science, School of Science, Aalto University, to be directly followed by a new Centre of Excellence in Computational Inference (2012 - 2017).

The core function and strength of our Centres of Excellence is the ability to analyze and process extensive data sets coming from a number of various application fields using our own innovative and generic methods. Our research has concentrated on neurocomputing and statistical machine learning algorithms, with a number of applications. In the algorithmic research, we have attained a world class status over the years, especially in such unsupervised machine learning methods as the Self-Organizing Map and Independent Component Analysis.

Building on this solid methodological foundation, we apply the knowledge, expertise and tools to advance knowledge in other domains and disciplines. In the AIRC, we took a goal-oriented and interdisciplinary approach in targeting at the adaptive informatics problem. By adaptive informatics we mean a field of research where automated learning algorithms are used to discover the relevant informative concepts, components, and their mutual relations from large amounts of data. Access to the ever-increasing amounts of available data and its transformation to forms intelligible for the human user is one of the grand challenges in the near future.

The AIRC Centre of Excellence focussed on several adaptive informatics problems. One is the efficient retrieval and processing techniques for text, digital audio and video, and numerical data such as biological and medical measurements, which create valuable information sources. Another problem area are advanced multimodal natural interfaces. We are building systems that process multimodal contextual information including spoken and written language, images, videos, and explicit and implicit user feedback. Automated semantic processing of such information facilitates cost-effective knowledge acquisition and knowledge translation without the need to build the descriptions manually. Yet another problem, which we approach together with experts in brain science and molecular biology, is to develop and apply our algorithmic methods to problems in neuroinformatics and bioinformatics.

The Adaptive Informatics methodology that we focus on is to build empirical models of the data by using automated machine learning techniques, in order to make the information usable. The deep expertise on the algorithmic methods, gained over the years, is used to build realistic solutions, starting from the problem requirements. The application domains have been chosen because of our acquired knowledge in some of their core problems, because of their strategic importance in the near future, and because of their mutual interrelations. The algorithms are based on our own core expertise. Future research, which largely takes place in the new Center of Excellence in Computational Inference (COIN; 2012 - 2017), will continue to be novel, innovative, as well as inter- and multi-disciplinary, with a specific focus on shared research activities that will have a significant societal impact.

The AIRC Centre of Excellence consisted of five interrelated research groups: Algorithms and Methods, Bioinformatics and Neuroinformatics, Multimodal Interfaces, Computational Cognitive Systems, and Adaptive Informatics Applications (see Figure 1.1).

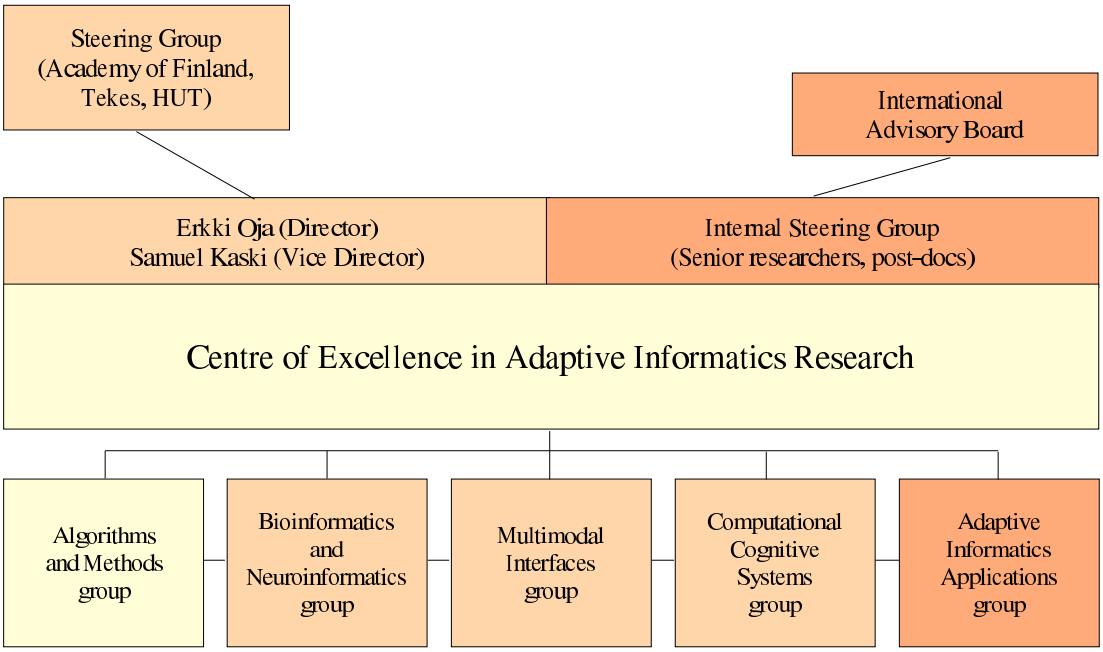


Figure 1.1: *The organization of the AIRC Centre of Excellence*

The Algorithms and Methods group conducted basic algorithmic research in adaptive informatics that relies heavily on computer science, mathematics and statistics, and was partly motivated by the research problems of other groups. In contrast, the groups of Bioinformatics and Neuroinformatics, Multimodal Interfaces and Computational Cognitive Systems formed an interdisciplinary research network with shared research interests in life and human sciences. The group of Adaptive Informatics Applications brought the research results into practice together with collaborating enterprises. This inter- and multi-disciplinary diversity facilitated a rich exchange of ideas, knowledge and expertise both within and between research groups. The ideas generated in one research group spark innovative ideas and research methods in other groups. This kind of ability to pool knowledge and resources between groups reduces duplication, saves time, and generates more powerful research methods and results. Altogether, it makes the Centre of Excellence a coherent whole. One proof of the success was that the core group of senior researchers, complemented by some other PI's, won already the fourth consequent Center of Excellence, COIN (Computational Inference).

Each group in AIRC and COIN has a wide range of national and international collaborators both in Academia and industry. Researcher training, graduate studies, and promotion of creative research is strongly emphasized, following the successful existing traditions.

The present Biennial Report 2010 - 2011 details the individual research projects of the five groups during the final two years of the six-year period of the AIRC. Additional information including demos etc. is available from our Web pages, `www.cis.hut.fi/research`.

Algorithms and Methods

Chapter 2

Bayesian learning of latent variable models

Juha Karhunen, Tapani Raiko, Alexander Ilin, Antti Honkela, Jaakko Luttinen, KyungHyun Cho

2.1 Bayesian modeling and variational learning

Unsupervised learning methods are often based on a generative approach where the goal is to find a latent variable model which explains how the observations were generated. It is assumed that there exist certain latent variables (also called in different contexts source signals, factors, or hidden variables) which have generated the observed data through an unknown mapping. The goal of generative learning is to identify both the latent variables and the unknown generative mapping.

The success of a specific model depends on how well it captures the structure of the phenomena underlying the observations. Various linear models have been popular, because their mathematical treatment is fairly easy. However, in many realistic cases the observations have been generated by a nonlinear process. Unsupervised learning of a nonlinear model is a challenging task, because it is typically computationally much more demanding than for linear models, and flexible models require strong regularization for avoiding overfitting.

In Bayesian data analysis and estimation methods, all the uncertain quantities are modeled in terms of their joint probability distribution. The key principle is to construct the joint posterior distribution for all the unknown quantities in a model, given the data sample. This posterior distribution contains all the relevant information on the parameters to be estimated in parametric models, or the predictions in non-parametric prediction or classification tasks [1, 2].

Denote by $\mathcal{H}$ the particular model under consideration, and by $\boldsymbol{\theta}$ the set of model parameters that we wish to infer from a given data set X. The posterior probability density $p(\boldsymbol{\theta}|X, \mathcal{H})$ of the parameters given the data X and the model $\mathcal{H}$ can be computed from the Bayes' rule

$$p(\boldsymbol{\theta}|X, \mathcal{H}) = \frac{p(X|\boldsymbol{\theta}, \mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})}{p(X|\mathcal{H})} \tag{2.1}$$

Here $p(X|\boldsymbol{\theta}, \mathcal{H})$ is the likelihood of the parameters $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathcal{H})$ is the prior pdf of the parameters, and $p(X|\mathcal{H})$ is a normalizing constant. The term $\mathcal{H}$ denotes all the assumptions made in defining the model, such as the choice of a particular model class and structure, specific noise model, etc.

The parameters $\boldsymbol{\theta}$ of a particular model $\mathcal{H}_i$ are often estimated by seeking the peak value of a probability distribution. The non-Bayesian maximum likelihood (ML) method uses to this end the distribution $p(X|\boldsymbol{\theta}, \mathcal{H})$ of the data, and the Bayesian maximum a posteriori (MAP) method finds the parameter values that maximize the posterior probability density $p(\boldsymbol{\theta}|X, \mathcal{H})$. However, using point estimates provided by the ML or MAP methods is often problematic, because the model order estimation and overfitting (choosing too complicated a model for the given data) are severe problems [1, 2].

Instead of searching for some point estimates, the correct Bayesian procedure is to use all possible models to evaluate predictions and weight them by the respective posterior probabilities of the models. This means that the predictions will be sensitive to regions where the probability mass is large instead of being sensitive to high values of the probability density [3, 2]. This procedure optimally solves the issues related to the model complexity and choice of a specific model $\mathcal{H}_i$ among several candidates. In practice, however, the differences between the probabilities of candidate model structures are often very large, and hence it is sufficient to select the most probable model and use the estimates or predictions

given by it.

A problem with fully Bayesian estimation is that the posterior distribution (2.1) has a highly complicated form except for in the simplest problems. Therefore it is too difficult to handle exactly, and some approximative method must be used. Variational methods form a class of approximations where the exact posterior is approximated with a simpler distribution [4, 2]. In a method commonly known as *Variational Bayes (VB)* [1, 3, 2] the misfit of the approximation is measured by the Kullback-Leibler (KL) divergence between two probability distributions $q(v)$ and $p(v)$. The KL divergence is defined by

$$D(q \parallel p) = \int q(v) \ln \frac{q(v)}{p(v)} dv \tag{2.2}$$

which measures the difference in the probability mass between the densities $q(v)$ and $p(v)$.

A key idea in the VB method is to minimize the misfit between the actual posterior pdf and its parametric approximation using the KL divergence. The approximating density is often taken a diagonal multivariate Gaussian density, because the computations become then tractable. Even this crude approximation is adequate for finding the region where the mass of the actual posterior density is concentrated. The mean values of the Gaussian approximation provide reasonably good point estimates of the unknown parameters, and the respective variances measure the reliability of these estimates.

A main motivation of using VB is that it avoids overfitting which would be a difficult problem if ML or MAP estimates were used. VB method allows one to select a model having appropriate complexity, making often possible to infer the correct number of latent variables or sources. It has provided good estimation results in the very difficult unsupervised (blind) learning problems that we have considered.

Variational Bayes is closely related to information theoretic approaches which minimize the description length of the data, because the description length is defined to be the negative logarithm of the probability. Minimal description length thus means maximal probability. In the probabilistic framework, we try to find the latent variables or sources and the nonlinear mapping which most probably correspond to the observed data. In the information theoretic framework, this corresponds to finding the latent variables or sources and the mapping that can generate the observed data and have the minimum total complexity. This information theoretic view also provides insights to many aspects of learning and helps to explain several common problems [5].

During the last two years, our research has extended to deep learning, which is not a Bayesian but a probabilistic latent variable analysis method. In deep learning one tries to find hierarchical representations of data, starting from observations towards more and more abstract representations. Deep learning can be cumbersome and difficult but on the other hand it can provide world record results in difficult classification problems. We have improved deep learning algorithms, making them more stable and robust against the choice of learning parameters. Deep learning is discussed in this chapter in its own subsection.

In the following subsections, we first discuss improvements in variational Bayesian learning, including a natural conjugate gradient algorithm which speeds up learning remarkably, as well as transformations of latent variables leading also to a faster convergence. After this we consider extensions of probabilistic principal component analysis (PCA) for treating missing values and achieving robustness in the presence of outliers. We then consider

time series modeling in bioinformatics to learn gene regulatory relationships from time series expression data. Our contributions to deep learning and Boltzmann machines are discussed in the next section. Finally, we have carried out some work on oscillatory neural networks, and applied our Bayesian methods to novelty detection in structural health monitoring and document classification utilizing relational information.

2.2 Algorithmic improvements for variational inference

Riemannian conjugate gradient

Variational methods for approximate inference in machine learning often adapt a parametric probability distribution to optimize a given objective function. This view is especially useful when applying variational Bayes (VB) to models outside the conjugate-exponential family. For them, variational Bayesian expectation maximization (VB EM) algorithms are not easily available, and gradient-based methods are often used as alternatives.

In previous machine learning algorithms based on natural gradients [6], the aim has been to use maximum likelihood to directly update the model parameters $\boldsymbol{\theta}$ taking into account the geometry imposed by the predictive distribution for data $p(\boldsymbol{X}|\boldsymbol{\theta})$. The resulting geometry is often very complicated as the effects of different parameters cannot be separated and the Fisher information matrix is relatively dense.

Recently, in [7], we propose using natural gradients for free energy minimisation in variational Bayesian learning using the simpler geometry of the approximating distributions $q(\boldsymbol{\theta}|\boldsymbol{\xi})$. Because the approximations are often chosen to minimize dependencies between different parameters $\boldsymbol{\theta}$, the resulting Fisher information matrix with respect to the variational parameters $\boldsymbol{\xi}$ will be mostly diagonal and hence easy to invert.

While taking into account the structure of the approximation, plain natural gradient in this case ignores the structure of the model and the global geometry of the parameters $\boldsymbol{\theta}$. This can be addressed by using conjugate gradients. Combining the natural gradient search direction with a conjugate gradient method yields our proposed *approximate Riemannian conjugate gradient (RCG)* method.

The RCG algorithm was compared against conjugate gradient (CG) and Riemannian gradient (RG) algorithms in learning a nonlinear state-space model [8]. The results for a number of datasets ranging from 200 to 500 samples of 21 dimensional speech spectrograms can be seen in Figure 2.1. The plain CG and RG methods were clearly slower than others and the maximum runtime of 24 hours was reached by most CG and some RG runs. RCG was clearly the fastest algorithm with the older heuristic method of [8] between these extremes. The results with a larger data set are very similar with RCG outperforming all alternatives by a factor of more than 10.

The experiments in [7] show that the natural conjugate gradient method outperforms both conjugate gradient and natural gradient methods by a large margin. Considering univariate Gaussian distributions, the regular gradient is too strong for model variables with small posterior variance and too weak for variables with large posterior variance. The posterior variance of latent variables is often much larger than the posterior variance of model parameters and the natural gradient takes this into account in a very natural manner.

Transformation of latent variables

Variational methods have been used for learning linear latent variable models in which observed data vectors $\mathbf{x}(t)$ are modeled as linear combination of latent variables $\mathbf{s}(t)$:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \boldsymbol{\mu} + \mathbf{n}(t), \quad t = 1, \ldots, N. \tag{2.3}$$

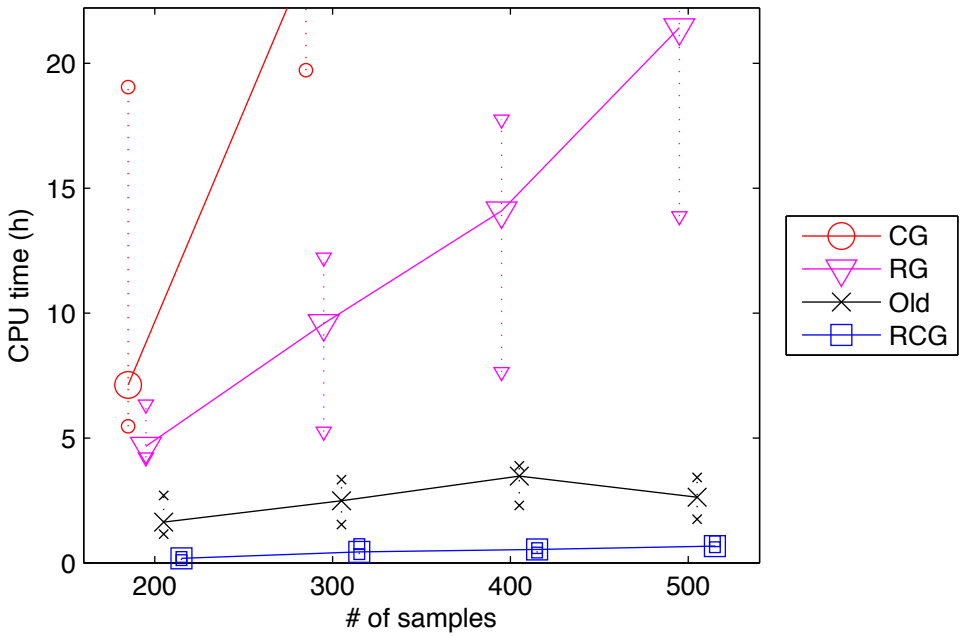


Figure 2.1: Convergence speed of the Riemannian conjugate gradient (RCG), the Riemannian gradient (RG) and the conjugate gradient (CG) methods as well as the heuristic algorithm (Old) with different data sizes. The lines show median times with 25 % and 75 % quantiles shown by the smaller marks. The times were limited to at most 24 hours, which was reached by a number of simulations.

The latent variables are assigned some prior distributions, such as zero-mean Gaussian priors with uncorrelated components in the basic factor analysis model. When VB learning is used, the true posterior probability density function (pdf) of the unknown variables is approximated using a tractable pdf factorized as follows:

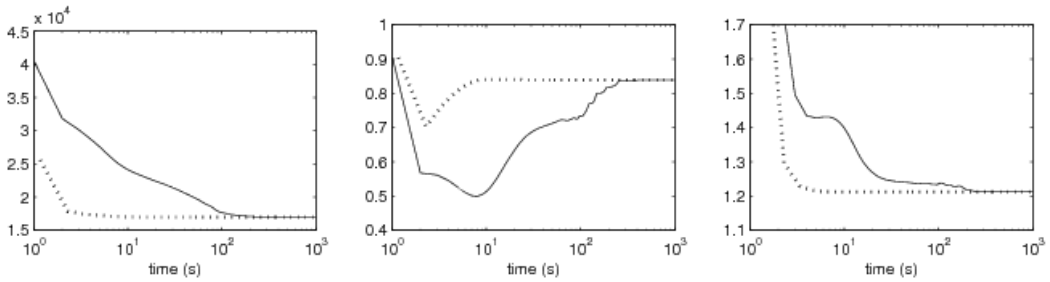
$$p\left(\boldsymbol{\mu}, \mathbf{A}, \mathbf{s}(1), ..., \mathbf{s}(N) \mid \{\mathbf{x}(t)\}\right) \approx q(\boldsymbol{\mu})q(\mathbf{A})q(\mathbf{s}(1)) \ldots q(\mathbf{s}(N)) \,.$$

This form of the posterior approximation q ignores the strong correlations present between the variables, which often causes slow convergence of VB learning.

Parameter-expanded VB (PX-VB) methods were recently proposed to address the slow convergence problem [9]. The general idea is to use auxiliary parameters in the original model to reduce the effect of strong couplings between different variables. The auxiliary parameters are optimized during learning, which corresponds to *joint* optimization of different components of the variational approximation of the true posterior. In this way strong functional couplings between the components are reduced and faster convergence is facilitated. One of the main challenges for applying the PX-VB methodology is to use proper reparameterization of the original model.

In our journal paper [10], we present a similar idea in the context of VB learning of factor analysis models. There we use auxiliary parameters $\mathbf{b}$ and $\mathbf{R}$ which translate and rotate the latent variables:

$$\begin{aligned} \mathbf{s}(t) &\leftarrow \mathbf{s}(t) - \mathbf{b} & \boldsymbol{\mu} &\leftarrow \boldsymbol{\mu} + \mathbf{A}\mathbf{b} \\ \mathbf{s}(t) &\leftarrow \mathbf{R}\mathbf{s}(t) & \mathbf{A} &\leftarrow \mathbf{A}\mathbf{R}^{-1} \,. \end{aligned}$$



(a) VB cost (b) Training RMSE (c) Test RMSE

Figure 2.2: Convergence of VB PCA tested on artificial data. The dotted and solid curves represent the results with and without the proposed transformations, respectively.

The optimal parameters $\mathbf{b}$ and $\mathbf{R}$ which minimize the misfit between the posterior pdf and its approximation can then be computed analytically. This corresponds to joint optimization of factors $q(\mathbf{s}(t))$. In our paper, we show that the proposed transformations essentially perform centering and whitening of the hidden factors taking into account their posterior uncertainties.

We tested the effect of the proposed transformations by applying the VB PCA model to an artificial dataset consisting of $N = 200$ samples of normally distributed 50-dimensional vectors $\mathbf{x}(t)$. Figure 2.2 shows the minimized VB cost and the root mean squared error (RMSE) computed on the training and test sets during learning. The curves indicate that the method first overfits providing a solution with an unreasonably small RMSE. Later, learning proceeds toward a better solution yielding smaller test RMSE. Note that using the proposed transformations reduced the overfitting effect at the beginning of learning, which led to faster convergence to the optimal solution.

2.3 Extensions of probabilistic PCA

PCA of large-scale datasets with many missing values

Principal component analysis (PCA) is a classical data analysis technique. Some algorithms for PCA scale better than others to problems with high dimensionality. They also differ in the ability to handle missing values in the data. In our recent paper [11], a case is studied where the data are high-dimensional and a majority of the values are missing. In the case of very sparse data, overfitting becomes a severe problem even in simple linear models such as PCA. Regularization can be provided using the Bayesian approach by introducing prior for the model parameters. The PCA model can then be identified using, for example, maximum a posteriori estimates (MAPPCA) or variational Bayesian (VBPCA) learning.

In [11], we study different approaches to PCA for incomplete data. We show that faster convergence can be achieved using the following rule for the model parameters:

$$\theta_i \leftarrow \theta_i - \gamma \left(\frac{\partial^2 C}{\partial \theta_i^2} \right)^{-\alpha} \frac{\partial C}{\partial \theta_i},$$

where α is a control parameter that allows the learning algorithm to vary from the standard gradient descent ($\alpha = 0$) to the diagonal Newton's method ($\alpha = 1$). These learning rules can be used for standard PCA learning and extended to MAPPCA and VBPCA.

The algorithms were tested on the Netflix problem (`http://www.netflixprize.com/`), which is a task of predicting preferences (or producing personal recommendations) by using other people's preferences. The Netflix problem consists of movie ratings given by 480189 customers to 17770 movies. There are 100480507 ratings from 1 to 5 given, and the task is to predict 2817131 other ratings among the same group of customers and movies. 1408395 of the ratings are reserved for validation. Thus, 98.8% of the values are missing.

We used different variants of PCA in order to predict the test ratings in the Netflix data set. The obtained results are shown in Figure 2.3. The best accuracy was obtained using VB PCA with a simplified form of the posterior approximation (VBPCAd in Figure 2.3). That method was also able to provide reasonable estimates of the uncertainties of the predictions.

Robust PCA for incomplete data

Standard PCA is known to be sensitive to outliers in the data because it is based on minimisation of a quadratic criterion such as the mean-square representation error. Thus, corrupted or atypical observations may cause the failure of PCA, especially for data sets with missing values. A standard way to cope with this problem is replacing the quadratic cost function of PCA a function which grows more slowly.

In [12], we present a new robust PCA model based on the Student-t distribution and show how it can be identified for data sets with missing values. We make the assumption that the outliers can arise independently in each sensor (i.e. for each dimension of a data vector). This assumption is different to the previously introduced techniques [13] and it turns out to be important for modeling incomplete data sets. The proposed model can improve the quality of the principal subspace estimation and provide better reconstructions of missing

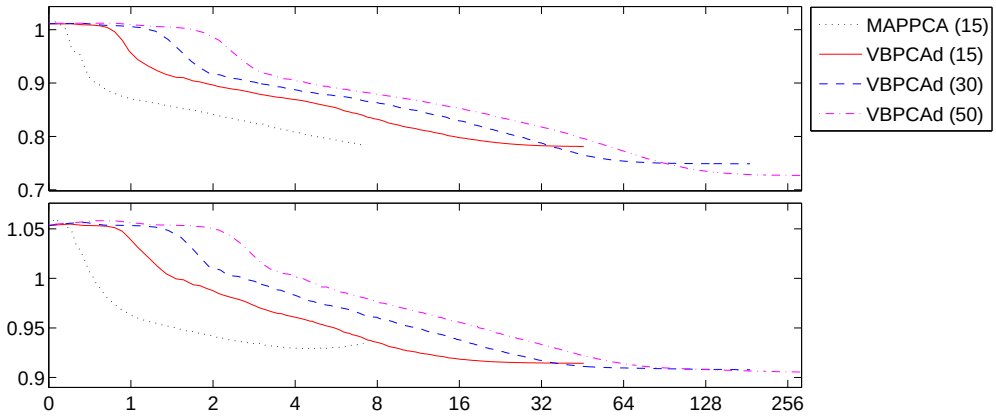


Figure 2.3: Root mean squared errors for the Netflix data (y-axis) plotted against the processor time in hours. The upper plot shows the training error while the lower plot shows the error for the probing data provided by Netflix. The time scale is linear from 0 to 1 and logarithmic above 1.

values. The model can also be used to remove outliers by estimating the true values of their corrupted components from the uncorrupted ones.

We tested the robust PCA model on the Helsinki Testbed data set which at the moment of our studies contained many atypical measurements and missing values. The model was used to estimate four principal components of the temperature measurements from 79 stations in Southern Finland. Figure 2.4 presents the reconstruction of the data using our robust PCA model for four different stations. The reconstructions look very reasonable with most of the outliers being removed.

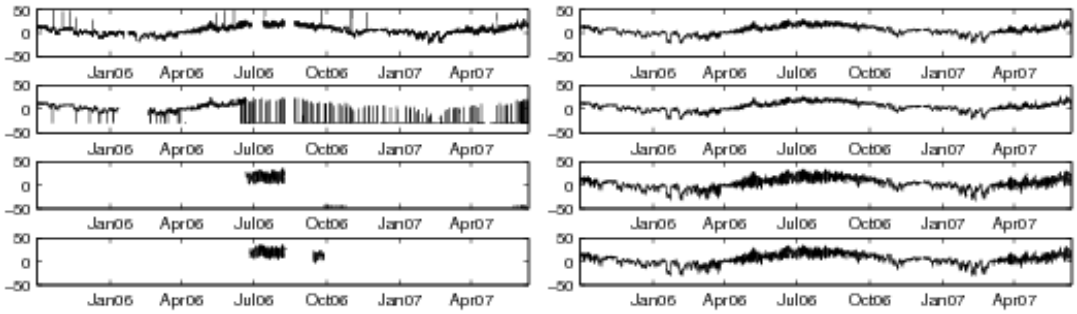


Figure 2.4: Four example signals from the Helsinki Testbed dataset and their reconstructions using the proposed robust PCA.

The results of this study are presented in more detail in the journal manuscript [14]. More traditional methods for robust PCA, also with missing values, have been studied in [15]. They are usually much easier to apply compared with Bayesian methods but less effective.

2.4 Gaussian process models of gene expression and gene regulation

Bayesian methods are well-suited for analysis of molecular biology data as the data sets practically always consist of very few samples with a high noise level. We have studied models of gene transcription regulation based on time series gene expression data in collaboration with Neil D. Lawrence and Magnus Rattray of the University of Sheffield. This is a very challenging modelling task as the time series are very short, typically at most a dozen time points.

Extending the model of [16] of single input motif systems, i.e. where a single transcription factor regulates a number of genes, we have developed a method of ranking putative targets of transcription factors based on expression data [17]. This is achieved by imposing a Gaussian process (GP) prior on the latent continuous time transcription factor gene expression profile, which drives a linear ODE model of transcription factor protein translation and target gene transcription. This linear ODE model leads to a joint GP model for all observable gene expression values and allows exact marginalisation of the latent functions. Candidate target genes can be ranked using model likelihood.

We have applied the model to genome-wide ranking of potential target genes of transcription factors. Fig. 2.5 shows results from experiments with key regulators of *Drosophila* mesoderm and muscle development. They show very high accuracy in terms of enrichment of detected transcription factor binding near the predicted target genes [17]. An implementation of the method is available in Bioconductor for R [18].

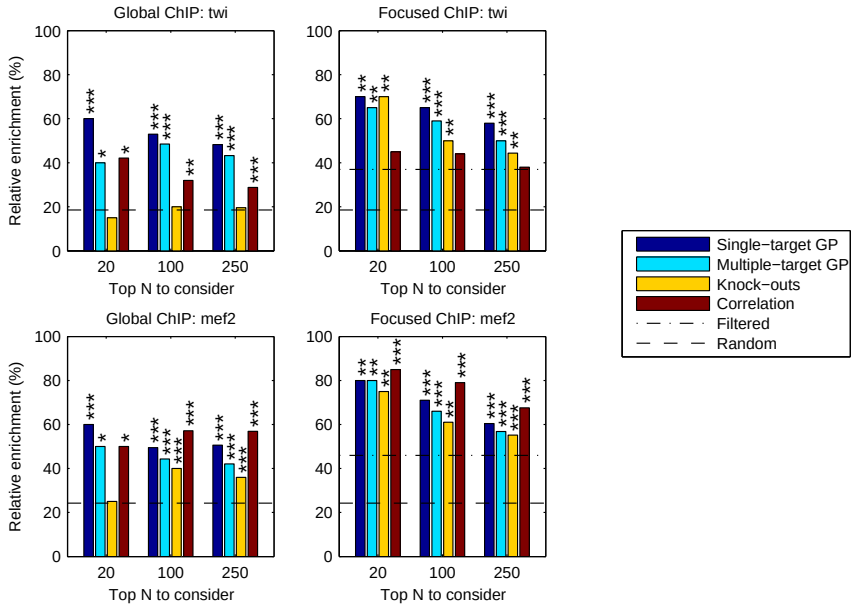


Figure 2.5: Evaluation results from [17] of two variants of the proposed GP-based ranking methods and two alternatives showing the relative frequency of positive predictions among N top-ranking targets ("global" evaluations) and among N top genes with annotated expression in mesoderm or muscle tissue ("focused" evaluations) for two studied transcription factors. The dashed line denotes the frequency in the full population and the dash-dot line within the population considered in focused evaluation. The bars show the frequency of targets with ChIP-chip binding within 2000 base pairs of the gene. p-values of results significantly different from random are denoted by '***': $p < 0.001$, '**': $p < 0.01$, '*': $p < 0.05$.

2.5 Deep learning and Boltzmann machines

Deep learning has gained its popularity recently as a way of learning complex and large probabilistic models [25]. Especially, deep neural networks such as a deep belief network and a deep Boltzmann machine have been applied to various machine learning tasks with impressive improvements over conventional approaches.

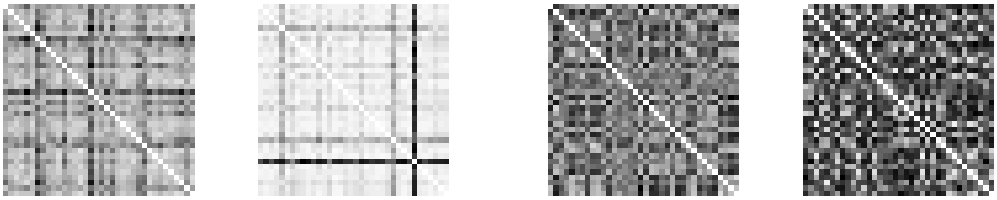
Deep neural networks are characterized by the large number of layers of neurons and by using layer-wise unsupervised pretraining to learn a probabilistic model for the data. A deep neural network is typically constructed by stacking multiple restricted Boltzmann machines (RBM) so that the hidden layer of one RBM becomes the visible layer of another RBM. Layer-wise pretraining of RBMs then facilitates finding a more accurate model for the data. Various papers (see, e.g., [26], [25] and references therein) empirically confirmed that such multi-stage learning works better than conventional learning methods.

Unfortunately, even training a simple RBM which consists of only two layers of visible and hidden neurons is known to be difficult [31, 32]. This problem is often evidenced by the decreasing likelihood during learning. These failures have discouraged using RBMs and its extensions such as deep Boltzmann machines for more sophisticated and variety of machine learning tasks.

In our recent conference papers [28], we have proposed to use parallel tempering, an ad-

vanced Markov-chain Monte-Carlo sampling, as a replacement of a simple Gibbs sampling in obtaining samples from a model distribution defined by an RBM. It was shown that a better model with higher log-likelihood could be found using the stochastic gradient method based on PT compared to a widely-used method of minimizing contrastive divergence.

Additionally to the problem of using a simple Gibbs sampling we have determined other possible problems that discourage using an RBM as a building block for building a deep neural network. In [29] we identified density of training samples and learning hyper-parameters, such as a learning rate and an initialization of parameters, as two sources of difficulty in training RBMs. Furthermore, we also discovered that the conventional form of an energy function of Gaussian-Bernoulli RBM (GRBM) is defected in some sense that learning becomes easily unstable, in [27].



Traditional Gradient Enhanced Gradient

Figure 2.6: The angles between the update directions for the weights of an RBM with 36 hidden neurons. White pixels correspond to small angles, while black pixels correspond to orthogonal directions. From left to right: traditional gradient after 26 updates, traditional gradient after 364 updates, enhanced gradient after 26 updates, and enhanced gradient after 364 updates.

We have derived a new update direction for training RBMs, called enhanced gradient, in [29]:

$$w_{ij} \leftarrow w_{ij} + \eta_w \nabla_e \, w_{ij} \tag{2.4}$$

$$b_i \leftarrow b_i + \eta_b \nabla_e \, b_i \tag{2.5}$$

$$c_j \leftarrow c_j + \eta_c \nabla_e \, c_j, \tag{2.6}$$

where w_{ij}, b_i and c_j are weight between a visible neuron i and a hidden neuron j and biases for a visible neurons i and a hidden neuron j, respectively.

The enhanced gradient makes learning based on the stochastic gradient invariant to the density of training samples as well as the sparsity of hidden neurons. It turned out that the enhanced gradient is more robust to the choice of learning hyper-parameters and makes the gradient per hidden neuron more orthogonal to each other as can be see in Figure 2.6. It was shown to help avoid a common degenerate case where most hidden neurons learn a bias.

Also in [29], we proposed a new adaptation mechanism, call adaptive learning rate, for choosing a learning rate on-the-fly. The adaptive learning rate greedily adapts the learning rate while learning parameters by maximizing the locally estimated log-likelihood. Together with the enhanced gradient, it shows in Figure 2.7 that more stable and better models can be trained.

All three approaches– parallel tempering, the enhanced gradient, and the adpative learning rate– have been shown to work with extensions of RBMs. In [27], we showed that these

methods can be directly applied to a GRBM which replaces a binary visible neuron of an RBM with a Gaussian neuron. Furthermore, we showed that a hierarchical version of Boltzmann machines called deep Boltzmann machines (DBM) can readily use the proposed approaches in [30].

Additionally to studying Boltzmann machines for deep learning, a method of transforming a standard multi-layer perceptron by introducing linear shortcut connections and proposing transformations in non-linearities was proposed in [33]. It was shown in the paper that with the proposed transformations a faster convergence to a state-of-the-art performance can be achieved.

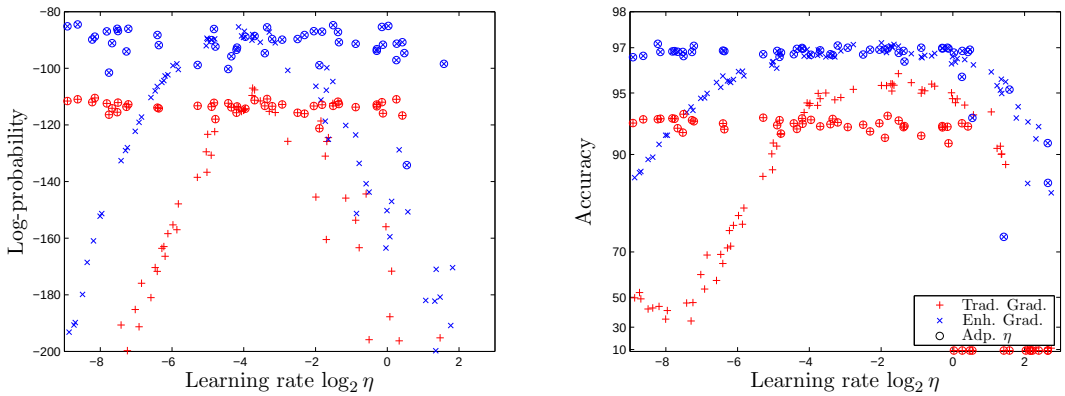


Figure 2.7: Log-probabilities and classificiation accuracies of test data for different initializations of the learning rate. The models were trained on MNIST using the stochastic gradient with parallel tempering.

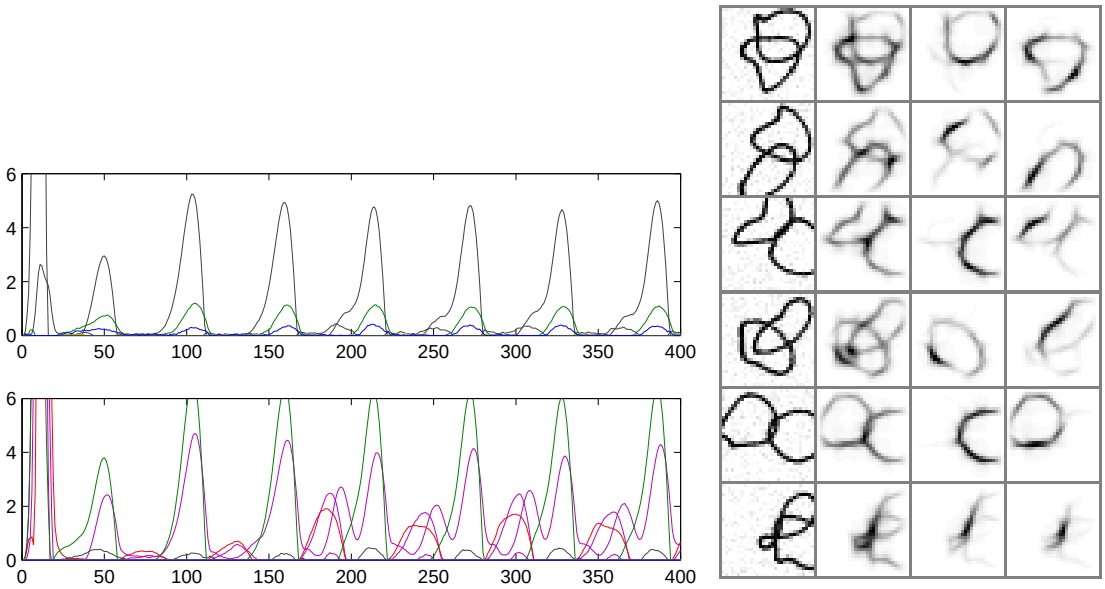


Figure 2.8: Left: The neural signals corresponding to the middle patch (top) and the patch below it (bottom) plotted as a function of time. The activities are given for the top-most data sample. Right: The segmentation result obtained from NMF analysis of the signals. First column from the left is the data, second column is the reconstruction from the feature activities, third and fourth columns are segmented objects.

2.6 Oscillatory neural networks

In [34] we studied the emergent properties of an artificial neural network which combines segmentation by oscillations and biased competition for perceptual processing. The aim was to progress in image segmentation by mimicking abstractly the way how the cerebral cortex works. In our model, the neurons associated with features belonging to an object start to oscillate synchronously, while competing objects oscillate with an opposing phase.

The overall structure of our network is such that there are so called areas that correspond to patches in the image. The areas get bottom-up input from the pixels. The areas should be connected to each other with local interactions only, that is, there is no hierarchy or global signals. The different areas should work in the same way, using the same algorithms. The emergent properties of the network are confirmed by experiments with artificial image data as seen in Figure 2.8.

2.7 Applications in climate science

We applied the Bayesian methodology for several problems in climate science.

In our papers [19, 21], we consider the problem of historical reconstruction of climate fields, which is a problem of infilling missing values in the observational data. We take the statistical approach and propose a probabilistic model called Gaussian-process factor analysis (GPFA). The model is based on standard matrix factorization

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \text{noise} = \sum_{d=1}^{D} \mathbf{w}_{:d}\mathbf{x}_{d:}^{\mathrm{T}} + \text{noise},$$

where $\mathbf{Y}$ is a data matrix in which each row contains measurements in one spatial location and each column corresponds to one time instance. The goal is to learn the model parameters $\mathbf{W}$, $\mathbf{X}$ from available observations in order to reconstruct the missing values in $\mathbf{Y}$. Each $\mathbf{x}_{d:}$ is a vector representing the time series of one of the D factors whereas $\mathbf{w}_{:d}$ is a vector of loadings which are spatially distributed. We assume that both factors $\mathbf{x}_{d:}$ and corresponding loadings $\mathbf{w}_{:d}$ have prominent structures that we model using the Gaussian process methodology [20]. The model is identified in the framework of variational Bayesian learning and high computational cost of GP modeling is reduced by using sparse approximations derived in the variational methodology.

Another problem studied in our group is parametric tuning of climate models. Climate models contain closure parameters which can act as effective "tuning handles" of the simulated climate. These appear in physical parameterization schemes where unresolved variables are expressed by predefined parameters rather than being explicitly modeled. In the current climate model tuning process, best expert knowledge is used to define the optimal closure parameter values, based on observations, process studies, large eddy simulations, etc.

Our research group participates in the Academy of Finland project called "Novel advanced mathematical and statistical methods for understanding climate" (NOVAC, 2010-2013), whose goal is to develop algorithmic ways for closure parameter estimation. We focus on the atmospheric model ECHAM5 but the methodology is generic and applicable in any multi-scale problem with similar closure parameters [22].

The uncertainties of the closure parameters are estimated using Markov chain Monte Carlo (MCMC) simulations [23]. The MCMC approach is, however, computationally very expensive and only maximally optimized MCMC techniques make the approach realistic in practice. We develop new tools based on adaptive algorithms, multiple computational grids, parallel chains as well as methods based on early rejection.

The central problem in closure parameter estimation is how to formulate the likelihood function. This task is not trivial because of the chaotic nature of climate models. Climate model simulations quickly diverge from observations, which makes classical parameter estimation based on direct comparison of model simulations and observations inefficient. Our initial approach to circumvent the chaoticity problem was to formulate the likelihood function in terms of summary statistics. In [23], the likelihood is evaluated by comparing some temporal and spatial averages of observed and simulated data. Several summary statistics potentially useful for climate model tuning have been studied in [24].

2.8 Other Applications

We applied nonlinear factor analysis to novelty detection for structural health monitoring in [35]. In vibration-based structural health monitoring damage in structure is tried to detect from damage-sensitive features. Because neither prior information nor data about expected damage are normally available, damage detection problem must be solved by using a novelty detection approach. Features, which are sensitive to damage, are often sensitive to environmental and operational variations. Therefore elimination of these variations is essential for reliable damage detection. At present many of the damage detection methods are linear, though it has been shown that many of the vibration changes in structures are bilinear or nonlinear. We proposed to use nonlinear factor analysis to detect damage via elimination of external effectsfrom damage features. The effectiveness of the proposed method was demonstrated by analyzing the experimental Z24 Bridge data with a comparison to a linear method [35]. It was shown that elimination of adverse effects and damage detection are feasible.

In [36], we studied document classification utilising relational information. Two major types of relational information can be utilized in automatic document classification as background information: relations between terms, such as ontologies, and relations between documents, such as web links or citations in articles. We introduced a model where a traditional bag-of-words type classifier is gradually extended to utilize both of these information types. The experiments with data from the Finnish National Archive show that classification accuracy improves from 70% to 74% when the General Finnish Ontology YSO is used as background information, without using relations between documents.

References

[1] D. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[2] C. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[3] H. Lappalainen and J. Miskin. Ensemble learning. In M. Girolami, editor, *Advances in Independent Component Analysis*, Springer, 2000, pages 75–92.

[4] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In M. Jordan, editor, *Learning in Graphical Models*, MIT Press, 1999, pages 105–161.

[5] A. Honkela and H. Valpola. Variational learning and bits-back coding: an information-theoretic view to Bayesian learning. *IEEE Transactions on Neural Networks*, 15(4):267–282, 2004.

[6] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[7] A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen. Approximate Riemannian Conjugate Gradient Learning for Fixed-Form Variational Bayes. In Journal of Machine Learning Research (JMLR), volume 11, pages 3235–3268, November 2010.

[8] H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692, 2002.

[9] Y. Qi, T. S. Jaakkola. Parameter expanded variational Bayesian methods. In *Advances in Neural Information Processing Systems 19*, pp. 1097–1104, Cambridge, MA, 2007.

[10] J. Luttinen and A. Ilin. Transformations in variational Bayesian factor analysis to speed up learning. *Neurocomputing*, 73(7-9):1093–1102, 2010.

[11] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. In Journal of Machine Learning Research (JMLR), volume 11, pages 1957–2000, July, 2010.

[12] J. Luttinen, A. Ilin, and Juha Karhunen. Bayesian robust PCA for incomplete data. In *Proc. of the 8th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2009)*, pp. 66–73, Paraty, Brazil, March 2009.

[13] C. Archambeau, N. Delannay, M. Verleysen. Robust probabilistic projections. In *Proc. of the 23rd International Conference on Machine Learning (ICML 2006)*, pp. 33-40, New York, NY, USA, 2006.

[14] J. Luttinen, A. Ilin, and Juha Karhunen. Bayesian robust PCA for incomplete data. Revised version submitted to *Neural Processing Letters*, 2012.

[15] J. Karhunen. Robust PCA methods for complete and missing data. *Neural Network World*, 21(5):357–392, 2011.

[16] P. Gao, A. Honkela, M. Rattray, and N. D. Lawrence. Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics* 24(16):i70–i75, 2008.

[17] A. Honkela, C. Girardot, E. H. Gustafson, Y.-H. Liu, E. E. M. Furlong, N. D. Lawrence, and M. Rattray. Model-based method for transcription factor target identification with limited data. *Proc Natl Acad Sci U S A* 107(17):7793–7798, 2010.

[18] A. Honkela, P. Gao, J. Ropponen, M. Rattray, and N. D. Lawrence. tigre: Transcription factor Inference through Gaussian process Reconstruction of Expression for Bioconductor. *Bioinformatics* 27(7):1026–1027, 2011.

[19] J. Luttinen and A. Ilin. Variational Gaussian-process factor analysis for modeling spatio-temporal data. In *Advances in Neural Information Processing Systems 22*, 2009.

[20] C. E. Rasmussen, C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.

[21] A. Ilin and J. Luttinen. Variational Gaussian-process factor analysis for modeling spatio-temporal data. In *Proc. of the 11th Meeting on Statistical Climatology*, Edinburgh, Scotland, 2010.

[22] H. Haario, E. Oja, A. Ilin, H. Järvinen and J. Tamminen Novel advanced mathematical and statistical methods for understanding climate (NOVAC). In *Proc. of the 11th Meeting on Statistical Climatology*, Edinburgh, Scotland, 2010.

[23] H. Järvinen, P. Räisänen, M. Laine, J. Tamminen, A. Ilin, E. Oja, A. Solonen, and H. Haario. Estimation of ECHAM5 climate model closure parameters with adaptive MCMC. *Atmospheric Chemistry and Physics Discussion*, Vol. 10, No 5, pp. 11951–11973, 2010.

[24] J. Saarimäki. Performance Metrics for the Atmospheric Model ECHAM5. Master's thesis, Aalto University, 2010

[25] Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, 2:1–127.

[26] Salakhutdinov, R. (2009). *Learning Deep Generative Models*. PhD thesis, University of Toronto.

[27] Cho, K., Ilin, A., and Raiko, T. (2011a). Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines. In *Proceedings of the Twentith International Conference on Artificial Neural Networks*, ICANN 2011.

[28] Cho, K., Raiko, T., and Ilin, A. (2010). Parallel Tempering is Efficient for Learning Restricted Boltzmann Machines. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2010)*, pages 3246 – 3253, Barcelona, Spain.

[29] Cho, K., Raiko, T., and Ilin, A. (2011b). Enhanced Gradient and Adaptive Learning Rate for Training Restricted Boltzmann Machines. In *Proceedings of the Twenty-seventh International Conference on Machine Learning*, ICML 2011.

[30] Cho, K., Raiko, T., and Ilin, A. (2011c). Gaussian-bernoulli deep boltzmann machine. In *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, Sierra Nevada, Spain.

[31] Fischer, A. and Igel, C. (2010). Empirical analysis of the divergence of Gibbs sampling based learning algorithms for restricted Boltzmann machines. In *Proceedings of the 20th international conference on Artificial neural networks: Part III*, ICANN'10, pages 208–217, Berlin, Heidelberg. Springer-Verlag.

[32] Schulz, H., Müller, A., and Behnke, S. (2010). Investigating Convergence of Restricted Boltzmann Machine Learning. In *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*.

[33] Raiko, T., Valpola, H., and LeCun, Y. (2011). Deep Learning Made Easier by Linear Transformations in Perceptrons. In *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, Sierra Nevada, Spain.

[34] T. Raiko and H. Valpola. Chapter 7: Oscillatory Neural Network for Image Segmentation with Biased Competition for Attention. In From Brains to Systems: Brain-Inspired Cognitive Systems 2010, Advances in Experimental Medicine and Biology, volume 718, pages 75–86, Springer New York, 2011.

[35] V. Lämsä and T. Raiko. Novelty Detection by Nonlinear Factor Analysis for Structural Health Monitoring. In IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010), pages 468–473, Kittilä, Finland, August, 2010.

[36] K. Nyberg, T. Raiko, T. Tiinanen, and E. Hyvönen. Document Classification Utilising Ontologies and Relations between Documents, In Proceedings of the Eighth Workshop on Mining and Learning with Graphs (MLG 2010), Washington DC, USA, July, 2010.

Chapter 3

Blind and semi-blind source separation

Erkki Oja, Juha Karhunen, Alexander Ilin, Zhirong Yang, Jaakko Luttinen, He Zhang, Jarkko Ylipaavalniemi, Tele Hao

3.1 Introduction

Erkki Oja

What is Blind and Semi-blind Source Separation? Blind source separation (BSS) is a class of computational data analysis techniques for revealing hidden factors, that underlie sets of measurements or signals. BSS assumes a statistical model whereby the observed multivariate data, typically given as a large database of samples, are assumed to be linear or nonlinear mixtures of some unknown latent variables. The mixing coefficients are also unknown.

By BSS, these latent variables, also called sources or factors, can be found. Thus BSS can be seen as an extension to the classical methods of Principal Component Analysis and Factor Analysis. BSS is a much richer class of techniques, however, capable of finding the sources when the classical methods, implicitly or explicitly based on Gaussian models, fail completely.

In many cases, the measurements are given as a set of parallel signals or time series. Typical examples are mixtures of simultaneous sounds or human voices that have been picked up by several microphones, brain signal measurements from multiple EEG sensors, several radio signals arriving at a portable phone, or multiple parallel time series obtained from some industrial process. But BSS has other applications as well: it turns out that clustering, or finding mutually similar subsets of the dataset, can also be addressed as a linear source separation problem, when suitable constraints are added to the model. This also applies to the related problem of graph partitioning.

Perhaps the best known single methodology in BSS is Independent Component Analysis (ICA), in which the latent variables are nongaussian and mutually independent. However, also other criteria than independence can be used for finding the sources. One such simple criterion is the non-negativity of the sources. Sometimes more prior information about the sources is available or is induced into the model, such as the form of their probability densities, their spectral contents, etc. Then the term "blind" is often replaced by "semiblind".

Our earlier contributions in ICA research. In our ICA research group, the research stems from some early work on on-line PCA, nonlinear PCA, and separation, that we were involved with in the 80's and early 90's. Since mid-90's, our ICA group grew considerably. This earlier work has been reported in the previous Triennial and Biennial reports of our laboratory from 1994 to 2009 [1]. A notable achievement from that period was the textbook "Independent Component Analysis" by A. Hyvärinen, J. Karhunen, and E. Oja [2]. It has been very well received in the research community; according to the latest publisher's report, over 6000 copies had been sold by August, 2011. The book has been extensively cited in the ICA literature and seems to have evolved into the standard text on the subject worldwide. In Google Scholar, the de facto standard for citations in the ICT field, the book has received over 6700 citations (April 2012). In 2005, the Japanese translation of the book appeared (Tokyo Denki University Press), and in 2007, the Chinese translation (Publishing House of Electronics Industry).

Another tangible contribution has been the public domain FastICA software package [3]. This is one of the few most popular ICA algorithms used by the practitioners and a

standard benchmark in algorithmic comparisons in ICA literature.

In the reporting period 2010 - 2011, ICA/BSS research stayed as one of the core projects in the laboratory, with the pure ICA theory waning and being replaced by several new directions in blind and semiblind source separation. In this Chapter, we present two such novel directions.

Section 3.2 introduces some theoretical advances on Nonnegative Matrix Factorization undertaken during the reporting period, especially on the new Projective Nonnegative Matrix Factorization (PNMF) principle, which is a principled way to perform approximate nonnegative Principal Component Analysis.

Section 3.3 introduces novel results in finding independent and dependent sources from two related data sets. It is based on a combination of Canonical Correlation Analysis and ICA.

Quite another way to formulate the BSS problem is Bayesian analysis. This is covered in the separate Chapter **??**.

References

[1] Triennial and Biennial reports of CIS and AIRC. `http://www.cis.hut.fi/research/reports/`.

[2] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. J. Wiley, 2001.

[3] The FastICA software package. `http://www.cis.hut.fi/projects/ica/fastica/`.

3.2 Non-negative Low-Rank Learning

Zhirong Yang, He Zhang, Zhanxing Zhu, and Erkki Oja

Enforcing nonnegativity in linear factorizations [4] has proven to be a powerful principle for multivariate data analysis, especially sparse feature analysis, as shown by the well-known Nonnegative Matrix Factorization (NMF) algorithm by Lee and Seung [2]. Their method minimizes the difference between the data matrix $\mathbf{X}$ and its non-negative decomposition $\mathbf{WH}$. Yuan and Oja [11] proposed the Projective NMF (PNMF) method which replaces $\mathbf{H}$ in NMF with $\mathbf{W}^T\mathbf{X}$. Empirical results indicate that PNMF is able to produce more spatially localized, part-based representations of visual patterns.

Recently, we have extended and completed the preliminary work with the following new contributions [5]: (1) formal convergence analysis of the original PNMF algorithms, (2) PNMF with the orthonormality constraint, (3) nonlinear extension of PNMF, (4) comparison of PNMF with two classical and two recent algorithms [10, 1] for clustering, (5) a new application of PNMF for recovering the projection matrix in a nonnegative mixture model, (6) comparison of PNMF with the approach of discretizing eigenvectors, and (7) theoretical justification of moving a term in the generic multiplicative update rule. Our in-depth analysis shows that the PNMF replacement has positive consequences in sparseness of the approximation, orthogonality of the factorizing matrix, decreased computational complexity in learning, close equivalence to clustering, generalization of the approximation to new data without heavy re-computations, and easy extension to a nonlinear kernel method with wide applications for optimization problems. We have later demonstrated that combining orthogonality and negativity works well in graph partitioning [3].

In NMF, the matrix difference was originally measured by the Frobenius matrix norm or the unnormalized Kullback-Leibler divergence (I-divergence). Recently we have significantly extended NMF to a much larger variety of divergences with theoretically convergent algorithms. In [7], we have presented a generic principle for deriving multiplicative update rules, as well as a proof of the convergence of their objective function, that applies for a large variety of linear and quadratic NMF problems. The proposed principle only requires that the NMF approximation objective function can be written as a sum of a finite number of monomials, which is a mild assumption that holds for many commonly used approximation error measures. As a result, our method turns the derivation, which seemingly requires intense mathematical work, into a routine exercise that could be even readily automated using symbolic mathematics software. In our practice [8], both theoretical and practical advantages indicate that there would be good reasons to replace the I-divergence with normalized Kullback-Leibler for NMF and its variants. The PNMF method can also be generalized to the α-divergence family [6].

Automatic determination of the low-rank in NMF is a difficult problem. In [9], we have presented a new algorithm which can automatically determine the rank of the projection matrix in PNMF. By using Jeffrey's prior as the model prior, we have made our algorithm free of human tuning in finding algorithm parameters. Figure 3.1 visualizes the learned basis of the *Swimmer* dataset.

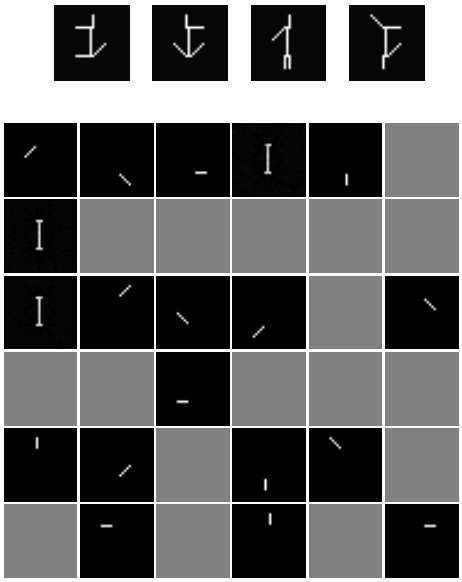


Figure 3.1: (Top) Some sample images of Swimmer dataset; (Bottom) 36 basis images of Swimmer dataset. The gray cells correspond to matrix columns whose L_2-norms are zero or very close to zero.

References

[1] I. Dhillon, Y. Guan, and B. Kulis. Kernel kmeans, spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556, Seattle, WA, USA, 2004.

[2] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[3] E. Oja and Z. Yang. Orthogonal nonnegative learning for sparse feature extraction and approximate combinatorial optimization. *Frontiers of Electrical and Electronic Engineering in China*, 5(3):261–273, 2010.

[4] P. Paatero P and U. Tapper. Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[5] Z. Yang and E. Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Transaction on Neural Networks*, 21(5):734–749, 2010.

[6] Z. Yang and E. Oja. Projective nonnegative matrix factorization based on α-divergence. *Journal of Artificial Intelligence and Soft Computing Research*, 1(1):7–16, 2011.

[7] Z. Yang and E. Oja. Unified development of multiplicative algorithms for linear and quadratic nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 22(12):1878–1891, 2011.

[8] Z. Yang, H. Zhang, Z. Yuan, and E. Oja. Kullback-leibler divergence for nonnegative matrix factorization. In *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN2011)*, pages 250–257, Espoo, Finland, 2011.

[9] Z. Yang, Z. Zhu, and E. Oja. Automatic rank determination in projective nonnegative matrix factorization. In *Proccedings of the 9th International Conference on Latent Variable Analysis and Signal Separation (LVA 2010)*, pages 514–521, St. Malo, France, 2010.

[10] S. X. Yu and J. Shi. Multiclass spectral clustering. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, volume 2, pages 313–319, 2003.

[11] Z. Yuan and E. Oja. Projective nonnegative matrix factorization for image compression and feature extraction. In *Proceedings of 14th Scandinavian Conference on Image Analysis (SCIA)*, pages 333–342, Joensuu, Finland, 2005.

3.3 Finding dependent and independent sources from two related data sets

Juha Karhunen, Tele Hao, and Jarkko Ylipaavalniemi

We have considered in two papers [3, 4] extension of independent component analysis (ICA) and blind source separation (BSS) for separating mutually dependent and independent components from two different but related data sets. This problem is important in practice, because such data sets are commonplace in real-world applications. We propose a new method which first uses canonical correlation analysis (CCA) [2] for detecting subspaces of independent and dependent components. The data sets are then mapped onto these subspaces. Even plain CCA can provide a coarse separation in simple cases, and we can justify this. Better separation results are obtained by applying some suitable ICA or BSS method [1] to the mapped data sets. These methods can utilize somewhat different properties of the data such as non-Gaussianity, temporal correlatedness, or nonstationary depending on the characteristics of the data.

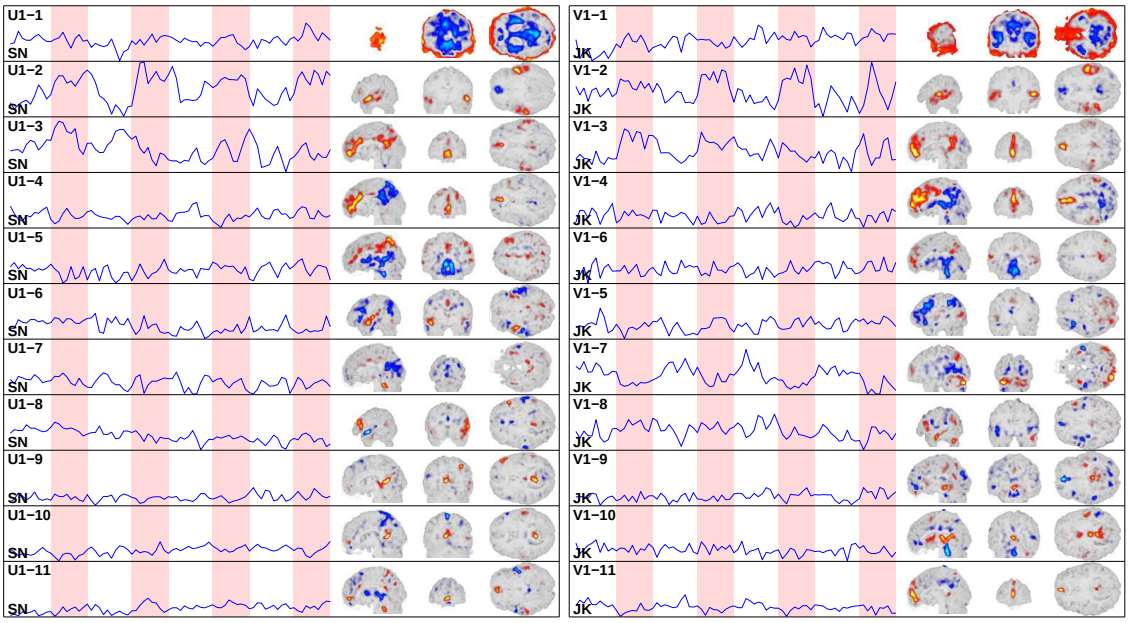
The proposed method is straightforward to implement and computationally not too demanding. CCA preprocessing improves often quite markedly the separation results of the chosen ICA or BSS method especially in difficult separation problems. Not only are the signal-to-noise ratios of the separated sources clearly higher, but CCA also helps a method to separate sources that it alone is not able to separate. In [3, 4], we present experimental results for several well-known ICA and BSS methods for synthetically constructed source signals [5] which are quite difficult to separate for most ICA and BSS methods. Furthermore, we have applied our method successfully to real-world robot grasping data in [3].

In [4], we tested the usefulness of our method with data taken from a functional magnetic resonance imaging (fMRI) study [6], where it is described in more detail. We used the measurements of two healthy adults while they were listening to spoken safety instructions in 30 s intervals, interleaved with 30 s resting periods. In these experiments we used slow feature analysis (SFA) [7] for post-processing the results given by CCA, because it gave better results than the most widely used standard ICA method FastICA [1].

Fig. 3.2 shows the results of applying our method to the two datasets and separating 11 components from the subspaces of dependent components. The consistency of the components across the subjects is quite good. The first component shows a global hemodynamic contrast, that may also be related to artifacts originating from smoothing the data in the standard preprocessing. The activity of the second component is focused on the primary auditory cortices. The third and fourth components show both positively and negatively task-related activity around the anterior and posterior cingulate gyrus. These first results are promising and in good agreement with the the ones reported in [6].

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Wiley, 2001.



(a) (b)

Figure 3.2: Experimental results with fMRI data. Each row shows one of the 11 separated components. The activation time-course with the stimulation blocks for reference, shown on the left, and the corresponding spatial pattern on three coincident slices, on the right. Components from (a) the first and (b) the second dataset.

[2] A. Rencher. *Methods of Multivariate Analysis, 2nd ed.*. Wiley, 2002.

[3] J. Karhunen and T. Hao. Finding dependent and independent components from two related data sets. In *Proc. of the IEEE Int. Joint Conf. on Neural Networks (IJCNN 2011)*, pp. 457–466, San Jose, California, USA, August 2011.

[4] J. Karhunen, T. Hao, and J. Ylipaavalniemi. A canonical correlation analysis based method for improving BSS of two related data sets. To appear in *Proc. of the 10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVAICA 2012)*, Tel-Aviv, Israel, March 2012.

[5] A. Hyvärinen. A unifying model for blind separation of independent sources. *Signal Processing*, 85(7):1419–1427, 2005.

[6] J. Ylipaavalniemi et al. Analyzing consistency of independent components: An fMRI illustration. *NeuroImage*, 39:169–180, 2008.

[7] L. Wiskott and T. Sejnowski. Slow feature analysis: unsupervised learning of invariances. *Neural Computation*, 14:715–770, 2002.

Chapter 4

Multi-source machine learning

Samuel Kaski, Mehmet Gönen, Arto Klami, Gayle Leen, Jaakko Peltonen, Ilkka Huopaniemi, Melih Kandemir, Suleiman A. Khan, Kristian Nybo, Juuso Parkkinen, Tommi Suvitaival, Jaakko Viinikanoja, Seppo Virtanen, Yusuf Yaslan

4.1 Introduction

We develop statistical machine learning methods for extracting useful regularities from large, high-dimensional data sets. In practical computational data analysis tasks a common problem is lack of sufficient amount of representative data. If there was enough data, modern statistical machine learning toolboxes would contain powerful approaches to building flexible models that do not make strong assumptions about data, but given little data we need to seek alternative ways to bring in more information. Our approach is combining various sources of information.

In many applications, for instance in molecular biology and neuroinformatics, there is data available in public or special-purpose databanks, but the problem is that not everything is relevant. We are developing new machine learning methods capable of learning from *multiple data sources* containing only *partially relevant* data, and generalizing to new contexts. The methods extend and generalize the current approaches called multi-view, multi-way and multi-task learning, on structured and unstructured domains.

Moreover, we have developed new principles and methods for the task of *visualizing* high-dimensional data; this task is central in any knowledge discovery process.

4.2 Multi-view and multi-way learning

Multi-view learning tells how several data sources, or views, can be combined to extract more relevant information. We build Bayesian latent variable models for the task of extracting statistical dependencies between multiple views of the same objects, for example to capture relationships between images and their captions, or between expressions of genes and chemical descriptors of drugs.

In the completely unsupervised case, we are given only the data matrices of co-occurring data, and the task is to mine for dependencies between them. For combining two views we have earlier introduced the **Bayesian Canonical Correlation Analysis** (CCA) model, which finds linear components capturing correlations between the views while modeling the variation specific to each view by separate noise components. To extend the range of potential applications, we have extended the Bayesian CCA model to mixtures of robust CCAs [1] and to generic exponential family noise models [2]. Recently, we introduced a considerably more efficient version of Bayesian CCA [3], which is applicable also to very large dimensionalities. Our novel solution builds on an efficient variational approximation, enabled by reformulating the problem as a group-wise sparse latent component model. Besides working with linear models, we have also presented a nonparametric Bayesian clustering model for similar setups [4].

The problem of analysing dependencies between more than two views is considerably mode difficult. Most solutions seek relationships between all views, whereas most practical applications will not satisfy that assumption. Recently we introduced the **Group Factor Analysis** problem, where the task is to find dependencies between all possible subsets of the views. By building on the group-wise sparsity assumption used for CCA we were able to derive a model that finds efficiently all types of dependencies present in the data collection, even though their potential number grows exponentially as a number of views [5]. The model is illustrated in Figure 4.1.

The task in the **matching** problem is to infer the co-occurrence of the samples from the data set itself. For example, given a collection of documents written in two languages, we might want to learn which documents correspond to each other. In [6] we show how such a match or alignment can be learned simultaneously while learning a model that maximizes the dependency between the two views, by an algorithm that alternates between learning the match and learning a subspace in which the samples can be compared with each other. We also demonstrated how multiple matching solutions can be combined to learn a consensus match over multiple data set instances, to learn a match between metabolites of two species.

The samples co-occurring in the multiple views can also be associated with covariates (labels); then the analysis problem becomes to discover how the different populations indicated by the labels differ from each other, akin to analysis of variance (ANOVA). The problem is particularly difficult in the "large p, small n" case ubiquitous in computational molecular biology, of having a high dimensionality p and a small sample size n. In [7] we introduced a solution combining both multi-view and multi-way learning, by building a Bayesian model that models the covariate effects in the latent space, assuming the views to be conditionally independent given the latent variable, similarly as in the above models. This kind of models and their applications in computational systems biology are discussed in detail in Chapter 5.

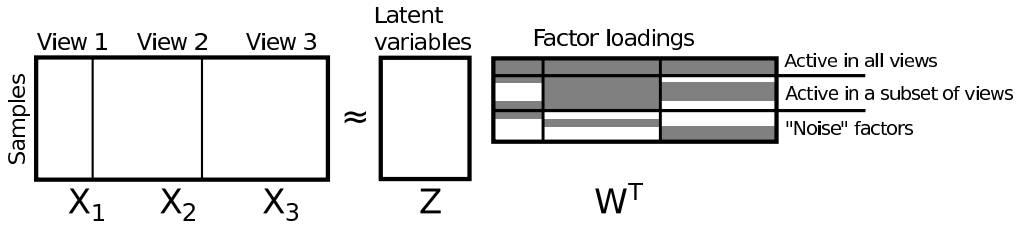


Figure 4.1: Illustration of the group factor analysis of three data sets or views. The feature-wise concatenation of the data sets $\mathbf{X}_i$ is factorized as a product of the latent variables $\mathbf{Z}$ and factor loadings $\mathbf{W}$. The factor loadings are group-wise sparse, so that each factor is active (gray shading, indicating $\mathbf{f}_{m,k} = 1$) only in some subset of views (or all of them). The factors active in just one of the views model the structured noise, variation independent of all other views, whereas the rest model the dependencies. The nature of each of the factors is learned automatically, without needing to specify the numbers of different factor types (whose number could be exponential in the number of views) beforehand.

References

[1] Jaakko Viinikanoja, Arto Klami, and Samuel Kaski. Variational Bayesian mixture of robust CCA models. In Aristides Gionis José Luis Balcázar, Francesco Bonchi and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases. Proceedings of European Conference, ECML PKDD 2010*, volume III, pages 370–385, Berlin, 2010. Springer.

[2] Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian exponential family projections for coupled data sources. In Peter Grunwald and Peter Spirtes, editors, *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (2010)*, pages 286–293, 2010. AUAI Press.

[3] Seppo Virtanen, Arto Klami, and Samuel Kaski. Bayesian CCA via group sparsity. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 457–464, New York, NY, 2011. ACM.

[4] Simon Rogers, Arto Klami, Janne Sinkkonen, Mark Girolami, and Samuel Kaski. Infinite factorization of multiple non-parametric views. *Machine Learning*, 79(1-2):201–226, 2010.

[5] Seppo Virtanen, Arto Klami, Suleiman A. Khan, and Samuel Kaski. Bayesian group factor analysis. In *Proceedings of AISTATS'12*, 2012. Preliminary version available as arXiv:1110.3204.

[6] Abhishek Tripathi, Arto Klami, Matej Orešič, and Samuel Kaski. Matching samples of multiple views. *Data Mining and Knowledge Discovery*, 23:300–321, 2011.

[7] Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, Matej Orešič, and Samuel Kaski. Multivariate multi-way analysis of multi-source data. *Bioinformatics*, 26:i391–i398, 2010. (ISMB 2010).

4.3 Multi-task learning

We have introduced two new multi-task learning setups, suitable for different scenarios, and solutions for them: *asymmetric multi-task learning* and *multi-task multiple kernel learning*.

4.3.1 Asymmetric multi-task learning

Multi-task learning is the setting where several collections of data samples are analyzed together; each collection represents a different learning task and comes from a different underlying distribution: for example, measurements of student performances in different schools, or scientific documents collected from different venues. The task is usually a supervised task, classification or regression, but may also be an unsupervised task such as clustering or density estimation. Often the data are high-dimensional and the number of data points in each individual task is too small for learning well the subtle distinctions necessary for good performance in the task.

Unlike in multi-view learning, in multi-task learning the individual samples in the data sources do not typically co-occur. Instead it is assumed that there are connections on the population level: if the underlying distributions in the tasks have similar properties (similar trends, groupings, manifolds, etc.) then learning the tasks together allows sharing the data between tasks, making possible learning of more complex models.

Typical multi-task learning solutions are based on treating all of the learning tasks symmetrically (with equal interest), for example by learning a hierarchical probabilistic model from all of the data collections where the models share parameters or priors of parameters; then all the data collections affect learning the shared parameters with an equal role. However, in many settings there is instead a task of interest (such as gene expression measurements of the current patient) where we wish to perform well and where test samples will come from, and other tasks are simply additional sources of information (such as historical records of earlier patients). In such settings the learning should be *asymmetric multi-task learning*: it should focus on learning the task of interest as well as possible, avoiding the danger of skewing the model of the task of interest in favor of modeling other tasks, which can happen in some symmetric approaches.

We have introduced a formalism for asymmetric multi-task learning, focusing on learning a classification task or regression task of interest with the help from auxiliary tasks that are related but are of less interest. On an intuitive level the idea is to *extract only the relevant information* of earlier data sets to help the learning of the task of interest. Technically, we use an intelligent mixture model, where each earlier task is explained partly by a *shared model* and partly by a task-specific *explaining-away model*. The task-of-interest, where everything is relevant, only uses the shared model, while other tasks are partly explained away by the explaining-away model.

Two kinds of methods were derived from this approach: a method for asymmetric multitask logistic regression [2], and a method for asymmetric multitask Gaussian process regression or classification [1]. In the logistic regression case, the model was formulated as

$$p_S(c|\mathbf{x}) = (1 - \pi_S)p^{shared}(c|\mathbf{x}) + \pi_S p_S^{explaining-away}(c|\mathbf{x})$$

where $\mathbf{x}$ are samples c are class labels, $p^{shared}(c|\mathbf{x})$ is a model shared between all tasks,

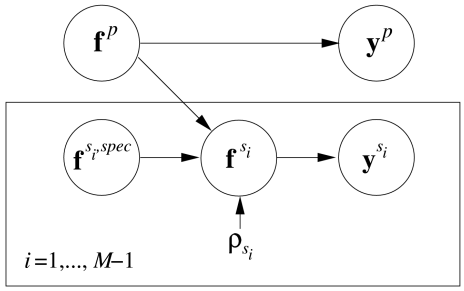


Figure 4.2: Graphical model of an asymmetric multi-task Gaussian process regression model, showing the relationship between the function values of the primary task (task of interest) and secondary tasks (other tasks).

$p_S^{explaining-away}(c|\mathbf{x})$ is a model to explain away non-relevant parts of task S and π_S is a mixture weight. In a Gaussian process regression context this can be similarly written as

$$y = f_S(\mathbf{x}) = f^{shared}(\mathbf{x}) + f_S^{explaining-away}(\mathbf{x})$$

where the y are regression targets, $f^{shared}(\mathbf{x})$ is a shared regression function and $f_S^{explaining-away}(\mathbf{x})$ is a function to explain away non-relevant regression variation of task S, and the functions are drawn independently from Gaussian process priors.

The methods were shown to outperform both naive approaches, such as single-task learning or pooling together all tasks, and also the nearest comparable symmetric multi-task learning approaches.

4.3.2 Multi-task multiple kernel learning

Empirical success of kernel-based learning algorithms is very much dependent on the kernel function used. Instead of using a single fixed kernel function, multiple kernel learning algorithms learn a combination of different kernel functions in order to obtain a similarity measure that better matches the underlying problem. We study multi-task learning problems and formulate a novel multi-task learning algorithm [3] that trains coupled but nonidentical multiple kernel learning models across the tasks. The proposed algorithm is especially useful for tasks that have different input and/or output space characteristics and is computationally very efficient. Empirical results on three data sets validate the generalization performance and the efficiency of our approach.

References

[1] Gayle Leen, Jaakko Peltonen, and Samuel Kaski. Focused Multi-task Learning Using Gaussian Processes. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases (proceedings of ECML PKDD 2011)*, Part II, pages 310–325, Berlin Heidelberg, 2011. Springer-Verlag. Winner of the ECML PKDD 2011 Best Paper Award in Machine Learning.

[2] Jaakko Peltonen, Yusuf Yaslan, and Samuel Kaski. Relevant subtask learning by constrained mixture models. *Intelligent Data Analysis*, 14:641–662, 2010.

[3] Mehmet Gönen, Melih Kandemir, and Samuel Kaski. Multitask Learning Using Regularized Multiple Kernel Learning. In Bao-Liang Lu, Liqing Zhang, and James Kwok, editors, *Proceedings of 18th International Conference on Neural Information Processing (ICONIP)*, Part II, pages 500–509, Berlin Heidelberg, 2011. Springer-Verlag.

4.4 Information visualization

Information visualization is an essential part of analysis of new data, especially in the first stages when strong hypotheses about the data have not been made yet. Many dimensionality reduction methods have been designed for tasks such as manifold learning and are not suitable for reducing the data much beyond the effective dimensionality of data. A visualization on a low-dimensional display cannot represent all aspects of high-dimensional data: it is then crucial to be able to quantify the errors that unavoidably occur in any visualization.

We have *formalized information visualization as a task of visual information retrieval* [1, 2], focusing on the specific task of retrieving similar items (retrieving neighborhood relationships) for a query item based on the visual display. In this task, all visualizations naturally end up with two kinds of errors, false neighbors and misses. The accuracy of such retrieval can be rigorously quantified using the information retrieval measures *precision* and *recall*. The analyst needs to specify a tradeoff between precision and recall (tradeoff between the costs of false neighbors and misses) to evaluate the goodness of visualizations. Moreover, generalizations of such measures can be directly set as an optimization goal, to produce visualizations that are optimal for information retrieval. We have also shown that optimizing visualizations for information retrieval can be done in the framework of generative modeling [3]. We have created nonlinear embeddings optimal for information retrieval, and have shown that they outperform existing visualization methods in the information retrieval tasks, and also by commonly used indirect measures.

We have applied the approach to visualization of graphs (graph layout) [4], fMRI data (Fig 4.3), and gene expression measurements (e.g., [5]).

References

[1] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.

[2] Samuel Kaski and Jaakko Peltonen. Dimensionality Reduction for Data Visualization. *IEEE Signal Processing Magazine*, 28(2):100–104, 2011.

[3] Jaakko Peltonen and Samuel Kaski. Generative Modeling for Maximizing Precision and Recall in Information Visualization. In Geoffrey Gordon, David Dunson, and Miroslav Dudik, editors, *Proceedings of AISTATS 2011, the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP, vol. 15*, 2011.

[4] Juuso Parkkinen, Kristian Nybo, Jaakko Peltonen, and Samuel Kaski. Graph Visualization With Latent Variable Models. In *Proceedings of MLG 2010, the Eighth Workshop on Mining and Learning with Graphs*, pages 94-101, New York, NY, USA, 2010. ACM.

[5] Jaakko Peltonen, Helena Aidos, Nils Gehlenborg, Alvis Brazma, and Samuel Kaski. An information retrieval perspective on visualization of gene expression data with ontological annotation. In *Proceedings of ICASSP 2010*, pages 2178-2181, 2010. IEEE.

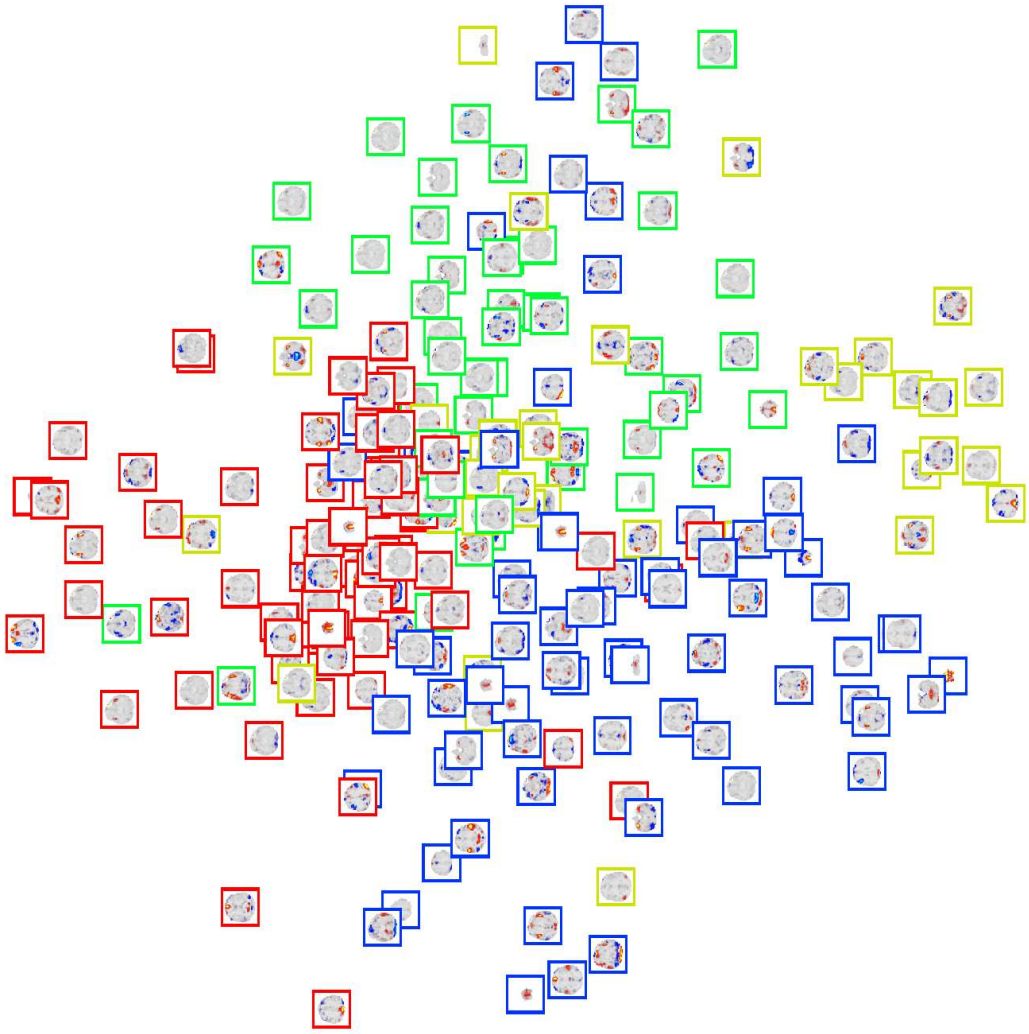


Figure 4.3: Visualization of fMRI whole-head volumes from an experiment with several people experiencing multiple stimuli. The visualization has been optimized for information retrieval of similar (neighbor) images from the visualization. The four stimuli types (red: tactile, yellow: auditory tone, green: auditory voice, blue: visual) have become separated in the visualization; the two auditory stimuli types are arranged close-by as is intuitively reasonable. An axial slice is shown for each whole-head volume, chosen so that the shown slice contains the highest-activity voxel.

Bioinformatics and Neuroinformatics

Chapter 5

Bioinformatics

Samuel Kaski, Elisabeth Georgii, Arto Klami, José Caldas, Ali Faisal, Ilkka Huopaniemi, Suleiman Ali Khan, Leo Lahti, Juuso Parkkinen, Tommi Suvitaival

5.1 Introduction

The accumulation of different types of high-throughput measurement data yields unprecedented opportunities to study specific biological questions in context of the big picture of genome biology. It remains a major challenge how to process, analyze, and exploit this wealth of data such that the findings generate useful biomedical hypotheses and advance our understanding of cellular processes. As the number of molecular players, such as genes and metabolites, is extremely large compared to the number of available measurements, deciphering their function and functional interactions is highly non-trivial. In addition, regulatory elements within the genomic sequence are known only to a small extent, and the number of potential candidate regions is enormous. Thus, a main concern of computational systems biology is to detect statistical relationships between data points as well as variables in high-dimensional and potentially heterogeneous data spaces.

Our research focuses on three major topics. The first theme is translational modeling in medical studies; the ultimate goal is to predict biological responses to treatment across different tissues as well as from model organisms to human. The second theme is data-driven comparison and retrieval of gene expression experiments; measurements from different laboratories and different biological conditions are brought together in a common modeling framework to discover similarities or dissimilarities of samples regarding their transcriptional characteristics. The third theme is data integration, modeling of heterogeneous data sets that provide multiple views on the same biological samples or entities, e.g., gene expression measurements, genome methylation profiles, and copy number changes. The task is to detect shared aspects as well as source-specific aspects by looking at dependencies between the views. The three topics are described in more detail in the following sections. In addition, we have developed probabilistic models for decomposing biological networks into functional modules [12], and for estimating probe reliability in microarray measurements [8].

We have worked in close collaboration with VTT (Prof. M. Orešič), Haartman Institute (Prof. S. Knuutila), European Bioinformatics Institute EBI (Prof. A. Brazma), Department of Biological and Environmental Sciences at University of Helsinki (Prof. J. Kangasjärvi), Institute for Molecular Medicine Finland FIMM (Prof. O. Kallioniemi), and Institute of Biomedicine (Dr. Sampsa Hautaniemi).

5.2 Translational modeling for molecular medicine

We develop probabilistic machine learning methods for translational tasks motivated by research questions in molecular medicine. We are addressing computational modeling problems, with the aim to ultimately assist in approaching the following tasks: (i) to predict the response to disease and its medical treatments in the complex biological system of a human being, based on experiments on model organisms and cell lines, and (ii) to decrease the need for invasive operations on human patients, by detecting dependencies between views that are hard and easy to observe (e.g., study of the state of an inner organ based on blood levels).

In our research, we have developed machine learning methods for estimating the effects of multiple experimental factors. These solutions take ANOVA-type modeling beyond the possibilities of the classical approaches. We have presented ways of detecting similar responses between multiple tissues of a patient, and between the patient and a model organism. We have utilized the novel methods in current metabolomic studies of human diseased and their medical treatments.

Disease-related dependencies between multiple tissues. We developed a data fusion method [5], which allows us to detect dependencies between multiple tissues of a biological organism that are related to known experimental factors, such the disease status and medical treatment (Figure 5.1b). Many diseases such as cancer may be located in a specific organ whose state is not directly observable without invasive operations. Our method provides a way of making predictions of the hard-to-measure tissue via more readily observable samples, such as from the blood.

Multi-way modeling made possible for heterogeneous clinical data sets. We have developed ANOVA-type modeling of responses to multiple experimental factors further for heterogeneous time series data [6] (Figure 5.1c). One of the major complications in the analysis of clinical studies of humans has been the heterogenity of individual histories in the medical records. By utilizing dynamical generative models, we separated progression into disease from normal aging-related development of individuals. For another clinical study [18], our ANOVA-type modeling approach for high-dimensional data was extended to the repeated measures setting, where the blood levels of each patient are observed both before and after the medical treatment.

Matching objects of multiple views. The major obtacle for translational studies is the lack of one-to-one mapping between the biological systems of the different organisms. Even the approximate mapping is often unknown. We have developed a novel matching algorithm for simultaneously (i) learning a metric to maximize the dependency between two data sets, and (ii) matching the objects between the data sets [16].

Disease-related responses across several species. The goals of translational cross-species modeling are to (i) find similarities between the responses measured in two domains (human, model organism), and (ii) predict the outcome of a new intervention in one of these domains based on a similar realized experiment in the other domain. An important application lies in pharmaceutics, where the effect of a new drug on the development of

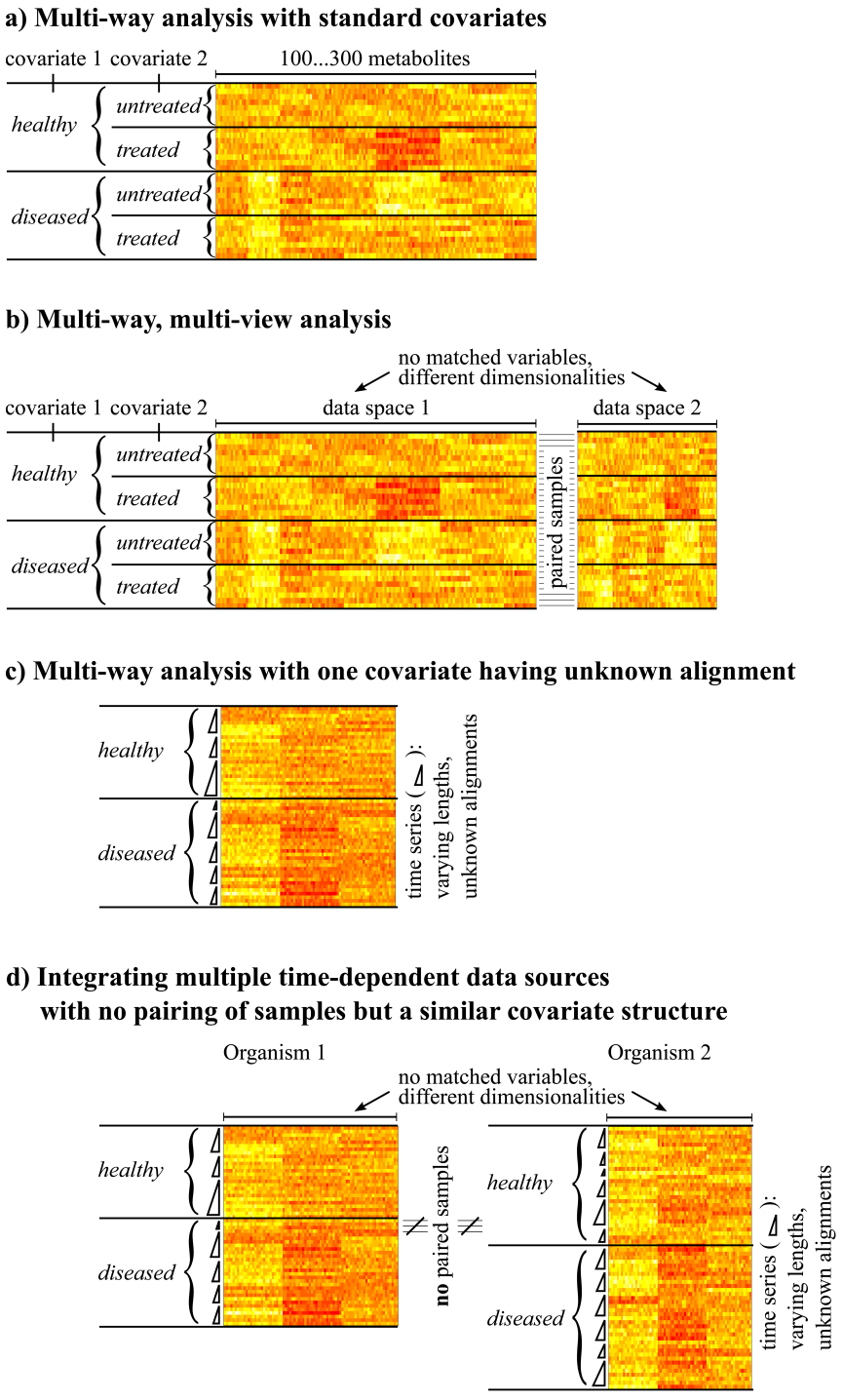


Figure 5.1: Illustration of the four data analysis tasks common in translational modeling for molecular medicine.

a human disease is studied with animals or tissues grown in laboratory, and later further tested with clinical studies on human patients. From the computational perspective, the translation of experiments is an unsolved problem, as neither variables nor samples are matched between the domains (Figure 5.1d).

We have introduced a model for matching groups of variables between the two domains based on their similarity in responses to relevant experimental factors [6]. Further, we separated domain-specific responses from the responses shared by the domains [13, 14].

Model organism study for type 1 diabetes. Metabolic development in children progressing into Type 1 Diabetes (T1D) is not well understood. As members of an interdisciplinary consortium, we have studied the development of T1D through a model organism [15]. By comparing the results with a similar follow-up study of human children, we found out that before the onset of the disease, female NOD mice exhibit the same lipidomic pattern as pre-diabetic human children. The results suggest alternative metabolic-related pathways as therapeutic targets to prevent the disease. These biological findings were made possible by methods for learning a data-driven model that maps human and mouse lipidomes.

We have proposed translational modeling methods for detecting dependencies between heterogeneous data sets as well as estimating and predicting effects of relevant experimental factors across domains. The methods have been designed to work with high-dimensional biological real-world data.

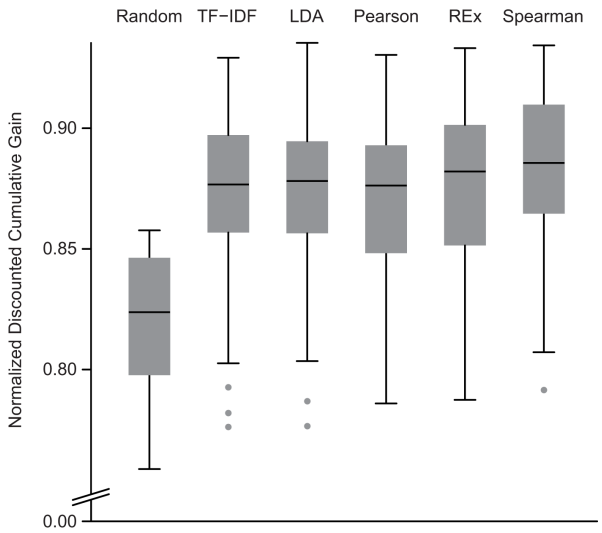


Figure 5.2: Figure taken from Caldas *et al.* [1]. Data-driven retrieval performance based on Normalized Discounted Cumulative Gain (NDCG); a measure of effectiveness of a search engine. The box plot summarize the distribution NDCG results for 219 interpretable query comparisons. "LDA" corresponds to our earlier method while REx corresponds to [1].

5.3 Data-driven comparison and retrieval of gene expression experiments

Considerable effort has been spent on collecting gene expression measurements into huge public repositories. This has opened up the door to large-scale comparisons and meta-analysis of data from different experiments. We have developed probabilistic methods that assist in these analysis tasks. In addition, we have introduced the concept of model-based retrieval of relevant biological experiments.

Content-based retrieval of relevant experiments. In previous work, we developed the first prototype of a content-based retrieval engine for biological experiments (REx: Retrieval of Relevant Experiments). To complement keyword search functionalities provided by most repositories for retrieval of similarly annotated studies, we developed probabilistic machine learning methods that relate gene expression studies through their actual measurement data, along with visualization tools that allow exploring and interpreting the results. The "model of biology" underlying our retrieval method is both data- and knowledge-driven: we use enrichment analysis for known functional gene sets (pathways) to obtain a representation of expression data that is comparable across measurement platforms. In [1], we extended the REx work to handle arbitrary experimental designs and to use a more accurate approach for modeling the activity of gene sets. In addition, the new REx model takes into account correlations in the activity of gene expression patterns. We also proposed a novel performance evaluation approach that is based on the Experimental Factor Ontology (EFO) of the ArrayExpress database and thereby much more scalable than manual relevance classification.

In a thorough comparison with alternative methods, REx performs competitively (see Figure 5.2). The advantage of our method lies in the interpretability of search results in terms of differential expression patterns. A previously unknown connection between

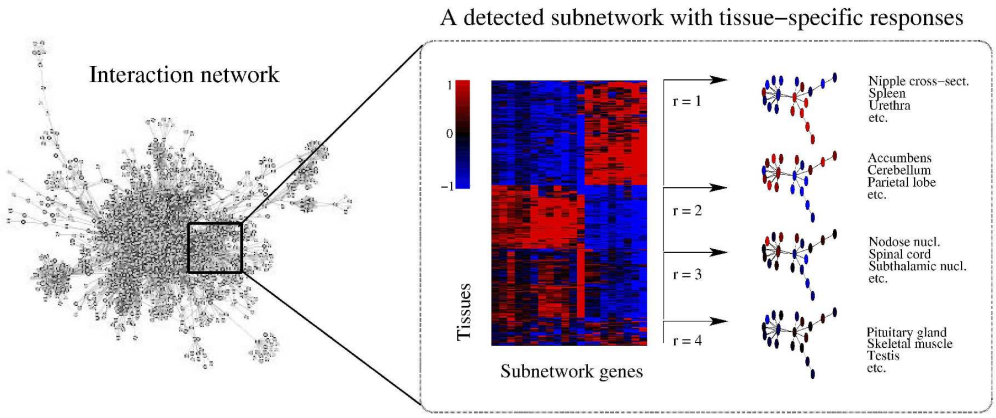


Figure 5.3: Figure taken from [10]. Organism-wide analysis of transcriptional responses in a human pathway interaction network reveals physiologically coherent activation patterns and tissue-specific regulation. One of the subnetworks and its tissue-specific responses, as detected by the NetResponse algorithm is shown. The expression of each gene is visualized with respect to its mean level of expression across all samples.

differential expression of the *SIM2s* gene and malignant pleural mesothelioma (MPM) suggested by our method in one of the case studies was experimentally verified in a new set of mesothelioma samples. Our work shows that the relatively unexplored paradigm of data-driven information retrieval in transcriptomics data offers the possibility of obtaining novel biological findings based on existing data, and holds the potential to ultimately accelerate biomedical research. A further extension of REx based on targeted regulatory models of gene expression has been submitted for publication.

Network-guided transcriptional response patterns. Different biological conditions (and tissues) can share the same cellular processes, which can be characterized by coordinated up- and down-regulation patterns in a specific set of genes, building so-called transcriptional signatures. Pathways and functional gene sets stored in public databases are typically not provided with information on the biological context of activation and generally too broad to define condition-dependent transcriptional signatures. We have developed an algorithm to detect gene sets that partition the biological conditions into groups where each group is characterized by a coherent activation pattern, modeled by a specific underlying signature (see Figure 5.3) [10]. The patterns are learned directly from gene expression data; to guide the analysis towards biologically interpretable signatures, the method exploits a network of known gene interactions to incrementally build larger candidate gene sets.

Hierarchical biclustering. Biclustering is the computational task of simultaneously clustering objects and inferring which features of the objects contribute to the grouping. Biclustering approaches are very popular in gene expression analysis, assisting in simultaneously uncovering relationships among biological samples and among genes. Our approach [2, 3] has two main contributions to the biclustering world: First, it applies the Bayesian framework to rigorously account for noise and uncertainty. Second, it learns a hierarchical tree structure for the samples, assigning characteristic genes to the nodes in the hierarchy. The model additionally yields natural information retrieval relevance

measures that can be used for relating samples to a query, making it eligible for the REx applications described above. The method outperformed four state-of-the-art biclustering procedures on a large miRNA data set.

5.4 Detection of dependencies between heterogeneous biological data types

Living cells are extremely complex systems, and hence integration of information from multiple sources is needed for accurate identification of underlying biological processes. We consider the data fusion problem of combining two or more data sources where each source consists of vector-valued measurements from the same object or sample, but on different variables. The task is to detect aspects that are shared between different sources as well as source-specific components.

Bayesian Group Factor Analysis for understanding drug action mechanisms. We decomposed the dependencies between drug structures and their biological responses in multiple diseases, using a novel method called *Group Factor Analysis (GFA)* [17]. Unlike standard QSAR methods, which relate drug properties and univariate responses, we find relationships between a set of structural descriptors of drugs and their genome-wide responses. GFA is a novel extension to factor analysis that models dependencies between sets of variables ("views") instead of variables, representing them as group-wise sparse factors (see Figure 5.4). Unlike existing methods, GFA finds sparse factors shared by subsets of views (most interesting) along with those shared by all and those specific to one view.

In [7], we present details of the decomposed drug response relationships. With GFA, we are able to find factors that capture variation between chemical descriptors and biological responses in one, two, or all three diseases. The factors form hypotheses about drug response patterns, allowing us to relate specific chemical descriptors with targeted cellular responses. We find four main types of factors: (i) Factors shared by the chemical view and a subset (one or two) of the cell lines. These factors give hypotheses for drug responses specific to cancer type and are hence the most interesting ones. (ii) Factors shared by all cell lines and the chemical space, representing drug effects common to all three subtypes of cancer. (iii) Factors shared by all cell lines but not the chemical space. They are either drug effects not captured by the specific chemical descriptors used, or common biological response to the modulation of two or more different targets which can not be captured by any common chemical description. (iv) Factors specific to one view represent "biological noise". Our analysis shows that the discovered factors not only capture meaningful biological dependencies but are also more predictive of protein targets than similarly but individually analyzed chemical and biological response spaces.

Survival-associated biomarkers from multi-view functional genomics data. Genomic instability is a hallmark of cancer and high-throughput measurements of copy-number variation data have become commonplace in cancers. Given that copy-number alterations are noisy, one of the most successful approaches in increasing the reliability of putative driver genes involved in tumor progression and drug resistance is integration of copy number data with transcriptomics data. In [11], we demonstrate the benefits of using a systematic computational framework to include algorithms that enable indentification of context and clinically important patient groups. The results provide genes and genomic regions that have survival effect in Glioblastoma or a clinically defined subset, such as temozolomide-treated patients, and thus facilitate translation of large-scale biomedical data to knowledge.

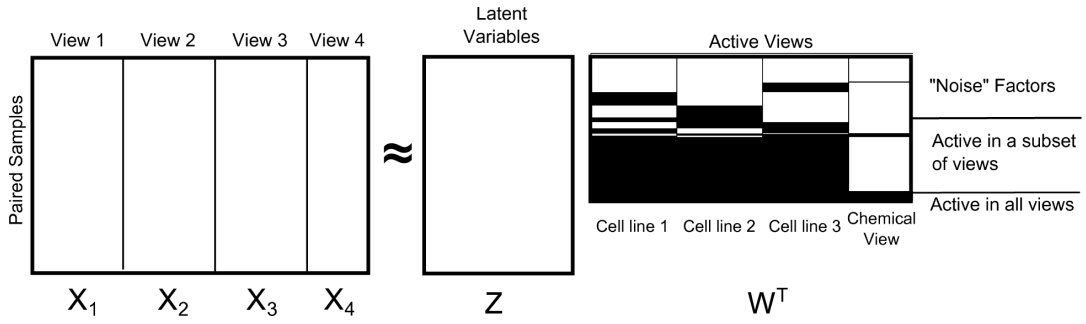


Figure 5.4: Illustration of the group factor analysis on three cell lines (diseases) and chemical view. The feature-wise concatenation of the data sets X_i is factorized as a product of the latent variables Z and factor loadings W. The factor loadings are group-wise sparse, so that each factor is active (gray shading) only in some subset of views (or all of them). The factors active in just one of the views model the structured noise, that is, variation independent of all other views, whereas the rest model the dependencies. The W shows the activity of the GFA factors in the 3 diseases (cell line 1: HL60, 2: MCF7, 3: PC3) and the drug descriptors (chemical view).

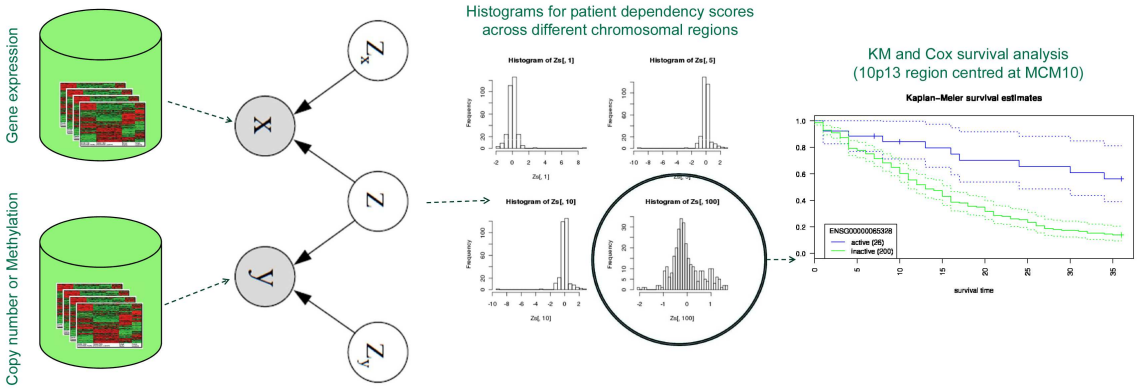


Figure 5.5: The analysis pipeline: A. Plate diagram for the canonical correlation analysis that captures the shared patterns Z from two data sources X and Z. B. Histograms of patients' contribution for four different genomic regions that have significantly high dependence scores; these histograms are used to form patient groups based on quantile clustering of the histogram. C. Sample Kaplan-Meier survival curve comparing the two patient groups for the genomic region centered at MCM10 gene; the patients with high dependence score have better survival than patients with low dependence scores, X-axis: months, Y-axis: percentage of GBM patients alive, dotted lines: 95% confidence intervals.

In [4], we present details of the approach used to identify potential genomic regions (or biomarkers) that effectively stratify patients in low and high survival groups. We first identify chromosomal regions that have high dependency between gene expression, methylation, and copy number changes, and then form patients groups from the regions and check whether the identified genomic aberrations have survival association (see Figure 5.5). The integration model is based on our earlier method for constrained canonical correlation analysis [9]. In addition, we incorporate suitable priors that model the positive correlation between gene expression and copy number data and the negative correlation between gene expression and methylation data. Furthermore, we incorporate sample-specific covariates in advanced survival analysis techniques. Results on Glioblastoma multiforme (GBM)

patient measurements identify known and novel genomic regions that may contribute to GBM progression and drug resistance.

References

[1] J. Caldas, N. Gehlenborg, E. Kettunen, A. Faisal, M. Ronty, A.G. Nicholson, S. Knuutila, A. Brazma and S. Kaski. Data-Driven Information Retrieval in Heterogeneous Collections of Transcriptomics Data Links SIM2s to Malignant Pleural Mesothelioma. *Bioinformatics*, 28(2):i246–i253, 2012.

[2] J. Caldas and S. Kaski. Hierarchical generative biclustering for microRNA expression analysis. *Journal of Computational Biology*, 18:251–261, 2011.

[3] J. Caldas and S. Kaski. Hierarchical generative biclustering for microRNA expression analysis. In *Research in Computational Molecular Biology, Proceedings of 14th Annual International Conference RECOMB 2010*, pages 65–79. Springer, 2010.

[4] A. Faisal, R. Louhimo, L. Lahti, S. Hautaniemi and S. Kaski. Biomarker discovery via dependency analysis of multi-view functional genomics data. In *NIPS 2011 workshop "From Statistical Genetics to Predictive Models in Personalized Medicine"*, 2011. Extendend abstract.

[5] I. Huopaniemi, T. Suvitaival, Janne Nikkilä, M. Orešič, and S. Kaski. Multivariate multi-way analysis of multi-source data. *Bioinformatics*, 26:i391–i398, 2010. (ISMB 2010)

[6] I. Huopaniemi, T. Suvitaival, M. Orešič, and S. Kaski. Graphical multi-way models. In J.L. Balcázar, F. Bonchi, A. Gionis and M. Sebag, editors, *Machine Learning and Knowledge Discovery in Databases. Proceedings of the European Conference on Machine Learning, ECML PKDD 2010*, volume I, pages 538–553. Springer, 2010.

[7] S.A. Khan, S. Virtanen, A. Klami, K. Wennerberg and S. Kaski. Decomposing Drug Response Patterns using Bayesian Group Factor Analysis. In *NIPS 2011 workshop "Machine Learning in Computational Biology"*, 2011. Abstract.

[8] L. Lahti, L. Elo, T. Aittokallio, and S. Kaski. Probabilistic analysis of probe reliability in differential gene expression studies with short oligonucleotide arrays. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8:217–225, 2011.

[9] L. Lahti and O.-P. Huovilainen. Dependency modeling toolkit. In *MLOSS workshop at ICML 2010*, 2010. Computer program.

[10] L. Lahti, J. Knuuttila, and S. Kaski. Global modeling of transcriptional responses in interaction networks. *Bioinformatics*, 26:2713–2720, 2010.

[11] R. Louhimo, V. Aittomäki, A. Faisal, M. Laakso, P. Chen, K. Ovaska, E. Valo, L. Lahti, V. Rogojin, S. Kaski, and S. Hautaniemi. Systematic use of computational methods allows stratifying treatment responders in glioblastoma multiforme. In *Proceedings of CAMDA 2011 conference, "Critical Assessment of Massive Data Analysis"*, 2011.

[12] J. Parkkinen and S. Kaski. Searching for functional gene modules with interaction component models. *BMC Systems Biology*, 4:4, 2010.

[13] T. Suvitaival, I. Huopaniemi, M. Orešič, and S. Kaski. Cross-species translation of multi-way biomarkers. In T. Honkela, W. Duch, M. Girolami, and S. Kaski, editors, *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN), Part I*, volume 6791 of *Lecture Notes in Computer Science*, pages 209–216. Springer, 2011.

[14] T. Suvitaival, I. Huopaniemi, M. Orešič, and S. Kaski. Detecting similar high-dimensional responses to experimental factors between human and model organism. In *NIPS 2011 workshop "From Statistical Genetics to Predictive Models in Personalized Medicine"*, 2011. Extendend abstract.

[15] M. Sysi-Aho, A. Ermolov, P.V. Gopalacharyulu, A. Tripathi, T. Seppänen-Laakso, J. Maukonen, I. Mattila, S.T. Ruohonen, L. Vähätalo, L. Yetukuri, T. Härkönen, E. Lindfors, J. Nikkilä, J. Ilonen, O. Simell, M. Saarela, M. Knip, S. Kaski, E. Savontaus, and M. Orešič. Metabolic regulation in progression to autoimmune diabetes. *PLoS Computational Biology*, 7:e1002257, 2011.

[16] A. Tripathi, A. Klami, M. Orešič, and S. Kaski. Matching samples of multiple views. *Data Mining and Knowledge Discovery*, 23:300–321, 2011.

[17] S. Virtanen, A. Klami, S.A. Khan, and S. Kaski. Bayesian Group Factor Analysis. To appear in *Fifteenth International Conference on Artificial Intelligence and Statistics AISTATS* 2012, *pre-print at: arXiv:1110.3204 [stat.ML]*

[18] L. Yetukuri, I. Huopaniemi, A. Koivuniemi, M. Maranghi, A. Hiukka, H. Nygren, S. Kaski, M.-R. Taskinen, I. Vattulainen, M. Jauhiainen, and M. Orešič. High Density Lipoprotein structural changes and drug response in lipidomic profiles following the long-term fenofibrate therapy in the FIELD substudy. *PLoS ONE*, (8):e23589, 2011.

Chapter 6

Neuroinformatics

Ricardo Vigário, Miguel Almeida, Nicolau Gonçalves, Nima Reyhani, Jarkko Ylipaavalniemi, Jayaprakash Rajasekharan, Jaakko Viinikanoja, Seppo Virtanen, Arto Klami, Mikko Kurimo, Samuel Kaski, Erkki Oja

6.1 Introduction

Neuroinformatics has been defined as *the combination of neuroscience and information sciences to develop and apply advanced tools and approaches essential for a major advancement in understanding the structure and function of the brain.* Aside from the development of new tools, the fields of application include often the analysis and modelling of neuronal behaviour, as well as the efficient handling and mining of scientific databases. The group aims at proposing algorithmic and methodological solutions for the analysis of elements and networks of functional brain activity, addressing several forms of communication mechanisms. Motivation and application areas include the understanding of ongoing brain activity and the neuronal responses to complex natural stimulation.

From a methodological viewpoint, the neuroinformatics group has studied properties of source separation methods, such as their reliability and extensions to subspaces. We have also assessed the suitability of such methods to the analysis of electrophysiological recordings (EEG and MEG), and functional magnetic resonance images (fMRI). We proposed also methods for the study of phase synchrony within the central nervous system, and between this and the peripheral nervous system. We have also developed methods for the analysis of neural responses of natural stimulation, based on a novel approach of capturing statistical dependencies between brain activity and the stimulus itself.

In addition to the analysis of fMRI recordings from natural stimulation, we have been also involved in the analysis of single trial event-related MEG data. Albeit its significantly higher temporal resolution, the signal-to-noise ratios are typically very poor, and averaging across hundreds of stimuli is often required. We currently search as well for efficient tissue segmentation of structural MRI.

In addition to these ongoing but stable research topics, we have made a pilot study in document mining. The goal is to extract, in a semi-automatic manner, functional information from neuroscience journals, hence reducing the dependence on curator intervention. We have also continued our research on approaches for segmentation of tissues in multispectral magnetic resonance images, with particular interest to the automatic detection and delineation of degenerative pathologies.

Research reported in this section has been carried out in collaboration with experts in neuroscience and cardiology. In the following, we highlight some of the results attained in the reported years.

References

[1] Almeida, M., J. Bioucas-Dias, and R. Vigário. Independent Phase Analysis: Separating Phase-Locked Subspaces. In *Proc. 9th Int. Conf. on Latent Variable Analysis and Source Separation (LVA/ICA'2010)*, St. Malo, France, 2010.

[2] Almeida M., J.-H. Schleimer, J. Bioucas-Dias, and R. Vigário. Source Separation of Phase-Locked Signals. *IEEE Trans. Neural Networks 22*, 1419–1434, 2011.

[3] Almeida M., R. Vigário, and J. Bioucas-Dias. Phase Locked Matrix Factorization. In *Proc. 19th European Signal Processing Conference (EUSIPCO'2011)*, Barcelona, Spain, 2011.

[4] A. Klami, P. Ramkumar, S. Virtanen, L. Parkkonen, R. Hari, and S. Kaski. ICANN/PASCAL2 challenge: MEG mind reading – overview and results. In *Proceedings of ICANN/PASCAL2 Challenge: MEG Mind Reading*, Aalto University Publication series SCIENCE + TECHNOLOGY 29/2011, pages 3–19, Espoo, 2011.

[5] M. Koskinen, J. Viinikanoja, M. Kurimo, A. Klami, S. Kaski, and R. Hari. Identifying fragments of natural speech from the listener's MEG signals. *Human Brain Mapping*, DOI:10.1002/hbm.22004, 2012.

[6] Rajasekharan J., U. Scharfenberger, N. Gonçalves, and R. Vigário. Image approach towards document mining in neuroscientific publications. In *Proc. 9th Int. Symposium in Advances in Intelligent Data Analysis (IDA'2010)*, Tucson, AZ, USA, 2010.

[7] Reyhani N., R. Vigário, J. Ylipaavalniemi, and E. Oja. A consistency and asymptotic normality of fastICA and bootstrap fastICA. *Signal Processing*, DOI:10.1016/j.sigpro.2011.11.025, 2012.

[8] Deville, Y., C. Jutten, and R. Vigário. Overview of source separation applications. In P. Comon and C. Jutten (Eds.), *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, pp. 639–681. Academic Press, 2010.

[9] Virtanen S., A. Klami, and S. Kaski. Bayesian CCA via group sparsity. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 457–464, New York, NY, 2011. ACM.

6.2 Natural stimuli and decoding

Natural stimuli are increasingly being used in neuroscientific experiments in order to study more complex brain activity, such as the brain's response to viewing movies, listening to free-flowing speech, or even being engaged in a discussion. New computational methods are needed for analysis of such experiments, since it is no longer feasible to assume single features of the experimental design to alone account for the brain activity. Instead, the stimulus itself becomes another source of data with rich features.

We have developed novel machine learning models (see Chapter 4 for more details) for analysis of MEG and fMRI response to natural stimuli. The models find latent representations that describe functional patterns in the brain data that correlate with rich feature representations of the stimuli. Besides providing interpretable components, the models are useful for decoding brain activity [9]. Given a brain measurement (for example, one TR in fMRI or a small time-window in MEG) the goal is to tell what kind of a stimulus the subject was exposed to. As a practical example, in [5] we used Bayesian Canonical Correlation Analysis (CCA) to model the MEG response to natural speech, and managed to identify which short segment of speech the subject was hearing with high accuracy even for segments lasting only for a few seconds (Fig. 6.1).

We also organized a PASCAL2 challenge on MEG decoding [4], to demonstrate the feasibility of single-trial MEG decoding and to provide public benchmark data. The task was to identify which of five types of video the subject was seeing based on just two seconds of data, and the best teams reached almost 70% accuracy (chance level 23%).

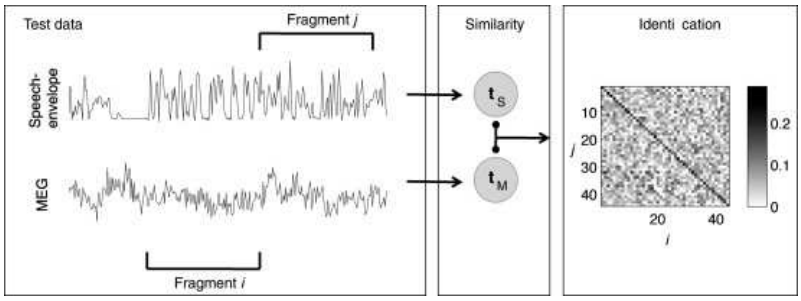


Figure 6.1: Bayesian CCA decodes speech by representing the speech fragments (left) in a joint latent space which enables directly comparing the MEG samples with speech envelopes (middle). The image is reproduced from [5] with permission.

6.3 Phase synchrony

Interest in phase synchronization phenomena has a long history, when studying the interaction of complex, natural or artificial, dynamic systems. Although not completely adopted, synchronization was attributed a role in the interplay between different parts of the central nervous system as well as across central and peripheral nervous systems. Such phenomena can be quantified by the phase locking factor, which requires knowledge of the instantaneous phase of an observed signal. Yet, observations are often mixtures of underlying phenomena, which destroys sources' phases.

Algorithms for Synchrony Source Separation
During the reported years, we extended the set of algorithmic tools for the identification of phase synchronous phenomena. And studied these tools in terms of deviations from the ideal modelled situations, when synchrony is affected by significant amounts of noise. Our earlier methods dealt with the extraction of sources phase-locked to a reference signal, the clustering of a population of oscillators into synchronous sub-populations, as well as the extraction of phase-locked subspaces, following an approach akin to the underlying considerations in independent component analysis. A summary of the said methods appeared in [2].

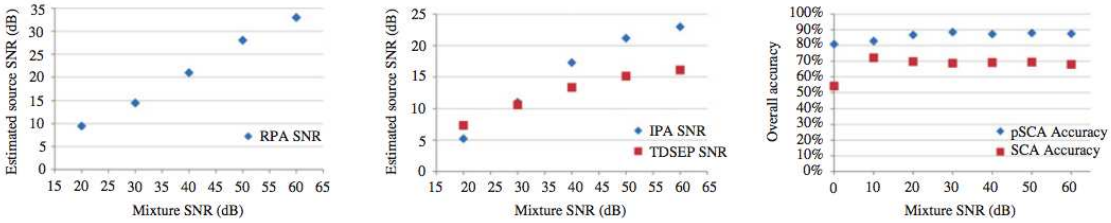


Figure 6.2: Study of noise robustness for three proposed phase-based algorithms: RPA (left), IPA and TDSEP (middle), and pSCA and SCA (right). From [2].

In [3], a new and very fast algorithm was proposed, based on a matrix factorisation approach. The separation is done through a minimisation problem involving three variables: the mixing matrix, the source time-dependent amplitudes, and their relative phases.

Phase-locked subspaces
We have further started to study the problem of blind separation of sources, when these are organized in subspaces. In this structure, sources in different subspaces have zero phase synchrony with each other, whereas sources in the same subspace exhibit full phase synchrony. Note that traditional source separation methods should fail in such generative model. The two-stage algorithm proposed in [1] performed remarkably well when in low-noise situations.

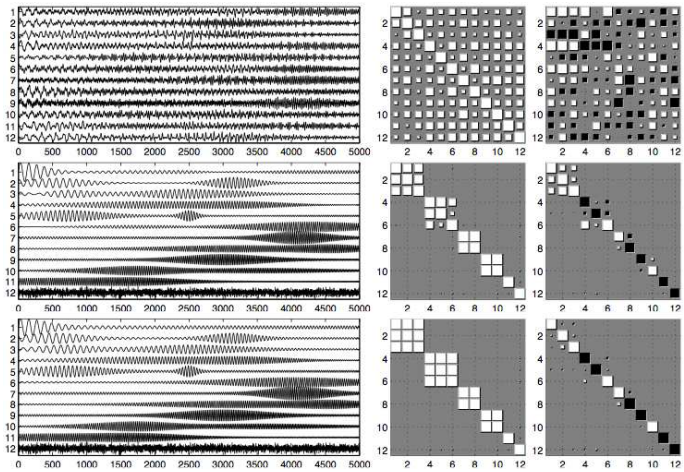


Figure 6.3: Measured (mixed) signals (first row, left); phase locking factors between those and the mixing matrix (middle); and the mixing matrix (right). (Second row) Sources resulting from TDSEP (left). Note that the inter.subspace PLFs (middle) are very close to zero, but the intra-subspace PLFs are not all close to 1. (Fourth row) Results found after the second stage of the algorithm. The estimated sources (left) are very similar to the original ones. This is corroborated by the PLFs between the estimated sources (middle) and the final unmixing matrix (right). From [1].

6.4 Document mining

There is an ever increase in the number of scientific publications in many areas in general, and in neurosciences in particular. Hundreds of articles are published each month. When comparing the results one obtains with a given experimental setup and existing information in literature, one may validate, integrate or confront different opinions and theories. The compilation of such a vast amount of information is not only crucial, but currently also rather human-intensive.

With that in mind, we have conducted a pilot study on document mining of journal publications reporting results on fMRI experiments. We have focused on the image content of the articles. The rather positive preliminary results reported in [6] suggest that a more systematic use of the methodology, and its improvement may help as well reducing the amount of curating work required for the construction of functional databases. We have been extending this research to hundreds of journal articles, and thousands of images, focusing on particular neurological conditions. The firs of our such results should appear soon.

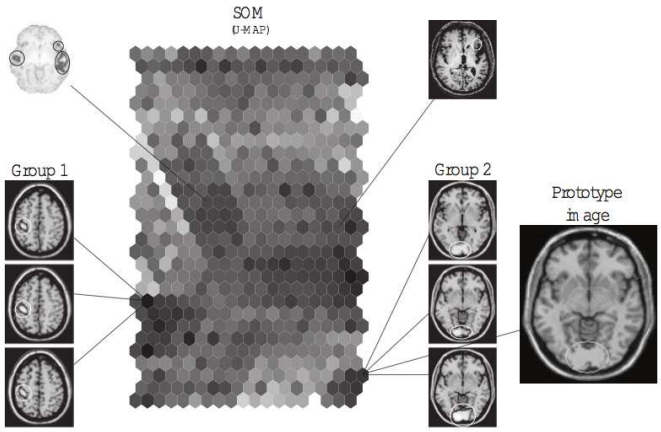


Figure 6.4: Self Organizing Map – U-matrix trained with 16 dimensional feature vectors, from a set of 100 images extracted from 11 journal papers. Two distinct cluster regions are observed at the lower left and right sides of the map. The prototype image, depicted in the upper left corner fits the expected cluster.

Multimodal interfaces

Chapter 7

Content-based information retrieval and analysis

Erkki Oja, Jorma Laaksonen, Markus Koskela, Ville Viitaniemi, Mats Sjöberg, Xi Chen, Satoru Ishikawa, Matti Karppa, Mikko Kurimo, Ville Turunen

7.1 Introduction

The Content-Based Information Retrieval Research Group studies and develops efficient methods for content-based and multimodal information retrieval and analysis tasks and implements them in the PicSOM[1] content-based information retrieval (CBIR) system. In the PicSOM CBIR system, parallel Self-Organizing Maps (SOMs) and Support Vector Machine (SVM) classifiers have been trained with separate data sets obtained from the multimodal object data with different feature extraction techniques. The different classifiers and their underlying feature extraction schemes impose different similarity measures and categorizations on the images, videos, texts and other media objects.

7.2 Semantic concept detection from images and videos

Extracting semantic concepts from multimedia data has been studied intensively in recent years. The aim of the research on the multimedia retrieval research community has been to facilitate semantic indexing and concept-based retrieval of unannotated multimedia content. The modeling of mid-level semantic concepts is often essential in supporting high-level indexing and querying on multimedia data as such concept models can be trained off-line with considerably more positive and negative examples than what are available at interactive query time.

In the course of previous years we have outlined and implemented our generic PicSOM system architecture for multimedia retrieval tasks. Detection of concepts from multimedia data—e.g. images and video shots—forms an important part of the architecture and we have formulated it as a standard supervised machine learning problem. Our concept detection technology is fundamentally based on fusion of a large number of elementary detections, each based on a different low-level audiovisual feature extracted from the multimedia data [1, 2].

During the period 2010–2011 we have continued our work in improving the bag of visual words (BoV) techniques for concept detection [3] and our participation in the annual TRECVID video analysis evaluations[2]. We have also applied our general-purpose algorithm for visual category recognition to the recognition of indoor locations [4]. Indoor localization is an important application in many emerging fields, such as mobile augmented reality and autonomous robots. A number of different approaches have been proposed, but arguably the prevailing method is to combine camera-based visual information to some additional input modalities, such as laser range sensors, depth cameras, sonar, stereo vision, temporal continuity, odometry, and the floorplan of the environment. We evaluated our method with other location recognition systems in the ImageCLEF 2010 RobotVision contest.

As a joint work together with the Speech Recognition Research Group, we have participated in the *Next Media* TIVIT ICT SHOK since 2010. We have applied our content-based video analysis and continuous speech recognition systems for the analysis of television broadcast material provided by the Finnish Broadcasting Company YLE. Figure 7.1 illustrates the results of the analysis for one regional news broadcast. On the left we can see how the temporal structure of the program has been revealed based on the clustering of

[1] `http://www.cis.hut.fi/picsom`
[2] `http://trecvid.nist.gov/`

visible human faces. In the right subfigure, the detected visual concepts are show on the top, the continuous speech recognition output on the bottom and the recognized name of the person on the right.

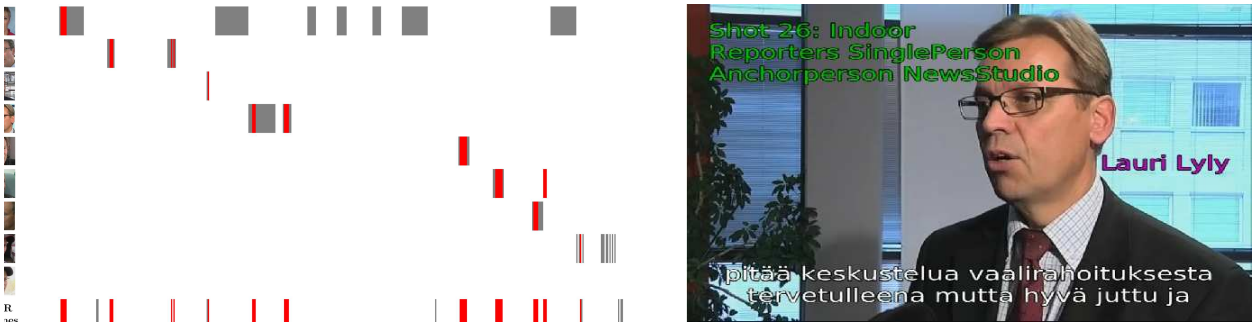


Figure 7.1: Temporal analysis of a news program based on clustering of facial images and the resulting content annotations.

7.3 Content-based video analysis and annotation of Finnish Sign Language

In January 2011 a new project *Content-based video analysis and annotation of Finnish Sign Language (CoBaSiL)*, funded by the Academy of Finland for four years, was started. In our joint work with the University of Jyväskylä and the Finnish Association of the Deaf, we are applying our methods of video content processing for the analysis, indexing and recognition of recorded Finnish Sign Language. In the project we study the use of computer vision techniques to recognize and analyze first the body parts of the signer and then his or her hand locations, shapes and gestures and facial expressions. Figure 7.2 illustrates the results of the stages of face detection, skin-color recognition and shape modelling with active shape models in the processing chain [5].

The linguistic goal of the project is to identify the sign and gesture boundaries and to indicate which video sequences correspond to specific signs and gestures [6]. This will facilitate indexing and construction of an example-based open-access visual corpus of the Finnish Sign Language for which there already exists large amounts of non-indexed video material. Currently we have concentrated our effort on studying the partially annotated material of the publicly available on-line dictionary of Finnish Sign Language, Suvi[3].

7.4 Image based linking

Augmenting the user's perception of her surroundings using a mobile device is a relatively new field of research which has been invigorated by the growth in number of capable mobile computing devices. These devices, while becoming increasingly small and inexpensive, allow us to use various computing facilities while roaming in the real world. In particular, ordinary mobile phones with integrated digital cameras are nowadays common, and even they can provide new ways to get access to digital information and services. Images or

[3] `http://suvi.viittomat.net/`

video captured by the mobile phone can be analyzed to recognize the object or scene appearing in the recording.

We studied new ways to get access to digital services for mobile phones in the Image Based Linking project in 2009–2011 [7, 8]. These kinds of methods can be used for various purposes linking digital information to the physical world. Possible application areas include outdoor advertising, additional digital material to magazine and newspaper articles, tourist applications, and shopping. Our focus in the project was on a use case with a magazine publisher as the content provider. Several target images can exist on the same page of a magazine, each linked to different extra information. Consequently, the target images may be rather small in print, and the captured photos may be highly blurred and out-of-focus. An example of matching such photos to the images in the magazine database is shown in Figure 7.3.

References

[1] Ville Viitaniemi, Mats Sjöberg, Markus Koskela, and Jorma Laaksonen. Concept-based video search with the PicSOM multimedia retrieval system. Technical Report TKK-ICS-R39, Aalto University School of Science and Technology, Department of Information and Computer Science, Espoo, Finland, December 2010.

[2] Ville Viitaniemi, Mats Sjöberg, Markus Koskela, and Jorma Laaksonen. Automatic video search using semantic concepts. In *Proceedings of 8th European Conference on Interactive TV and Video (EuroITV 2010)*, Tampere, Finland, June 2010.

[3] Ville Viitaniemi and Jorma Laaksonen. Region matching techniques for spatial bag of visual words based image category recognition. In *Proceedings of 20th International Conference on Artificial Neural Networks (ICANN 2010)*, volume 6352 of *Lecture Notes in Computer Science*, pages 531–540, Thessaloniki, Greece, September 2010. Springer Verlag.

[4] Mats Sjöberg, Markus Koskela, Ville Viitaniemi, and Jorma Laaksonen. Indoor location recognition using fusion of SVM-based visual classifiers. In *Proceedings of 2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 343–348, Kittilä, Finland, August-September 2010.

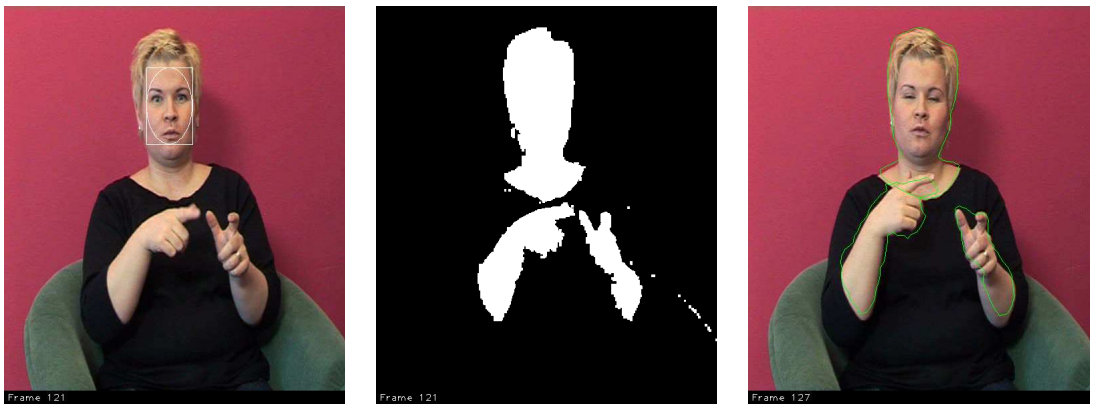


Figure 7.2: Example frames from the sign language video material. From left to right: Face detection, skin-color filtering, active shape models of skin regions.

Figure 7.3: An example of matching an image captured with a mobile phone and the corresponding magazine page.

[5] Matti Karppa, Tommi Jantunen, Markus Koskela, Jorma Laaksonen, and Ville Viitaniemi. Method for visualisation and analysis of hand and head movements in sign language video. In C. Kirchhof, Z. Malisz, and P. Wagner, editors, *Proceedings of the 2nd Gesture and Speech in Interaction conference (GESPIN 2011)*, Bielefeld, Germany, 2011. Available online as `http://coral2.spectrum.uni-bielefeld.de/gespin2011/final/Jantunen.pdf`.

[6] Tommi Jantunen, Markus Koskela, Jorma Laaksonen, and Päivi Rainò. Towards automated visualization and analysis of signed language motion: Method and linguistic issues. In *Proceedings of 5th International Conference on Speech Prosody*, Chicago, Ill. (USA), May 2010.

[7] Xi Chen, Markus Koskela, and Jouko Hyväkkä. Image based information access for mobile phones. In *Proceedings of 8th International Workshop on Content-Based Multimedia Indexing*, Grenoble, France, June 2010.

[8] Xi Chen and Markus Koskela. Mobile visual search from dynamic image databases. In *Proceedings of Scandinavian Conference on Image Analysis (SCIA 2011)*, Ystad, Sweden, May 2011.

Chapter 8

Automatic speech recognition

Mikko Kurimo, Kalle Palomäki, Janne Pylkkönen, Ville T. Turunen, Sami Virpioja, Ulpu Remes, Heikki Kallasjoki, Reima Karhila, Teemu Ruokolainen, Tanel Alumäe, Sami Keronen, André Mansikkaniemi, Peter Smit, Rama Sanand Doddipatla, Seppo Enarvi

8.1 Introduction

Automatic speech recognition (ASR) means an automated process that inputs human speech and tries to find out what was said. ASR is useful, for example, in speech-to-text applications (dictation, meeting transcription, etc.), speech-controlled interfaces, search engines for large speech or video archives, and speech-to-speech translation.

Figure 8.1 illustrates the major modules of an ASR system and their relation to applications. In *feature extraction*, signal processing techniques are applied to the speech signal in order to dig out the features that distinguish different phonemes from each other. Given the features extracted from the speech, *acoustic modeling* provides probabilities for different phonemes at different time instants. *Language modeling*, on the other hand, defines what kind of phoneme and word sequences are possible in the target language or application at hand, and what are their probabilities. The acoustic models and language models are used in *decoding* for searching the recognition hypothesis that fits best to the models. Recognition output can then be used in various applications.

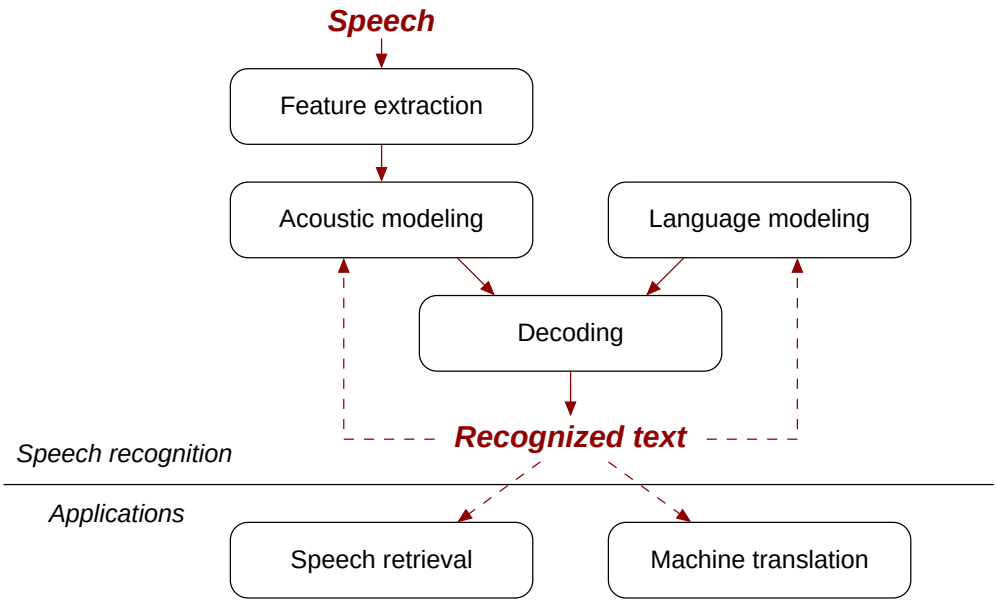


Figure 8.1: The main components of an automatic speech recognition system and their relation to speech retrieval and machine translation applications.

Our focus in ASR is large vocabulary continuous speech recognition (LVCSR). For several years, we have been developing new machine learning algorithms for each of the subfields and building a complete state-of-the-art recognizer to evaluate new methods and their impact. Originally, the recognizer was constructed for fluent and planned speech such as Finnish newsreading, where language models covering a very large vocabulary are required. Besides newsreading, other example tasks are political and academic speeches and other radio and television broadcasts where the language used is near the written style. Sofar, we have not seriously attempted to recognize Finnish spontaneous conversations, because enough Finnish training texts for learning the corresponding style do not exist. Our main training corpus for language modeling is the Finnish Language Bank at CSC. For acoustic modeling we use voice books, Finnish Broadcast Corpus at CSC and the SPEECON corpus.

In addition to the recognition of Finnish, we have performed experiments in English, Turkish and Estonian. To make this possible we have established research relations to different top speech groups in Europe and U.S., e.g. University of Colorado, International Computer Science Institute ICSI, Stanford Research Institute SRI, IDIAP, University of Edinburgh, University of Sheffield, University of Cambridge, Bogazici University, Tallinn University of Technology, and Nagoya Institute of Technology. The forms of collaboration have included researcher exchanges, special courses, workshops and joint research projects. We have also participated in several top international and national research projects funded by EU, Academy of Finland, Tekes, and our industrial partners. In the close collaboration with our Natural Language Processing group 10 we are also organizing an international competition called Morpho Challenge to evaluate the best unsupervised segmentation algorithms for words into morphemes for information retrieval, statistical machine translation, LVCSR and language modeling in different languages. This challenge project is funded by EU's PASCAL network and described in Chapter 10.

In the EU FP7 project called EMIME 2008-2011, the aim was to develop new technologies for spoken multilingual integration, such as speech-to-speech translation systems. This has broadened the field of the group to include some aspects of text-to-speech synthesis (TTS), such as supervised and unsupervised adaptation in the same way as in ASR. Successors of this project include a new EU FP7 project Simple4All which aims at developing unsupervised machine learning tools for rapid data-driven development for new TTS systems by adaptation and a new project Perso which aims at developing new Finnish TTS systems by adaptation.

Other new openings in the group are developing adaptation methods for special purpose dictation (e.g. in medical domain in Mobster project), using ASR in various multimodal human-computer interaction (e.g. in augmented reality in UI-ART project), and audiovisual indexing (e.g. television broadcasts in NextMedia project).

8.2 Training and adaptation of acoustic models

Acoustic modeling in automatic speech recognition means building statistical models for some meaningful speech units based on the feature vectors computed from speech. In most systems the speech signal is first chunked into overlapping 20-30 ms time windows at every 10 ms and the spectral representation is computed from each frame. A commonly used feature vector consists of Mel-frequency cepstral coefficients (MFCC) which are the result of the discrete cosine transform (DCT) applied to the logarithmic Mel-scaled filter bank energies. Local temporal dynamics can be captured by concatenating the first and second order delta features (time differences) to the basic feature vector.

The acoustic feature sequence in ASR is typically modeled using hidden Markov models (HMM). In a simple system each phoneme is modeled by a separate HMM, where the emission distributions of the HMM states are Gaussian mixtures (GMMs). In practice, however, we need to take the phoneme context into account. In that case each phoneme is modeled by multiple HMMs, representing different neighboring phonemes. This leads easily to very complex acoustic models where the number of parameters is in order of millions.

Estimating the parameters of complex HMM-GMM acoustic models is a very challenging task. Traditionally maximum likelihood (ML) estimation has been used, which offers simple and efficient re-estimation formulae for the parameters. However, ML estimation does not provide optimal parameter values for classification tasks such as ASR. Instead, discriminative training techniques are nowadays the state-of-the-art methods for estimating the parameters of acoustic models. They offer more detailed optimization criteria to match the estimation process with the actual recognition task. The drawback is increased computational complexity. Our implementation of the discriminative acoustic model training allows using several different training criteria such as maximum mutual information (MMI) and minimum phone error (MPE) [1]. Also alternative optimization methods such as gradient based optimization and constrained line search [2] can be used in addition to the commonly used extended Baum-Welch method. Our recent research has concentrated on comparing the different optimization strategies and finding the most effective ways to train well-performing robust acoustic models [3].

As acoustic models have a vast amount of parameters, a substantial amount of data is needed to train these models robustly. In the case a model needs to be targeted to a specific speaker, speaker group or other condition, not always sufficient data is available. The generic solution for this is to use adaptation methods like Constrained Maximum Likelihood Linear Regression [4] to transform a generic model in to a specific model using a limited amount of data. In [5] and [6] this method was repeatedly applied to a model, so that first a transformation to a foreign accented model was made and successively a transformation to a speaker-specific model. These stacked transformations improved up to 30% recognition accuracy, depending on the accent and amount of available data for the speaker. In Figure 8.2 the improvement in word error rate is shown for different amounts of speaker adaptation data and for both a native and a mixed acoustic model.

Vocal Tract Length Normalization (VTLN) has become an integral part of the standard adaptation toolkit for ASR. This method approximates physical properties of each speaker's vocal tract and shifts accordingly the frequency components of the speech to be recognized. The simple old school way of applying VTLN was to warp the cut-off frequencies in the filter bank analysis, before transforming the frequency channels of the

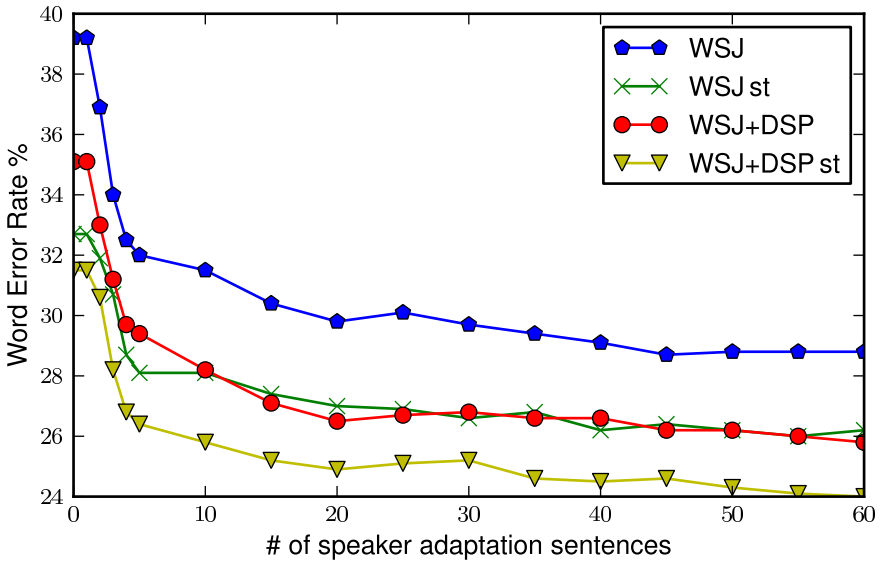


Figure 8.2: This figure shows the improvement that Stacked transformations (*st*) give over normal CMLLR adaptation. The WSJ is native English and the DSP dataset is Finnish-accented English speech. Stacked transformation have the most effect when only a small number of adaptation sentences is used.

speech sample to cepstral components. In the current approach, VTLN is represented as a CMLLR-style linear transformation on the conventional MFCC features. Using VTLN as a linear transformation on the MFCC features allowed us to study the curious interplay of CMLLR and VTLN adaptation methods and the use of VTLN to to boost other speaker adaptation methods [7].

Acoustic modeling of parametric speech synthesis

The rising paradigm of HMM-based statistical parametric speech synthesis relies on ASR-style acoustic modelling. Speech synthesis, or Text-To-Speech (TTS) models are more descriptive and less generalized than the ASR models. They try to accurately describe the numerous, variously stressed phones, and therefore the model sets are much larger than the ASR model sets. Training acoustic models for high-quality voice for a TTS system requires data of close to 1000 high-quality sentences from the target speaker. The adaptation of HMM-based TTS models is very similar to adaptation of ASR models. Maximum a posteriori (MAP) linear transformations are applied in similar fashion to ASR adaptation. A collaborative investigation using data from several languages showed that adapting a general voice is a practical and effective way to mimic a target speaker's voice[8].

The speech synthesis work related to the EMIME EU/FP7 project concentrated on the adaptation of HMM-based TTS models. The goal of the project was to personalize the output voice of a cross-lingual speech-to-speech system, to make it resemble the voice of the original speaker [9]. This is accomplished by adapting the acoustic features of the synthesis model set in one language (Source language, L1) and mapping these transformations to a second model set (Target language, L2). The goal of the Cross-Lingual Speaker Adaptation (CLSA) is to effectively model speakers' speech in another language. As a

person's speech in a foreign language depends, beside physical characteristics, also very much on the environmental factors - mostly how much and in what kind of linguistic environment has the speaker practised speaking the language, it is almost impossible to predict how a person would in reality sound in the second language. We investigated what kind of expectations listeners usually have about a speaker's voice in a second language, and particularly whether the listeners preferred a foreign- or native accented voice model for basis of adaptation, a very important aspect in real-life situation where only little data is available for adaptation [10].

References

[1] D. Povey and P. C. Woodland, Minimum Phone Error and I-smoothing for Improved Discriminative Training. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, pages I-105–108, 2002.

[2] P. Liu, C. Liu, H. Jiang, F. Soong, and R.-H. Wang, A constrained line search optimization method for discriminative training of HMMs. In *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 900–909, 2008.

[3] J. Pylkkönen, Investigations on Discriminative Training in Large Scale Acoustic Model Estimation. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, pp. 220–223, 2009.

[4] M.J.F. Gales, Maximum likelihood linear transformations for HMM-based speech recognition. In *Computer speech and language*, vol. 12, pp. 75–98, 1998.

[5] P. Smit and M. Kurimo, Using stacked transformations for recognizing foreign accented speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 5008–5011, May 2011.

[6] P. Smit, Stacked transformations for foreign accented speech recognition. *Masters' thesis*

[7] D.R. Sanand and M. Kurimo, A Study on Combining VTLN and SAT to Improve the Performance of Automatic Speech Recognition. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association, INTERSPEECH*, Florence, August 2011.

[8] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian,R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, M. Kurimo, Thousands of Voices for HMM-based Speech Synthesis. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, pp. 420–423, 2009.

[9] M. Wester, J. Dines, M. Gibson, H. Liang, Y.-J. Wu, L. Saheer, S. King, K. Oura, P. N. Garner, W. Byrne, Y. Guan, T. Hirsimäki, R. Karhila, M. Kurimo, M. Shannon, S. Shiota, J. Tian, K. Tokuda, and J. Yamagishi, Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In *Proceedings of the 7th ISCA Speech Synthesis Workshop, SSW7*, ISCA, September 2010.

[10] R. Karhila and M. Wester, Rapid adaptation of foreign-accented HMM-based speech synthesis. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association, INTERSPEECH*, Florence, August 2011.

8.3 Noise robust speech recognition

Despite the steady progress in speech technology, robustness to background noise remains a challenging research problem as the performance gap between automatic speech recognition and human listeners is widest when speech is corrupted with noise. The work presented in this section is focussed on methods that model the uncertainty in the observed or reconstructed (cleaned) speech features when the clean speech signal is corrupted with noise from an unknown source. In addition to the uncertainty-based methods presented here, we have continued the work on noise robust feature extraction using weighted linear prediction [1].

Missing feature approaches

The so called missing-feature methods are a special case of methods that use observation uncertainty or reliability in order to improve speech recognition performance in noisy conditions. The methods, which draw inspiration from the human auditory system, are based on the assumption that speech corrupted by noise can be divided to speech-dominated i.e. reliable regions and noise-dominated i.e. unreliable regions as illustrated in Figure 8.3. The clean speech information corresponding to the unreliable regions is assumed missing,

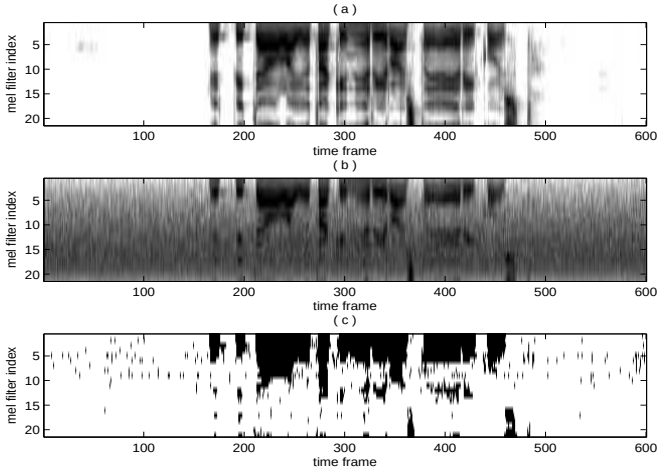


Figure 8.3: Logarithmic mel spectrogram of (a) an utterance recorded in quiet environment and (b) the same utterance corrupted with additive noise. The noise mask (c) constructed for the noisy speech signal indicates the speech dominated regions in black and the noise dominated regions in white.

which means that under additive noise assumption, the observed values determine an upper bound for the unobserved clean speech features but contain no further information regarding the missing values. In noise-robust speech recognition, the missing clean speech information is either marginalised over or reconstructed using missing-feature imputation techniques [2]. The reconstruction approach was compared with other noise-robust speech recognition methods in [3].

Reconstruction methods are based on modelling the statistical dependencies between clean speech features and using the model and the reliable observations to calculate clean speech estimates for the missing values. Recent improvements to missing-feature imputation are

due to modelling the temporal dependencies between clean speech features in consecutive frames. Processing the noisy speech in windows that span several time frames was first proposed in the exemplar-based sparse imputation (SI) framework [4]. SI outperformed the conventional GMM-based imputation method that used frame-based processing. Window-based processing was later introduced in the GMM-based framework in [5], and to investigate other approaches to temporal modelling, a nonlinear state-space model (NSSM) based framework was developed for missing-feature reconstruction in [6]. Both the window-based GMM and the NSSM imputation method outperformed frame-based GMM imputation in all experiments and outperformed SI when evaluated under loud impulsive noise.

In addition to work on improving the core missing feature methods, we have studied missing feature methods in models of human hearing. Related to this work, we proposed a model that explains the speech recognition performance of human listeners in a binaural listening scenario [7]. Furthermore, we have applied the missing-feature reconstruction methods developed for noise-robust speech recognition to extending the bandwidth of narrowband telephone speech to the high frequency band [8] and the low frequency band [9]. The latter study won the International Speech Communication Association award for the best student paper in Interspeech 2011.

Modelling uncertainty in reconstruction

In addition to using reliability estimates to determine reliable and unreliable features in missing-feature reconstruction, we have studied using another type of reliability estimates to improve the speech recognition performance when reconstructed or otherwise enhanced speech data is used. First, we have studied uncertainty estimation in the context of sparse imputation [10, 11]. Unlike the parametric methods that model clean speech using a GMM or NSSM, for example, the exemplar-based sparse imputation method does not provide for calculating a full posterior for the reconstructed features. We therefore investigated using a number of heuristic measures to represent the uncertainty related to the SI reconstruction performance. Similarly, we have developed a number of heuristic uncertainty measures for the exemplar-based sparse separation technique that uses a speech and noise dictionary to estimate clean speech features based on the noisy observations [12].

References

[1] S. Keronen, J. Pohjalainen, P. Alku, and M. Kurimo, Noise robust LVCSR feature extraction based on extended weighted linear prediction. Proc. INTERSPEECH, 2011.

[2] B. Raj and R. M. Stern, Missing-feature approaches in speech recognition. *IEEE Signal Processing Magazine*, vol. 22, pages 101–116, 2005.

[3] S. Keronen, U. Remes, K. J. Palomäki, T. Virtanen and M. Kurimo, Comparison of noise robust methods in large vocabulary speech recognition, Eusipco 2010.

[4] J. F. Gemmeke, B. Cranen, and U. Remes (2011). Sparse imputation for large vocabulary noise robust ASR. Computer Speech and Language, vol 25, issue 2, pp. 462-479, 2011.

[5] U. Remes, Y. Nankaku, and K. Tokuda, GMM-based missing feature reconstruction on multi-frame windows. Proc. INTERSPEECH, pp. 1665-1668, Florence, Italy, August 2011.

[6] U. Remes, K. J. Palomäki, T. Raiko, A. Honkela and M. Kurimo, Missing-feature reconstruction with bounded nonlinear state-space model, IEEE Signal processing letters, 18(10), 563-566, 2011

[7] K. J. Palomäki and G. J. Brown A computational model of binaural speech recognition: role of across-frequency vs. within-frequency processing and internal noise, Speech Communication, 53(6), 924-940, 2011

[8] H. Pulakka, U. Remes, K. J. Palomäki, M. Kurimo, P. Alku, Speech bandwidth extension using Gaussian Mixture Model-based estimation of the highband Mel spectrum. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 11), Prague, Czech Republic, May 22-27, 2011.

[9] H. Pulakka, U. Remes, S. Yrttiaho, K. Palomäki, M. Kurimo and P. Alku, Low-frequency bandwidth extension of telephone speech using sinusoidal synthesis and gaussian mixture model. In Proc. Interspeech 2011, Florence, Italy, Aug. 28-31, 2011.

[10] J. Gemmeke, U. Remes and K. J. Palomäki, Observation uncertainty measures for sparse imputation, Interspeech 2010.

[11] H. Kallasjoki, S. Keronen, G. J. Brown, J. F. Gemmeke , U. Remes and K. J. Palomäki, Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments, in International Workshop on Machine Listening in Multisource Environments, 2011.

[12] H. Kallasjoki, U. Remes, J. F. Gemmeke, T. Virtanen and K. J. Palomäki, Uncertainty measures for improving exemplar-based source separation, in Proc. Interspeech 2011.

8.4 Constraining and adapting language models

Early speech recognition systems used rigid grammars to describe the recognized language. Typically the grammar included a limited set of sentences used to command the system. Such language models do not scale for large vocabulary continuous speech recognition. Therefore modern recognizers, including the Aalto University recognizer, use statistical language models.

Constrained command languages are still useful in some spoken dialog applications, where commands are important to be recognized correctly, especially if the system cannot be adapted to a specific user group. We have succesfully built statistical language models from command grammars, modeled in Backus-Naur Form (BNF). Language models built in this way enable fast decoding and near perfect recognition accuracy.

When large-vocabulary speech recognition is applied in a specialized domain, the vocabulary and speaking style may substantially differ from those in the corpra that are available for Finnish language. Using additional text material from the specific domain, when estimating the language model, is beneficial, or even necessary for proper recognition accuracy. We have applied speech recognition to medical transcription. A huge collection of dental reports was received from In Net Oy, for estimating a language model specific to dental dictation. User tests are underway, but our benchmarks indicate large differences in accuracy between different users.

Collecting domain-specific texts is time-consuming and usually there's not enough data available to estimate a reliable language model. Most of the times we have to use the little in-domain data we have to adapt the general language model.

In a project aimed at developing a mobile dictation service for lawyers, we used law-related texts to train an in-domain language model [1]. Adapting the general language model with the in-domain model usually gave better results than just using either model separately. One of the key challenges of the project was still to find proper adaptation data. Even though the adaptation texts are of the targeted domain, the language of the real-life dictations can still be significantly different than the written text.

Language model adaptation usually consists of mixing or combining the probabilities of the general language model with the in-domain model. The most simple and popular LM adaptation method is linear interpolation. Linear interpolation is performed by simply calculating a weighted sum of the two models probabilities.

We have experimented with a more sophisticated LM adaptation method, which uses the information theory principle of maximum entropy (ME) to adapt the general language model with the in-domain model [2]. The key to this approach is that the global and domain-specific parameters are learned jointly. Domain-specific parameters are largely determined by global data, unless there is good domain- specific evidence that they should be different. We tested the method on English and Estonian broadcast news and experiments showed that the method consistently outperformed linear interpolation. The main drawback with this method is that it's very memory and time consuming.

The implementation of ME language model adaptation is freely available as an extension to the SRI language modeling toolkit [3].

References

[1] A. Mansikkaniemi. Acoustic Model and Language Model Adaptation for a Mobile Dictation Service. *Master's thesis, Aalto University*, 2010.

[2] T. Alumäe and M. Kurimo, Domain Adaptation of Maximum Entropy Language Models, Proceedings of the ACL 2010, Uppsala, Sweden, July 2010.

[3] T. Alumäe and M. Kurimo, Efficient Estimation of Maximum Entropy Language Models with N-gram features: an SRILM extension, Proceedings of Interspeech 2010, Chiba, Japan, September 2010.

8.5 Speech retrieval and indexing

Speech retrieval techniques enable users to find segments of interest from large collections of audio or video material. Automatic speech recognition is used to transforms the spoken segments in the audio to textual form. Information retrieval (IR) methods are used to index the text, and to perform searches on the material based on query words typed by the user. Since the amount of information in spoken form is very large and ever increasing, the methods developed have to be fast and robust to be able to process large amounts of variable quality material.

One complication in the speech retrieval process is the fact that the speech recognizer output will always have erroneous words. A special problem for speech retrieval are *out-of-vocabulary* (OOV) words – words that are not in the list of words the speech recognizer knows. Any OOV word in speech can not be recognized, and is replaced by similarly sounding but usually unrelated word. Since query words are chosen to be discriminative, they are often rare words such as proper names. But rare words are often also OOV, since the recognizer vocabulary is chosen so that a number of most common words are included.

This problem can be solved by using recognition units that are smaller than words, but that are large enough to be able to model the language. Morphs produced by the Morfessor algorithm have been proven to work well as such units. The speech recognizer language model is trained on a text corpus where the words are split to morphs, and the recognizer is then able to transcribe any word in speech by recognizing its component morphs. It is possible to join the morphs to words and use traditional morphological analyzers to find the base forms of the words for indexing. But since there will still be an amount of errors in the morph transcripts, especially when the spoken word is previously "unseen", a word that did not appear in the language model training corpus, using morphs as index terms will allow utilizing the partially correct words as well. In this case, query words are also split to morphs with Morfessor. Experiments using Finnish radio material show that morphs and base forms work about equally well as index terms, but combining the two approaches gives better results that either alone [1]. Table 8.1 shows an example how OOV words are recognized with word and morph language models.

Table 8.1: Example recognition results of two unseen query words at two different locations each. With the morph language model, it is possible to recognize correctly at least some of the morphs, which will match morphs in the query. With the word language model, the words are replaced by unrelated words.

Query word	Iliescun		Namibian	
- Translation	*Iliescu's*		*Namibia's*	
Morph query	ili escu n		na mi bi an	
Morph LM rec.	n ilja escu	ili a s kun	ami bi an	na min pi an
Word query	iliescun		namibia	
Word LM rec.	lieskoja	eli eskon	anjan	namin pian
Word lemmas	lieska	eli elää esko	anja	nami pian pia
- Translation	*flame*	*or live Esko*	*Anja*	*candy soon Pia*

Audio and video is typically distributed as a flow of material without any structure or indicators where the story changes. Thus, before indexing, the material needs to be automatically segmented into topically coherent speech documents. This can be done e.g. by measuring the lexical similarity of adjacent windows. Morphs were found to help in

the segmentation task as well when processing ASR transcripts [1].

Retrieval performance can be further improved by utilizing alternative recognition candidates from the recognizer [1]. Retrieval performance is decreased if a relevant term is misrecognized and is thus missing from the transcript. However, it is possible that the correct term was considered by the recognizer but was not the top choice. Thus, retrieval performance can be improved by extracting these alternative results from the recognizer and adding them to the index. A *confusion network* [2] provides a convenient representation of the competing terms along with a probability value for each term.

References

[1] V.T. Turunen, and M. Kurimo, Speech retrieval from unsegmented Finnish audio using statistical morpheme-like units for segmentation, recognition, and retrieval. *ACM Transactions on Speech and Language Processing*, Vol. 8, No. 1, pp. 1–25, October 2011.

[2] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech And Language*, 14:373–400, 2000.

Chapter 9

Proactive Interfaces

Samuel Kaski, Erkki Oja, Jorma Laaksonen, Mikko Kurimo, Arto Klami, Markus Koskela, Mehmet Gönen, Antti Ajanki, He Zhang, Melih Kandemir, Teemu Ruokolainen, Andre Mansikkaniemi, Jing Wu, Chiwei Wang

9.1 Introduction

The Proactive Interfaces research theme combines efforts of multiple research groups, including the Statistical Machine Learning and Bioinformatics group, lead by Professor Samuel Kaski, and the Content-Based Information Retrieval and Speech Recognition groups, lead by Professor Erkki Oja. Since 2008, major collaborative EU FP7, EIT ICT Labs and Aalto funded projects have been carried on which together form the AIRC flagship project *Proactive Interfaces*.

9.2 Inferring interest from implicit signals

Proactive systems anticipate the user's intentions and actions, and utilize the predictions to provide more natural and efficient user interfaces. One of the critical components in this loop is inferring the interests of the user, which is a challenging machine learning problem. Successful proactivity in varying contexts requires generalization from past experience. Generalization, on its part, requires suitable powerful (stochastic) models and a collection of data about relevant past history to learn the models.

We have studied inferring interest from eye movement patterns. Eye gaze location is a good proxy for attention but explicit eye movement control is tiresome. Therefore, we study methods that can infer relevance implicitly during normal viewing. Estimated relevance can be used as feedback for an information retrieval system.

We experimented with eye movements and other modalities as source of implicit feedback in image retrieval [1]. It is possible to predict relevant images relatively well from eye movements. We made a feasibility study on predicting the relevance of objects in a video from viewers' eye movements [2]. This setup is an extension of our earlier eye tracking studies on static text and image retrieval setups to dynamic scenes. Even with a relatively simple logistic regression predictor the eye movements predict the relevance with an encouraging accuracy.

The ability to infer relevance in dynamic scenes allows us to do proactive information retrieval in the context of the real world environment [3] which is a novel task. With modern data glasses, which have both augmentation and eye tracking capabilities, it is possible to track the user's attention on real and virtual objects and provide presently relevant information. The data glasses are provided to us by Nokia Research Center (NRC).

Other physiological signals than eye movements are also useful in inferring latent cognitive and emotional state. In [4] we show that learning a combined model of accelerometer, EEG, eye tracker and heart-rate sensors improves prediction accuracy over measurements from individual sensors.

In [5] we introduce a proactive retrieval interface for time-ordered image datasets such as personal lifelogs. Humans can effectively recognize familiar images and use them as reference points when navigating images on a timeline. The system further helps by making relevant images more salient. Relevance is estimated from explicit and implicit mouse movement features.

9.3 Eye-movement enhanced image retrieval

Personal Information Navigator Adapting Through Viewing (PinView)[1] was an EU FP7 funded three-year Collaborative Project coordinated by AIRC. It was started on 1 January 2008 and ended on 31 March 2011. The goal of PinView was a proactive personal information navigator that allows retrieval of multimedia – such as still images, text and video – from unannotated databases. During image browsing and searching with a task-dependent interface, the PinView system infers the goals of the user from explicit and implicit feedback signals and interaction, such as speech, eye movements and pointer traces and clicks, complemented with social filtering. The collected rich multimodal responses from the user are processed with new advanced machine learning methods to infer the implicit topic of the user's interest as well as the sense in which it is interesting in the current context.

The PinView consortium combined pioneering application expertise with a solid machine learning background in content-based information retrieval. Besides AIRC, the project consortium included University of Southampton (UK), University College London (UK), Montanuniversitaet Leoben (AU), Xerox Research Centre Europe (FR), and celum gmbh (AU).

The foremost output of the project was the PinView content-based image retrieval system, that uses (1) the LinRel algorithm for balancing the exploration–exploitation trade-off in image selection, and (2) the Multiple Kernel Learning algorithm for optimal use of available low-level image features for iterative online relevance feedback. The PinView method is able to make use of both explicit relevance feedback, given by pointer clicks on images, and implicit feedback obtained from estimated image relevances based on the user's eye movements while viewing retrieved images. Empirical evaluations have proven the efficiency and scalability of the PinView system in realistic small- and large-scale image retrieval experiments.

In a nutshell, a well-functioning novel search engine was implemented as illustrated in Figure 9.1 and scaling it to huge image collections was found to be feasible. User requirement studies were performed in the initial stage of the project. Later the PinView system was evaluated in four user studies that originated from genuine use case scenarios. These experiments showed that the gaze-based implicit relevance feedback clearly improved image retrieval accuracy and speed compared to the baseline of random browsing. As it can be expected that the price and size of eye tracking devices will continue diminishing while their accuracy and usability are concurrently improving, the effortless combination of browsing and proactive retrieval based on implicit gaze feedback will be useful and available on a large scale. During the project, seven peer-reviewed journal and 32 conference papers have been published by the PinView consortium, including e.g. [1, 6, 7].

9.4 Contextual information interfaces

Contextual information interfaces provide access to information that is relevant in the current context. They use sensory signals, such as gaze patterns, to track the user's context and foci of interest, and to predict what kind of information the user would need at the present time. The information is retrieved from databases and presented in a non-intrusive manner. Main challenges are extraction of context from visual and sensory data,

[1] `http://www.pinview.eu/`

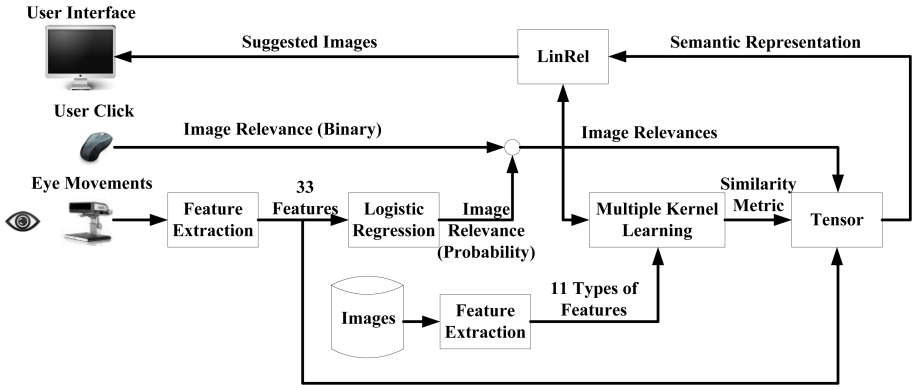


Figure 9.1: Main components and data flow in the PinView content-based image retrieval system that makes use of machine learning of implicit relevance feedback from eye movements.

construction of adaptive machine learning models that are able to utilize heterogeneous context cues to predict relevance, and undisturbing and easily understandable presentation of information. Novel statistical machine learning methods are used for multimodal information retrieval and for taking the context into account.

As a part of *Urban Contextual Information Interfaces with Multimodal Augmented Reality (UI-ART)* project[2], an interdisciplinary research project funded by Aalto *Multidisciplinary Institute of Digitalisation and Energy (MIDE)* programme, we have built a pilot system

[2] `http://mide.aalto.fi/en/UI-ART`

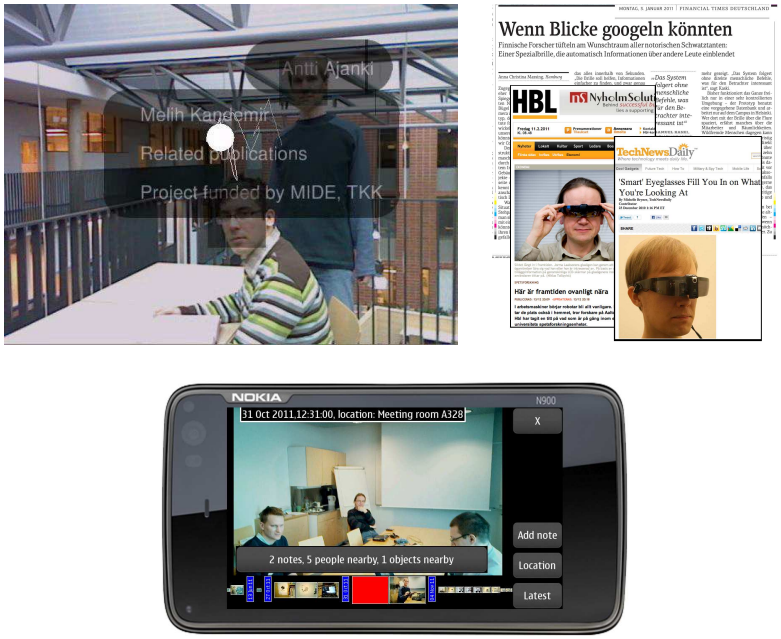


Figure 9.2: Top left: a near-eye display screenshot of the UI-ART contextual information interface. Top right: Our work received international media coverage in 2011. Bottom: Smart phone interface of the UI-ART system.

that retrieves and displays abstract information about people and real world objects in augmented reality [3]. As a pilot application scenario, we have implemented a guide that displays relevant information to a participant in a scientific workshop or meeting or a visitor at a university department. The interface consists of either a head-worn display with an integrated gaze-tracker or a smart phone that can be pointed towards an interesting object. People and objects in the view are recognized from the video feed [8] and information related to them is searched from a database. Retrieved textual annotations are augmented to the view and become part of the context the user can attend to. Evidence from gaze measurements and speech recognition is integrated to infer the user's current interests and annotations that match those are displayed. Figure 9.2 shows snapshots of the UI-ART system's augmented reality display.

In addition to the UI-ART project, the *Proactive Interfaces research group* participated in the *Device and Interoperability Ecosystem (DIEM)* research programme of the TIVIT ICT SHOK from July 2008 to December 2011. The project targeted to enable new services and applications for smart environments that comprise of digital devices containing relevant information for different purposes. The project involved Nokia Research Center (NRC) and Technical Research Centre of Finland (VTT) as collaborators. In 2011, the work was expanded to EIT ICT Labs' Smart Spaces thematic Action Line project *Pervasive Information, Interfaces, and Interaction (PI3)*, where co-operation was been carrier out with research groups from all EIT ICT Labs nodes.

References

[1] Peter Auer, Zakria Hussain, Samuel Kaski, Arto Klami, Jussi Kujala, Jorma Laaksonen, Alex P. Leung, Kitsuchart Pasupa, and John Shawe-Taylor. Pinview: Implicit feedback in content-based image retrieval. In Tom Diethe, Nello Cristianini, and John Shawe-Taylor, editors, *Proceedings of Workshop on Applications of Pattern Analysis*, volume 11 of *JMLR Workshop and Conference Proceedings*, pages 51–57, 2010.

[2] Melih Kandemir, Veli-Matti Saarinen, and Samuel Kaski. Inferring object relevance from gaze in dynamic scenes. In *Proceedings of ETRA 2010, ACM Symposium on Eye Tracking Research & Applications, Austin, TX, USA, March 22-24*, pages 105–108, New York, NY, 2010. ACM.

[3] Antti Ajanki, Mark Billinghurst, Hannes Gamper, Toni Järvenpää, Melih Kandemir, Samuel Kaski, Markus Koskela, Mikko Kurimo, Jorma Laaksonen, Kai Puolamäki, Teemu Ruokolainen, and Timo Tossavainen. An augmented reality interface to contextual information. *Virtual Reality*, 15(2-3):161–173, 2011.

[4] Mehmet Gönen, Melih Kandemir, and Samuel Kaski. Multitask learning using regularized multiple kernel learning. In Bao-Liang Lu, Liqing Zhang, and James Kwok, editors, *Proceedings of 18th International Conference on Neural Information Processing (ICONIP)*, volume 7063 of *Lecture Notes in Computer Science*, pages 500–509, Berlin / Heidelberg, 2011. Springer.

[5] Antti Ajanki and Samuel Kaski. Probabilistic proactive timeline browser. In Timo Honkela, Wlodzislaw Duch, Mark A. Girolami, and Samuel Kaski, editors, *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN), Part II*, Lecture Notes in Computer Science, pages 357–364, Berlin, 2011. Springer.

[6] Arto Klami. Inferring task-relevant image regions from gaze data. In Samuel Kaski, David J. Miller, Erkki Oja, and Antti Honkela, editors, *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 101–106. IEEE, 2010.

[7] He Zhang, Teemu Ruokolainen, Jorma Laaksonen, Christina Hochleitner, and Rudolf Traunmüller. Gaze- and speech-enhanced content-based image retrieval in image tagging. In *Proceedings of 21st International Conference on Artificial Neural Networks (ICANN 2011)*, Espoo, Finland, 2011.

[8] Jing Wu. Online face recognition with application to proactive augmented reality. Master's thesis, Aalto University School of Science and Technology, Department of Information and Computer Science, May 2010.

Chapter 10

Natural language processing

Krista Lagus, Mikko Kurimo, Timo Honkela, Sami Virpioja, Oskar Kohonen, Mari-Sanna Paukkeri, Tiina Lindh-Knuutila, Ville T. Turunen, Ilkka Kivimäki, Laura Leppänen, Sini Pessala, Santosh Tirunagari

10.1 Introduction

Work in the field of natural language processing involves several research themes that have close connections to work carried out in other groups, especially speech recognition (Chapter 8) and Computational Cognitive Systems groups (Chapter **??**). The objective of this research is to develop methods for learning general-purpose representations from text that can be applied to the recognition, understanding and generation of natural language. The results are evaluated in applications such as automatic speech recognition, information retrieval, and statistical machine translation.

During 2010–2011, our research has concentrated on minimally supervised and language-independent methods for morphology induction, keyphrase extraction, and creation and evaluation of vector space models.

10.2 Unsupervised and semi-supervised morphology induction

In the theory of linguistic morphology, morphemes are considered to be the smallest meaning-bearing elements of language, and they can be defined in a language-independent manner. It seems that even approximative morphological analysis is beneficial for many natural language applications dealing with large vocabularies, such as speech recognition and machine translation. These applications usually use *words* as vocabulary units. However, for highly-inflecting and agglutinative languages, this leads to very sparse data, as the number of possible word forms is very high.

Figure 10.1 shows the very different rates at which the vocabulary grows in various text corpora of the same size. For example, the number of different unique word forms in the Finnish corpus is considerably higher than in the English corpus. In addition to the language, the size of the vocabulary is affected by the genre.

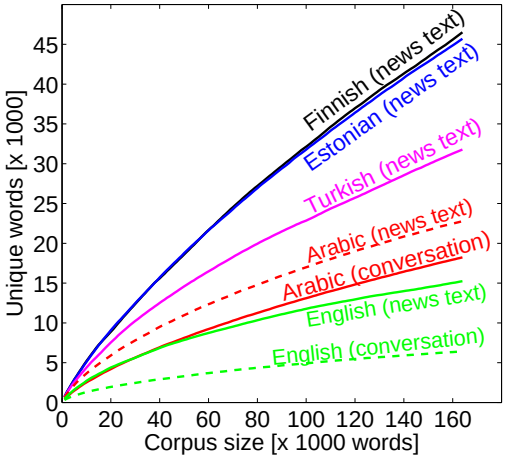


Figure 10.1: The number of different word forms (types) encountered in growing portions of running text (tokens) of various languages and text types.

Apart from practical use in various natural language processing applications, learning the phenomena underlying word construction in natural languages is an important question in psycholinguistics. Psycholinguistic questions regarding morphology include, for example, how the different word forms are learned, constructed, and stored in our mind in the so-called mental lexicon.

In 2010, we continued the series of Morpho Challenge competitions previously organized in 2005, 2007, 2008, and 2009. The objective of Morpho Challanges is to design statistical machine learning algorithms that discover discover the set of morphemes from which words are constructed [1]. The Morpho Challenge 2010 was funded by the EU Network of Excellence PASCAL2 Challenge Program. The evaluations included four languages and three evaluations: comparison to a linguistic gold standard, evaluation in an information retrieval task, and evaluation in a machine translation task. As a new task we introduced semi-supervised learning, in which a small set of linguistic gold standard morpheme analyses are provided as a training set. Four international groups participated in the Challenge, and the results and algorithms were published in a technical report [2].

Based on the Morpho Challenge results collected over five years, we have performed an extensive meta-evaluation of various evaluation methods for unsupervised learning of morphology [3]. Apart from comparing existing methods, we have further developed the evaluation methods and published evaluation software for the research community.

We have also continued to develop Morfessor [4], an unsupervised method for morphology induction. In *Allomorfessor* [5], Morfessor has been extended to account for the linguistic phenomenon of allomorphy. In allomorphy, an underlying morpheme-level unit has two or more surface realizations (e.g., "day" has an alternative surface form "dai" in "daily"). Allomorfessor has performed well in Morpho Challenge evaluations, although the amount of allomorphs found by the algorithm was limited.

In order to enable Morfessor to model complex morphological phonomena such as allomorphy, as well as to provide a reasonable baseline for the semi-supervised learning evaluated in Morpho Challenge 2010, we have also developed a semi-supervised learning algorithms for Morfessor [6]. The linguisic evaluation of Morpho Challenge shows that the accuracy of Morfessor improves rapidly already with small amounts of labeled data, surpassing the state-of-the-art unsupervised methods at 1000 labeled words for English and at 100 labeled words for Finnish. A further extension of the method has achieved the best published results for the semi-supervised learning setup of Morpho Challenge [7]. We have also studied the effect of word frequencies learning in generative models of morphology such as Morfessor, and found that using logaritmically dampened frequencies seem to provide better results than learning on word tokens and at least as good results than learning on word types [8].

Finally, in collaboration with the Brain Research Unit of the O.V. Lounasmaa laboratory at Aalto University, we have developed psycholinguistic framework for evaluating machine learning of morphology [9]. We use reaction times in a word recognition task as a proxy that provides an indirect measure of the underlying mental processing. In general, longer reaction times reflect more effortful cognitive processing. In comparison of several statistical models revealed that Morfessor Categories-MAP [4] provides an accurate and compact model for the reaction time data. Moreover, we observed a strong effect for the type and amount of the training data to the correlations. Figure 10.2 shows how Morfessor Categories-MAP predicts too high reaction times for abstract words such as *knowledge* and too low reactions times for concrete words such as *mother*.

References

[1] Mikko Kurimo, Sami Virpioja, Ville Turunen, and Krista Lagus. Morpho challenge 2005-2010: Evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 87-95, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[2] Mikko Kurimo, Sami Virpioja, and Ville T. Turunen (Eds.). *Proceedings of the Morpho Challenge 2010 workshop*. Technical Report TKK-ICS-R37, Aalto University School of Science and Technology, Department of Information and Computer Science, Espoo, Finland, September 2010.

[3] Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2), 2011.

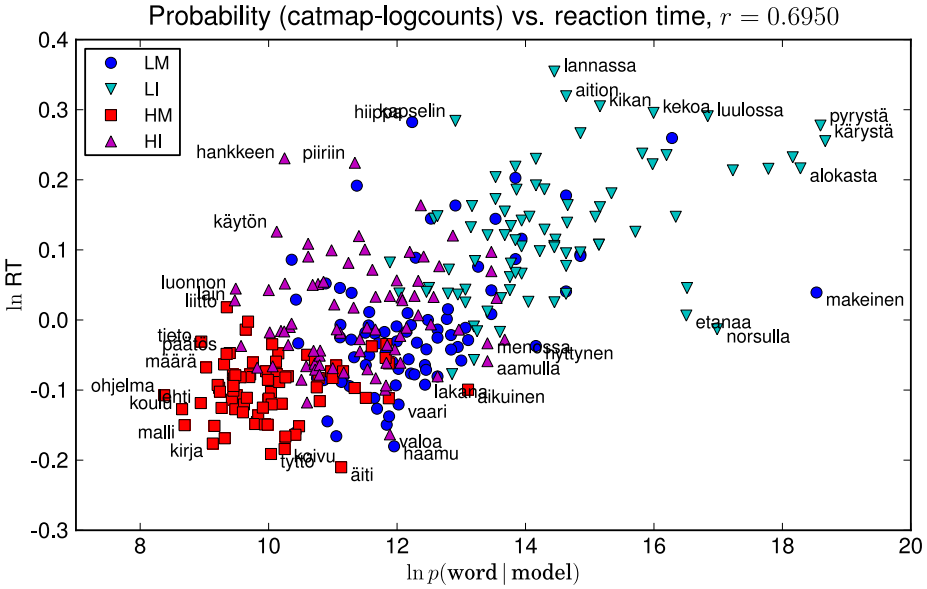


Figure 10.2: Scatter plot of reaction times and log-probabilities from Morfessor Categories-MAP [9]. The words are divided into four groups: low-frequency monomorphemic (LM), low-frequency inflected (LI), high-frequency monomorphemic (HM), and high-frequency inflected (HI). Words that have faster reaction times than predicted are often very concrete and related to family, nature, or stories: *tyttö* (girl), *äiti* (mother), *haamu* (ghost), *etanaa* (snail + partitive case), *norsulla* (elephant + adessive case). Words that have slower reaction times than predicted are often more abstract or professional: *ohjelma* (program), *tieto* (knowledge), *hankkeen* (project + genitive case), *käytön* (usage + genitive case), *hiippa* (miter), *kapselin* (capsule + genitive case).

[4] Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, Volume 4, Issue 1, Article 3, January 2007.

[5] Sami Virpioja, Oskar Kohonen, and Krista Lagus. Unsupervised morpheme analysis with Allomorfessor. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments: 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*, volume 6241 of Lecture Notes in Computer Science, pages 609-616. Springer Berlin / Heidelberg, September 2010.

[6] Oskar Kohonen, Sami Virpioja, and Krista Lagus. Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 78-86, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[7] Oskar Kohonen, Sami Virpioja, Laura Leppänen, and Krista Lagus. Semi-supervised extensions to Morfessor Baseline. In Mikko Kurimo, Sami Virpioja, and Ville T. Turunen, editors, *Proceedings of the Morpho Challenge 2010 Workshop*, pages 30–34, Espoo, Finland, September 2010. Aalto University School of Science and Technology, Department of Information and Computer Science. Technical Report TKK-ICS-R37. Extended abstract.

[8] Sami Virpioja, Oskar Kohonen, and Krista Lagus. Evaluating the effect of word frequencies in a probabilistic generative model of morphology. In Bolette Sandford Pedersen, Gunta NeÅ¡pore, and Inguna Skadina, editors, *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, volume 11 of NEALT Proceedings Series, pages 230-237. Northern European Association for Language Technology, Riga, Latvia, May 2011.

[9] Sami Virpioja, Minna Lehtonen, Annika Hultén, Riitta Salmelin, and Krista Lagus. Predicting reaction times in word recognition by unsupervised learning of morphology. In Timo Honkela, Wlodzislaw Duch, Mark Girolami, and Samuel Kaski, editors, *Artificial Neural Networks and Machine Learning — ICANN 2011*, volume 6791 of Lecture Notes in Computer Science, pages 275–282. Springer Berlin / Heidelberg, June 2011.

10.3 Keyphrase extraction

A language-independent keyphrase extraction method, *Likey*, extracts keyphrases from a document using phrase frequency ranks and comparison to a reference corpus. It has a light-weight preprocessing phase, whereas most of the other methods for keyphrase extraction are highly dependent on the language used and the need for preprocessing is extensive. Many of them need also a training corpus. On the contrary, *Likey* enables independence from the language used. It is possible to extract keyphrases from text in previously unknown language, provided that a suitable reference corpus is available. The method was further developed and applied for a set of scientific articles [1]. The evaluation was conducted against both author-provided and manually extracted keyphrases in the articles.

Learning taxonomic relations

As an application for the *Likey* keyphrase extraction method, a method for learning taxonomic relations from a set of text documents was developed [2]. *Likey* and two other methods for feature extraction were used to create document vectors for Wikipedia articles about animals in English and Finnish. The vectors were clustered hierarchically using the Self-Organizing Map (SOM). The resulting taxonomy were compared to a scientific classification of the animals, that can be seen in Figure 10.3.

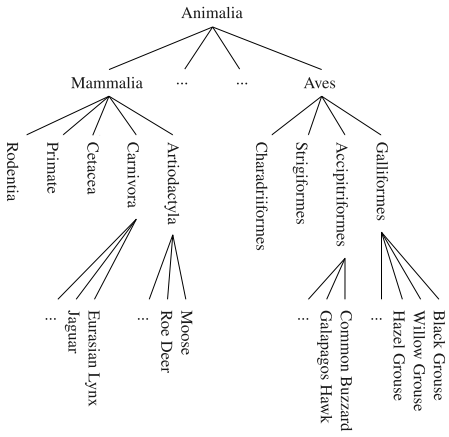


Figure 10.3: Part of the reference taxonomy.

References

[1] Mari-Sanna Paukkeri and Timo Honkela (2010) Likey: Unsupervised Language-Independent Keyphrase Extraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*. Association for Computational Linguistics. Uppsala, Sweden, July 2010.

[2] Mari-Sanna Paukkeri, Alberto Pérez García-Plaza, Sini Pessala, and Timo Honkela (2010) Learning taxonomic relations from a set of text documents. In *Proceedings of 5th International Symposium Advances in Artificial Intelligence and Applications (AAIA'10)*. Wisla, Poland, October 2010.

10.4 Vector space models of language

Vector space models are a standard way to represent documents or words as vectors of features. The model provides a solution to the problem of representing symbolic information (words) in numerical form for computational processing. In a vector space, similar items are close to each other, and the closeness can be measured using vector similarity measures. Vector space models are applied, for example, in various information retrieval tasks and text categorization tasks.

Dimensionality reduction in document clustering

In document clustering, semantically similar documents are grouped together. The dimensionality of document collections is often very large, thousands or tens of thousands of terms. Thus, it is common to reduce the original dimensionality before clustering. Cosine distance is widely seen as the best choice for measuring the distances between documents in k-means clustering. The effect of dimensionality reduction on different distance measures in document clustering was analysed in [1]. The results show that after dimensionality reduction into small target dimensionalities, such as 10 or below, the superiority of cosine measure does not hold. Also, for small dimensionalities, PCA dimensionality reduction method performs better than SVD. Further, the effect of l_2 normalization for different distance measures was studied. The experiments are run for three document sets in English and one in Hindi.

Analysis of adjectives in a word vector space

Large number of studies indicate that methods using co-occurrence data provide useful information on the relationships between the words, as words with similar or related meaning will tend to occur in similar contexts. This intuition has been carefully assessed, in particular, for nouns and verbs. In [2], we study how well the co-occurrence statistics provide a basis for automatically creating a representation for a group of adjectives as well. In this study, a the text collection used was extracted from English Wikipedia, and the evaluation was carried out with 72 adjectives which formed 36 antonym pairs (i.e. good-bad). Further, we compare three dimension reduction methods and their effect on the quality of final representation: The Principal Component Analysis, the Self-Organizing Map and Neighbor Retrieval Visualizer (NeRV). Figure 10.4 visualizes the adjectives and their neighbors after dimension reduction with the NeRV.

Vector space evaluation using CCA

The vector spaces are generated using different feature extraction methods for text data. However, evaluation of the feature extraction methods may be difficult. Indirect evaluation in an application is often time-consuming and the results may not generalize to other applications, whereas direct evaluations that measure the amount of captured semantic information usually require human evaluators or annotated data sets.

We have developed a novel direct evaluation method for vector space models of documents based on canonical correlation analysis (CCA) [3]. The evaluation method is based on

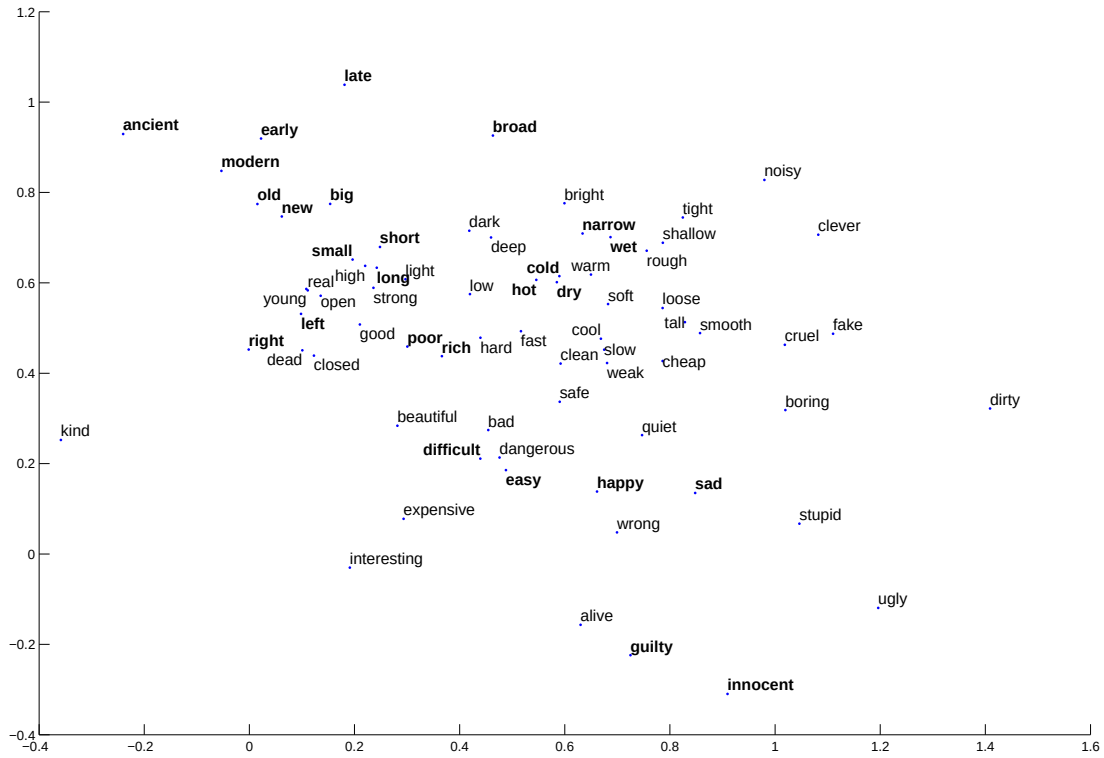


Figure 10.4: The set of adjectives used in the study projected into a 2-dimensional space using the Neighbor Retrieval Visualizer (NeRV) method. The words in bold have the antonym in their local neighborhood.

unsupervised learning, it is language and domain independent, and it does not require additional resources besides a parallel corpus.

CCA is a classical method for finding linear relationship between two data sets. In our setting, the two sets are parallel text documents in two languages. A good feature extraction method should provide representations that reflect the semantic contents of the documents. We assume that the underlying semantic contents is independent of the language, illustrated by the generation model on the left part of Figure 10.5. Then we can study which feature extraction methods capture the contents best by measuring the dependence between the representations of a document and its translation, illustrated on the right part of Figure 10.5.

In the case of CCA, the applied measure of dependence is correlation, which means that it can only find linear dependence. In a related study [4], we have shown that kernelized version of CCA outperforms linear CCA in a sentence matching task. Unfortunately, choosing the kernel and its parameters would require additional optimization step and held-out data for vector space evaluation.

We have demonstrated the proposed evaluation method on a sentence-aligned parallel corpus. The method was validated in three ways: (1) showing that the obtained results with bag-of-words representations are intuitive and agree well with the previous findings, (2) examining the performance of the proposed evaluation method with indirect evaluation methods in simple sentence matching tasks, and (3) in a quantitative manual evaluation of word translations. The results of the evaluation method correlate well with the results

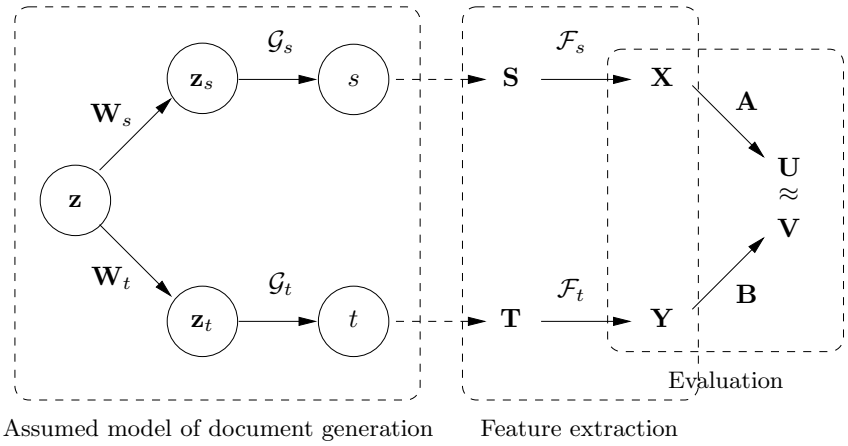


Figure 10.5: On the left: Assumed model for generation of documents s and t. Vector $\mathbf{z}$ in the language-independent semantic space $\mathcal{Z}$ is projected onto vectors $\mathbf{z}_s$ and $\mathbf{z}_t$ in the language-specific subspaces $\mathcal{Z}_s$ and $\mathcal{Z}_t$. Processes $\mathcal{G}_s$ and $\mathcal{G}_t$ generate document pairs from the respective subspaces. On the right: The process of evaluating feature extraction method $\mathcal{F}$ with CCA. The aligned document collections $\mathbf{S}$ and $\mathbf{T}$ are reduced to matrices $\mathbf{X}$ and $\mathbf{Y}$ of feature vectors using $\mathcal{F}$. Then $\mathbf{X}$ and $\mathbf{Y}$ are projected onto a common vector space using CCA.

of the indirect and manual evaluations.

References

[1] Mari-Sanna Paukkeri, Ilkka Kivimäki, Santosh Tirunagari, Erkki Oja, and Timo Honkela (2011) Effect of Dimensionality Reduction on Different Distance Measures in Document Clustering. In B.-L. Lu, L. Zhang, and J. Kwok (Eds.): ICONIP 2011, Part III, LNCS 7064, pp. 167-176. Springer-Verlag Berlin Heidelberg.

[2] Timo Honkela and Tiina Lindh-Knuutila and Krista Lagus (2010) Measuring Adjective Spaces. In K. Diamantaras, W. Duch, L. S. Iliadis (Eds.): Proceedings of ICANN 2010, Artificial Neural Networks, pp. 368-373. Springer Verlag Berlin Heidelberg.

[3] Sami Virpioja, Mari-Sanna Paukkeri, Abhishek Tripathi, Tiina Lindh-Knuutila, and Krista Lagus. Evaluating vector space models with canonical correlation analysis. *Natural Language Engineering*, to appear. Available on CJO 2011.

[4] Abhishek Tripathi, Arto Klami, and Sami Virpioja. Bilingual sentence matching using kernel CCA. In *Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010)*, pages 130–135, Kittilä, Finland, August 2010. IEEE.

Computational Cognitive Systems

Chapter 11

Computational Cognitive Systems

Timo Honkela, Krista Lagus, Marcus Dobrinkat, Oskar Kohonen, Mikaela Kumlander, Tiina Lindh-Knuutila, Ilari Nieminen, Mari-Sanna Paukkeri, Matti Pöllä, Juha Rautio, Sami Virpioja, Jaakko Väyrynen, Paul Wagner, Eric Malmi, Tero Tapiovaara, Tommi Vatanen, Ilkka Kivimäki, Laura Leppänen, Sini Pessala, Santosh Tirunagari

11.1 Introduction

Computational Cognitive Systems group conducts research on artificial systems that combine perception, action, reasoning, learning and communication. This area of research draws upon biological, cognitive and social system approaches to understanding cognition. Cognitive systems research is multidisciplinary and interdisciplinary. It benefits from sharing and leveraging expertise and resources between disciplines. Methodologically, statistical machine learning, pattern recognition and signal processing are central tools within computational cognitive systems research. Our research focuses on modeling and applying methods of unsupervised and semisupervised learning in the areas of conceptual modeling, multilingual and language independent language technology, and socio-cognitive modeling. Results related to language processing are reported in Section 10.

We approach conceptual modeling as a dynamic phenomenon. Among humans, conceptual processing takes place as an individual and social process. We attempt to model this dynamic and constructive aspect of conceptual modeling by using statistical machine learning methods. We also wish to respect the overall complexity of the theme, for instance, not relying on explicit symbolic representations are the only means relevant in conceptual modeling. Our machine translation research builds on the conceptual modeling research as well as on the research on adaptive language technology.

Socio-cognitive modeling is our newest research area which builds on 1) the experience and expertise in modeling complex phenomena related to language learning and use at cognitive and social levels and 2) strong national and international collaboration especially with the representatives of social sciences and humanities. Socio-cognitive modeling mainly merges aspects of computer science, social sciences and cognitive science. The basic idea is to model interlinked social and cognitive phenomena.

Summary of collaboration

We have worked in close collaboration with other groups in Adaptive Informatics Research Centre, lead by Prof. Erkki Oja and Prof. Samuel Kaski, in particular natural language processing and multimodal interfaces (Dr. Mikko Kurimo and Dr. Jorma Laaksonen).

The collaboration with *Aalto School of Economics* and *National Consumer Research Centre* that started in KULTA projects, has continued within Tekes-funded VirtualCoach project. The project focuses on wellbeing informatics and is discussed below in more detail.

In the area of multilingual language technology, META-NET Network of Excellence is a major effor (http://www.meta-net.eu). One objective is to build bridges to neighbouring technology fields such as machine learning and cognitive systems. The research agenda consist of four areas: (1) bringing more semantics into Machine Translation, (2) optimising the division of labour in hybrid Machine Translation, (3) exploiting the context for Translation, and (4) preparing a base for Machine Translation. The COG group has been actively involved in the third area. In overall, META-NET consists of 57 research centres from 33 countries. META-NET is coordinated by the German Research Center for Artificial Intelligence (DFKI). A Cognitive Systems blog description of a META-NET event, written by Jaakko Väyrynen is shown in Fig. 11.1 (http://cogsys.blogspot.com/2011/06/meta-forum-2011.html).

META-FORUM 2011

The META-FORUM 2011 is held June 27/28 in Budapest, Hungary. It is organized by META-NET, a Network of Excellence consisting of 47 research centers from 31 countries, which is dedicated to building the technological foundations of a multilingual European information society. The forum consists of keynote talks, several sessions, two panels and joined exhibition by industry and academia.

Thomas Hofmann from Google Research gave a keynote lecture on how they are bringing semantics to the web. He explained how interesting models can be scaled to web size by using curated knowledge bases (eg Freebase) and the output from accurate but slower parsers as the either the starting point to statistical analysis or training data for faster systems.

Bran Boguraev from IBM's TJ Watson Research Center talked about evidence based reasoning over natural language content. The example was of course their massively parallel Watson architecture for large-scale hypothesis generation, validation and scoring which tackles the long-standing challenge in AI to emulate human expertise.

Hans Uszkoreit introduced the outline for META-NET's strategic research agenda for language technology in Europe. The vision paper presents possible futures for innovative language technology applications in the following areas:

- language-transparent web and media experience,
- natural and inclusive interaction, and
- efficient information management.

Posted by Jaakko Väyrynen at 3:37 PM

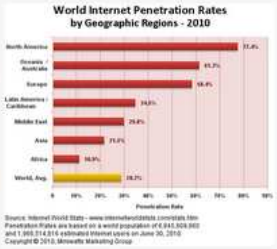
Figure 11.1: META-FORUM 2011 description in the widely read Cognitive Systems blog.

The COG group has actively participated the MultilingualWeb initiative that is concerned with standards and best practices that support the creation, localization and use of multilingual web-based information. The consortium is coordinated by W3C and includes companies such as Microsoft, Facebook, Opera and SAP. Fig. 11.2 gives an excerpt of a blog post to Cognitive Systems on a MultilingualWeb event, written by Matti Pöllaä (see http://cogsys.blogspot.com/2011/04/content-on-multilingual-web.html for more details).

Wednesday, April 20, 2011

Content on the Multilingual Web

The Internet world stats site is a fascinating source of information for predicting the trends of web usage in terms of language and locality. These statistics were frequently cited in Pisa, where the participants of the W3C coordinated MultilingualWeb project held its second workshop under the title *multilingual content in the web*.



To fully appreciate the demand for new developments in multilingual web technologies one only needs to look at the statistics of Internet penetration rates by geographic regions (see figure). In 2010, already **42% of Internet users are from Asia while the Internet penetration rate has only reached 21.5%**. Considering that the basic building blocks of web technology (browsers, content markup syntax) were originally developed for English, it is obvious that there is much work to be done in developing web standards that enable language interoperability from the nuts-and-bolts level of character encoding all the way to statistical machine translation.

While the MultilingualWeb project in itself is a small one, the European Commission has substantial investments in developing multilingual technologies in the near future. In specific, the project officer **Kimmo Rossi** counted that in the year 2015 there will be a total of 50 ongoing projects in the area of multilingual technologies. Recently, the EC conducted a survey with Eurobarometer to study how Europeans use languages online. One of the findings was that 44% of people feel that they miss important information online because they don't understand the content language. Stay tuned for the full report of the survey on Eurobarometer's site.

Figure 11.2: A description of a MultilingualWeb event.

MultilingualWeb has been organizing workshops open to the public and various communication channels, spreading information about what standards and best practices currently exist, and what gaps need to be filled. Fig. 11.3 shows a fragment of the result a thematic session. Its focus was "Semantic resources and machine learning for quality, efficiency and personalisation of accessing relevant information over language borders".

Figure 11.3: A fragment of a result from a MultilingualWeb thematic session.

An excellent example of the positive effects of international researcher exchange are the results that stem from the visit by young COG researchers, Tommi Vatanen and Eric Malmi, to CERN in Switzerland [5, 5].

In EIT ICT Labs, Dr. Krista Lagus has served as the Lead of Schools & Camps Catalyst Development. EIT ICT Labs is one of the first three Knowledge and Innovation Communities (KICs) selected by the European Institute of Innovation & Technology (EIT) to accelerate innovation in Europe. EIT aims to rapidly emerge as a key driver of EU's sustainable growth and competitiveness through the stimulation of world-leading innovation (http://eit.ictlabs.eu/).

Scientific events

In collaboration with other groups in AIRC, the COG group has been active in organizing national and international conferences. Two main events took place in 2011, International Conference on Artificial Neural Networks [49, 50] and Workshop on Self-Organizing Maps [3].

During ICANN 2011, META-NET workshop on Context in Machine Translation was organized to foster exchange of ideas and results in this area. The notion of context was meant to be understood broadly, including other modalities (like vision) in addition to the textual contexts.

The Context in Machine Translation Challenge is part of a series of challenges organized by the META-NET Network of Excellence (http://www.meta-net.eu), jointly by Aalto University (Finland), CNRS/LIMSI (France) and ILSP (Greece), supported by other network partners.

The COG group was also involved in organizing the Finnish Artificial Intelligence Confer-

ence, STeP 2011 [4].

VirtualCoach project

VirtualCoach – Paths of Wellbeing is a major project in the area of wellbeing informatics, lead by Dr. Krista Lagus (http://blog.pathsofwellbeing.com/). The VirtualCoach builds on the traditional methodological strengths of the COG group and AIRC, in general. Wellbeing informatics is an emerging area of research in which ICT methodologies are used to measure, analyze, and promote wellbeing of individuals. Examples of traditional applications include heart rate monitoring, tracking sports activities, analyzing the nutritional content of diets, and analyzing sleeping patterns with mobile technologies.

In the VirtualCoach project, a central topic is how to help people to find peer and professional support in a personalized manner. One approach is to build social media applications in which users can find stories that are potentially helpful in their individual life situations. The users may wish to develop their wellbeing further, or need to solve some problem that prevents them from achieving a satisfactory level of wellbeing.

The VirtualCoach project is a collaborative research effort with the National Consumer Research Center and several companies including createAmove, FlowDrinks, Futuria Consulting, If insurance company, Innotiimi, mutual pension insurance company Varma, MTV Media, Oppifi, Terveystalo Healthcare, and Vierumäki Sports Institute.

References

[1] Timo Honkela, Wlodzislaw Duch, Mark A. Girolami, and Samuel Kaski, editors (2011). Artificial Neural Networks and Machine Learning - Proceedings of ICANN 2011 - 21st International Conference on Artificial Neural Networks, Parts I and II. Springer.

[2] Mikael Kuusela, Eric Malmi, Risto Orava, and Tommi Vatanen (2011). Soft classification of diffractive interactions at the LHC. AIP Conference Proceedings, 1350(1):111-114.

[3] Jorma Laaksonen and Timo Honkela, editors (2011). Advances in Self-Organizing Maps - Proceedings of WSOM 2011, 8th International Workshop. Springer.

[4] Tapio Pahikkala, Jaakko Väyrynen, Jukka Kortela, and Antti Airola, editors (2010). Proceedings of the 14th Finnish Artificial Intelligence Conference, STeP 2010, Finnish Artificial Intelligence Society.

[5] Tommi Vatanen, Mikael Kuusela, Eric Malmi, Tapani Raiko, Timo Aaltonen, and Yoshikazu Nagai (2011). Fixed-background EM algorithm for semi-supervised anomaly detection. Technical report, Aalto University School of Science.

11.2 Learning to translate

Our research on multilinguality and machine translation (MT) uses novel methods that are based on adaptivity. An MT system is *learning to translate* rather than needs to be programmed to do so. The advances in statistical machine translation have shown that the adaptive paradigm can help in reducing the system development costs dramatically. However, these systems rely on representations that do not capture many relevant linguistic aspects, neither take into account the wealth of knowledge that is known about human cognitive processes related to natural language understanding, translation and interpretation.

An important context for the research and development work on multilinguality is META-NET. META-NET, a Network of Excellence consisting of 57 research centres from 33 countries, is dedicated to building the technological foundations of a multilingual European information society [1]. The research work in META-NET has been divided into four work packages. Our activities have focused on exploiting context in machine translation. During ICANN 2011 conference, a META-NET workshop on "Context in Machine Translation" was organized[2]. The objective was to foster exchange of ideas and results in this area. Here the notion of context is meant to be understood broadly, including other modalities (like vision) in addition to the textual contexts. An invited talk was given by Dr. Katerina Pastra entitled "Bridging language, action and perception: the cognitive context of machine translation". During the workshop a challenge on context in mt was announced. The challenge data set consists of documents from the JRC-ACQUIS Multilingual Parallel Corpus. Two language-pair directions are included in the data, English-¿Finnish and Greek-French. The constructed challenge training data set contains the document context, n-best lists for translated documents and additional contextual information as well as the reference translations.

Language identification of short text segments

For processing multilingual texts, it is important to identify the language of each document, sentence, or even word. There are many accurate methods for language identification of long text samples, but identification of very short strings still presents a challenge. In [1], we consider test samples that have only 5–21 characters. We show that a simple but efficient method, naive Bayes classifier based on character n-gram models, outperforms previous methods, when state-of-the-art language modeling techniques from automatic speech recognition research are applied. Using the Universal Declaration of Human Rights as a data set, we were able to conduct the experiments with as many as 281 languages.

Automatic machine translation evaluation

The normalized compression distance (NCD) has been further investigated as an automatic machine translation metric. It is based on character-sequence comparison of translated text and a reference translation, whereas most typical metrics (e.g. BLEU and NIST) operate on word-sequences. In [2], the NCD metric has been extended to include flexible word matching, which extends the references translations with synonyms for words.

[1] http://www.meta-net.eu/
[2] http://www.cis.hut.fi/icann11/con-txt-mt11/

Several possible extensions were tested in [3] in order to improve the evaluation metric, including multiple reference handling and segment replication. The metric also participated in the MetricsMATR 2010 machine translation evaluation shared task. MT evaluation metrics are themselves evaluated by measuring correlation between the automatic metric and known human evaluations of translations.

Automatic evaluation of machine translation (MT) systems requires automated procedures to ensure consistency and efficient handling of large amounts of data, and are essential for parameter optimization and statistical machine translation system development. In contrast to most MT evaluation measures, e.g., BLEU and METEOR, NCD provides a general information theoretic measure of string similarity.

NCD is an approximation of the uncomputable normalized information distance (NID), a general measure for the similarity of two objects. NID is based on the notion of Kolmogorov complexity, a theoretical measure for the information content of a string, defined as the shortest universal Turing machine that prints the string and stops. NCD approximates NID by the use of a compressor that is an upper bound of the Kolmogorov complexity.

Similar to the mBLEU extension of the BLEU metric, the same synonym handling module from METEOR was incorporated into the NCD metric. In our experiments, the resulting mNCD metric had consistently better correlation with human judgments of translation compared to the basic NCD metric.

An NCD-based metric was developed to handle multiple references in evaluation. It can be viewed as a generalization of the NCD metric, as they are equal with only one reference translations. It was shown to work better when two reference translations are available.

Domain adaptation for statistical machine translation

Statistical machine translation methodology is highly dependent of relevant parallel texts for training. However, available large parallel corpora are typically out-of-domain for many interesting translation tasks, such as news translation. This is especially true for less-resourced languages. Therefore methods that can utilize out-of-domain text

Four existing different domain adaptation methods were tested in [4]: language model (LM) adaptation, translation model (TM) adaptation, automatic post-editing and retraining with combined data. All tested methods except language model adaptation outperformed the baseline system trained with only the out-of domain data. The experiment were conducted with a larger out-of-domain Europarl parallel corpus and a small previously collected small corpus of Finnish Iltalehti news with their English translations.

Domain adaption can be accomplished at several locations in the statistical machine translation process. The simplest way is simply to pool all available data and to learn a single model based on it. It may not feasible if only models are available or the models are incompatible. Also, it may give too little emphasis on the small amount of in-domain data. *Language model adaptation* requires only monolingual target language data and affects only the selection of translations without providing any new translations possibilities for words or phrases. The adaptation can be done with a combination of the data or linear or log-linear interpolation of two or more language models. *Translation model adaptation* requires additional parallel data that can be either included in the existing data or the models can be joined with log-linear interpolation. The process is illustrated in

Figure 11.4. The *post-edit domain-adaption*, shown in Figure 11.5, learns another translation model from the output of the original translation system to correct or corrected translations, trying to statistically fix mistakes made by the original system.

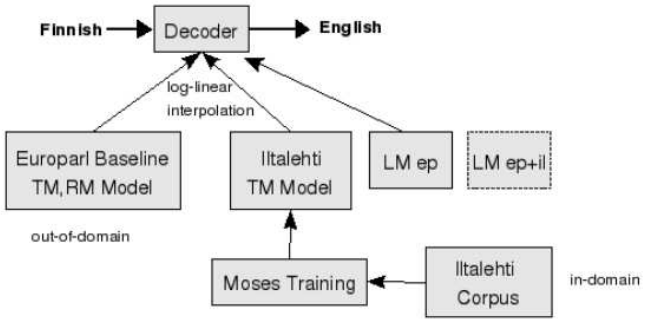


Figure 11.4: The process for domain adaptation with log-linear interpolation of baseline and in-domain translation models (TM). The wwo models operate on parallel inside one translation system.

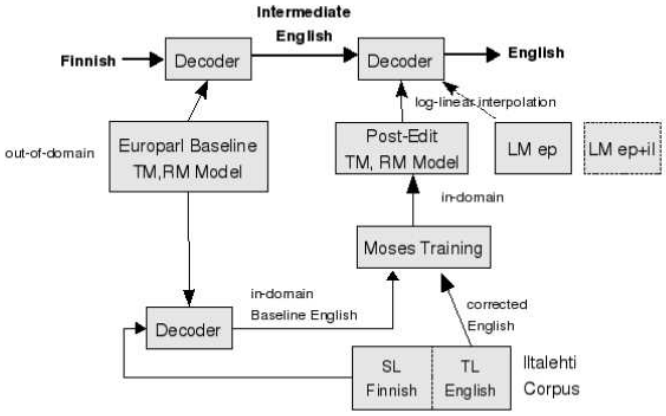


Figure 11.5: The process for post-edit domain adaptation. The two translation systems operate in sequence.

References

[1] Tommi Vatanen, Jaakko J. Väyrynen, and Sami Virpioja. Language identification of short text segments with n-gram models. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[2] Marcus Dobrinkat, Tero Tapiovaara, Jaakko Väyrynen, and Kimmo Kettunen. Evaluating machine translations using mNCD. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 80–85. Association for Computational Linguistics, 2010.

[3] Marcus Dobrinkat, Tero Tapiovaara, Jaakko Väyrynen, and Kimmo Kettunen. Normalized compression distance based measures for MetricsMATR 2010. In *Proceedings*

of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 343–348. Association for Computational Linguistics, 2010.

[4] Marcus Dobrinkat and Jaakko J. Väyrynen. Experiments with domain adaptation methods for statistical MT: From european parliament proceedings to finnish newspaper text. In *Proceedings of the 14th Finnish Artificial Intelligence Conference STeP 2010*, number 25 in Publications of the Finnish Artificial Intelligence Society, pages 31–38. Finnish Artificial Intelligence Society, 2010.

11.3 Socio-cognitive modeling

Socio-cognitive modeling is a new research area that merges aspects of computer science, social sciences and cognitive science. The basic idea is to model interlinked social and cognitive phenomena. Our focus has traditionally been in modeling individual cognition that learns and uses language, or in building models of language using statistical machine learning methods. Already for a long time, we have been interested in language and its use as a dynamic phenomenon rather than as a static structural object. Thereafter, we have widened our interest to language as a socio-cultural phenomenon that encodes human knowing and further to other socio-cognitive phenomena, however often related to language. In other words, cognition and intelligent activity are not only individual processes but ones which rely on socio-culturally developed cognitive tools. These include physical and conceptual artifacts as well as socially distributed and shared processes of intelligent activity embedded in complex social and cultural environments.

One approach in socio-cognitive modeling is social simulation. It aims at exploring and understanding of social processes by means of computer simulation. Social simulation methods can be used to to support the objective of building a bridge between the qualitative and descriptive approaches used in the social sciences and the quantitative and formal approaches used in the natural sciences. Collections of agents and their interactions are simulated as complex non-linear systems, which are difficult to study in closed form with classical mathematical equation-based models. Social simulation research builds on the distributed AI and multi-agent system research with a specific interest of linking the two areas.

Directions for e-science and science 2.0

Science 2.0 builds on the technologies of Web 2.0. Blogs, wikis and other social sharing and interaction tools allow scientists to interact and make their data and interpretations available for others in novel ways [1]. Science 2.0 continues and extends the tradition of publishing open source software and open access publishing of scientific articles. Local examples, stemming from the research of AIRC and its predecessors, include *SOM Toolbox* for Matlab (`http://www.cis.hut.fi/somtoolbox/`), *FastICA* for Matlab (`http://www.cis.hut.fi/projects/ica/fastica/`), *dredviz* software package for dimensionality reduction in information visualization (`http://www.cis.hut.fi/projects/mi/software/dredviz/`), and *Morfessor* software (`http://www.cis.hut.fi/projects/morpho/`).

Where Science 2.0 refers to new practices in conducting science with the help of communications and collaborations technologies, computational science builds on modeling and simulation of real world or anticipated phenomena based on massive data sets. Traditionally, this field has been dominated by applications related to natural sciences and engineering, but also human and social sciences have started to use computational models as a research tool.

An article, based on a keynote given by Timo Honkela in MASHS 2010 conference, discusses the themes described above and considers computational linguistics and computational economics as more specific examples [1]. Also a map of Finnish science was described and discussed (see Fig. 11.6). The map was created based on the contents of 3,224 application documents sent to Academy of Finland. The text contents were analyzed using an automatic terminology extraction method called Likey (see Section 10.1 and the section

on keyphrase extraction for more detauls). The SOM algorithm organized the documents into a map in which similar applications are close to each other and in which thematic areas emerged (Fig. 11.6).

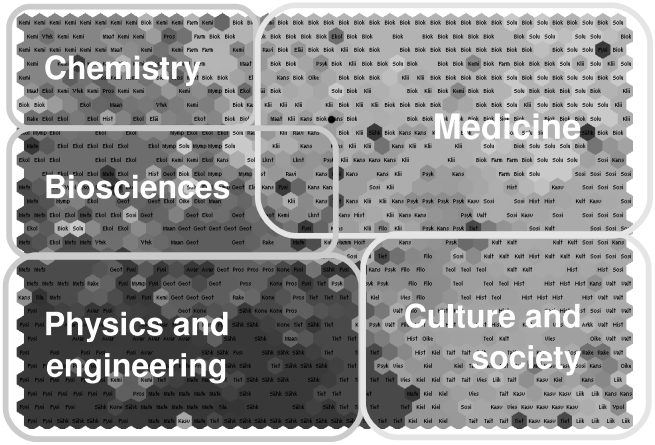


Figure 11.6: Map of Finnish science.

Text mining and qualitative analysis of history interviews

In collaboration with Dr. Petri Paju, we have explored the possibility of applying a text mining methods on a large qualitative source material concerning the history of information technology [3]. This data was collected in the Swedish documentation project "From Computing Machines to IT." We applied text mining on the interview transcripts of this Swedish documentation project. Specifically, we seeked to group the interviews according to their central themes and affinities and pinpoint the most relevant interviews for specific research questions. In addition, we searched for interpersonal links between the interviews. We applied the SOM algorithm to create a similarity diagram of the interviews. In the article based on this research, we discussed the results in several contexts including the possible future uses of text mining in researching history [3].

Analysis of global nutrition, lifestyle and health situation

EIT ICT Labs Wellbeing Innovation Camp (WIC) 2010 lead into a number of results (see http://www.cis.hut.fi/wicamp/2010/). In two cases, research work leading to publications in the general area of wellbeing informatics [2, 4]. The Action Line "Health and Wellbeing" in EIT ICT Labs has very similar basic objectives as the VirtualCoach project, discussed earlier in this report (see Section 11.1). In this action line, it is acknowledged that health and wellbeing "needs to be approached in a holistic way that fosters mental and physical fitness and balance. Having healthy and caring relationships, as well as good daily habits and behavioural patterns, are just two of the many principles in this holistic approach." (http://eit.ictlabs.eu/action-lines/health-wellbeing/)

In the first Wellbeing Innovation Camp project, the relationship between nutrition, lifestyle and health situation around the world was studied. The dataset used in the analysis is comprised of statistics that can be divided into three categories namely health, diet and lifestyle. The first category contains information such as obesity prevalence, incidence

of tuberculosis, mortality rates and related variables in different countries. The dietary information includes consumption of proteins, sugar and milk products, and various other components of nutrition. The lifestyle category provides information related to the drinking and smoking habits, etc. One interesting finding was that there is a clear correlation between high consumption of sugar or sweeteners and a prevalence of cholesterol in men and women. This and other conclusions from the study are reported in [2].

Emotional semantics of abstract art and low-level image features

The second result of a project initiated at the Wellbeing Innovation Camp 2010, is an analysis on how well low-level image features can be used in predicting the emotional responses of subjects to abstract art [4]. This research was conducted in close collaboration with Jorma Laaksonen's research group. In this work, we studied people's emotions evoked by viewing abstract art images based on traditional low-level image features within a binary classification framework. Abstract art is used instead of artistic or photographic images because those contain contextual information that influences the emotional assessment in a highly individual manner. Whether an image of a cat or a mountain elicits a negative or positive response is subjective. After discussing challenges concerning image emotional semantics research, we empirically demonstrated that the emotions triggered by viewing abstract art images can be predicted with reasonable accuracy by machine using a variety of low-level image descriptors such as color, shape, and texture [4].

References

[1] Timo Honkela. Directions for e-science and science 2.0 in human and social sciences (2010). In Proceedings of MASHS 2010, Computational Methods for Modeling and Learning in Social and Human Sciences, pages 119–134. Multiprint.

[2] Yasir Mehmood, Mudassar Abbas, Xi Chen, and Timo Honkela (2011). Self-organizing maps of nutrition, lifestyle and health situation in the world. In Advances in Self-Organizing Maps - Proceedings of WSOM 2011, 8th International Workshop, pages 160–167. Springer.

[3] Petri Paju, Eric Malmi, and Timo Honkela (2011). Text mining and qualitative analysis of an it history interview collection. In History of Nordic Computing, IFIP Publications, pages 433–443. Springer.

[4] He Zhang, Eimontas Augilius, Timo Honkela, Jorma Laaksonen, Hannes Gamper, and Henok Alene (2011). Analyzing emotional semantics of abstract art using low-level image features. In Joao Gama, Elizabeth Bradley, and Jaakko Hollmén, editors, Advances in Intelligent Data Analysis X, volume 7014 of Lecture Notes in Computer Science, pages 413–423, Berlin/Heidelberg, Springer.

11.4 GICA: Grounded Intersubjective Concept Analysis

We have introduced a novel method to analyze and make visible differences among people regarding how they conceptualize the world [1, 3]. The Grounded Intersubjective Concept Analysis (GICA) method first employs either a conceptual survey or a text mining step to elicit particular ways in which terms and associated concepts are used among individuals. The subsequent analysis and visualization reveals potential underlying groupings of people, objects and contexts. The GICA method extends the basic idea of the traditional term-document matrix analysis to include a third dimension of different individuals. This leads to a formation of a third-order tensor of Subjects x Objects x Contexts. Through flattening, these Subject-Object-Context (SOC) tensors can be analyzed using various computational methods. In the following, we introduce the GICA method and its background in some detail.

Introduction to epistemological subjectivity

When human communication as well as computational modeling of knowledge and language is considered, it is usually taken granted that the meaning of all symbols used in the communication or representation of knowledge is shared by all human and/or artificial agents. It is, however, quite straightforward to show empirically that this is not the case. Practical and theoretical limitations of traditional knowledge representation were already highlighted in an early project the objective of which was to developed a natural language database interface [2]. It started to be obvious that meaning needs to be defined contextually and also the subjective aspects is relevant. The word 'red' has different interpretations in different contexts such as "red shirt", "red skin", or "red wine". Subjectivity is particularly notable when abstract complex concepts such as 'computation', 'democracy', or 'sustainability' are considered.

In human communication, it is the occasional clear failure that allows us to see that understanding language is often difficult. In making the connection between a word and its typical and appropriate use, we humans rely on a long learning process. The process is made possible and guided by our genetic make-up, but its success essentially requires extensive immersion to a culture and contexts of using words and expressions. To the extent that these contexts are shared among individual language speakers, we are then able to understand each other. When our learning contexts differ, however, differences in understanding the concepts themselves arise and subsequent communication failures begin to take place. Two main failure types can be detected. The first type is false agreement, where on the surface it looks as if we agree, but in fact our conceptual difference hides the underlying difference in opinions or world views. The second type of problem caused by undiscovered meaning differences is false disagreement. If we are raised (linguistically speaking) in different sub-cultures, we might come to share ideas and views, but might have learned to use different expressions to describe them.

It is commonplace in linguistics to define semantics as dealing with prototypical meanings whereas pragmatics would be associated with meanings in context. For our purposes, this distinction is not relevant since interpretation of natural language expressions always takes place in some context, usually even within multiple levels of context including both linguistic and extra-linguistic ones. In the contrary case, that is, when an ambiguous word such as "break" appears alone without any specific context one can only try to guess which

of its multiple meanings could be in question. If there is even a short contextual cue — "break the law", or "have a break", or "how to break in billiards" – it is usually possible to arrive at a more accurate interpretation. Also the extralinguistic context of an expression usually helps in disambiguation.

Becoming conscious of individual differences as a way of increasing understanding

For the most part, people do not seem to be aware of the subjectivity of their perceptions, concepts, or world views. Furthermore, one might claim that we are more typically conscious of differences in opinions, whereas differences in perception or in conceptual level are less well understood. It is even possible that to be able to function efficiently it is best to mostly assume that my tools of communication are shared by people around me. However, there are situations where this assumption breaks to a degree that merits further attention. An example is the case when speakers of the same language from several disciplines, interest groups, or several otherwise closely knit cultural contexts come together to deliberate on some shared issues.

The background assumption of the GICA method innovation is the recognition that although different people may use the same word for some phenomenon, this does not necessarily mean that the conceptualization underlying this word usage is the same; in fact, the sameness at the level of names may hide significant differences at the level of concepts. Furthermore, there may be differences at many levels: experiences, values, understanding of the causal relationships, opinions and regarding the meanings of words. The differences in meanings of words are the most deceptive, because to discuss any of the other differences, a shared vocabulary which is understood in roughly the same way, is necessary. Often a difference in the meanings of used words remains unrecognized for a long time; it may, for instance, be misconstrued as a difference in opinions. Alternatively, a difference in opinions, or regarding a decision that the group makes, may be masked and remain unrecognized, because the same words are used seemingly in accord, but in fact in different meanings by different people. When these differences are not recognized during communication, it often leads to discord and unhappiness about the end result. As a result, the joint process may be considered to have failed in one or even all of its objectives.

Making differences in understanding visible

Our aim with the Grounded Intersubjective Concept Analysis (GICA) method is to devise a way in which differences in conceptualization such as described above can be made visible and integrated into complex communication and decision making processes. An attempt to describe the meaning of one word by relying on other words often fails, because the descriptive words themselves are understood differently across the domains. In fact, a domain may have a large number of words that have their specialized meanings. The more specific aims of this paper are to define the problem domain, to explain the processes of concept formation from a cognitive point of view based on our modeling standpoint, and to propose a methodology that can be used for making differences in conceptual models visible in a way that forms a basis for mutual understanding when different heterogeneous groups interact. Contexts of application are, for instance, public planning processes, environmental problem solving, interdisciplinary research projects,

product development processes, and mergers of organizations.

11.5 The GICA method

In the following, we present an overview of the GICA method based on conducting a conceptual survey among participants (see [1] for more details. A version in which text mining can be used to extract subjective information has been introduced recently [3].

The GICA method includes three main stages:

A Preparation,

B Focus session(s), and

C Knowledge to action activities.

These steps can be repeated iteratively. The focus sessions are supported with computational tools that enable the analysis and visualization of similarities and differences in the underlying conceptual systems.

Subjectivity cube

In the GICA method, the idea of considering some items or objects such as words in their contexts is taken a step further. As we have in the introductory section of this paper aimed to carefully show, subjectivity is an inherent aspect of interpretation. In order to capture the aspect of subjectiveness, we add a third dimension to the analysis. Namely, we extend the set of observations, $objects \times contexts$, into $objects \times contexts \times subjects$, i.e. we additionally consider what is the contribution of each subject in the context analysis.

Adopting the notation and terminology for *tensors* (multiway arrays), the *order of a tensor* is the number of the array dimensions, also known as ways or modes. As GICA dataset is observed under varied conditions of three factors, these form the ways of the order-three tensor $\mathcal{X} \in \mathcal{R}^{O \times C \times S}$, where O, C, S are the number of values (levels) in ranges $\{o_1, o_2, \ldots, o_O\}$, $\{c_1, c_2, \ldots, c_C\}$ and $\{s_1, s_2, \ldots, s_S\}$ of the categorical factor variables object **o**, context **c** and subject **s** respectively. An element of the tensor, $x_{ijk} \in \mathcal{R}$, is the individual observation under certain values (o_i, c_j, s_k) taken by the factors. $\mathcal{R}$ is the range of the observed variable.[3]

11.5.1 Obtaining subjectivity data

A central question in GICA is how to obtain the data on subjectivity for expanding an object-context matrix into the tensor that accounts additionally for subjectivity. The basic idea is that for each element in the object-context matrix one needs several subjective evaluations. Specifically, *the GICA data collection* measures for each subject s_k the relevance x_{ijk} of an object o_i in a context c_j, or, more generally, the association x_{ijk} between object and context.

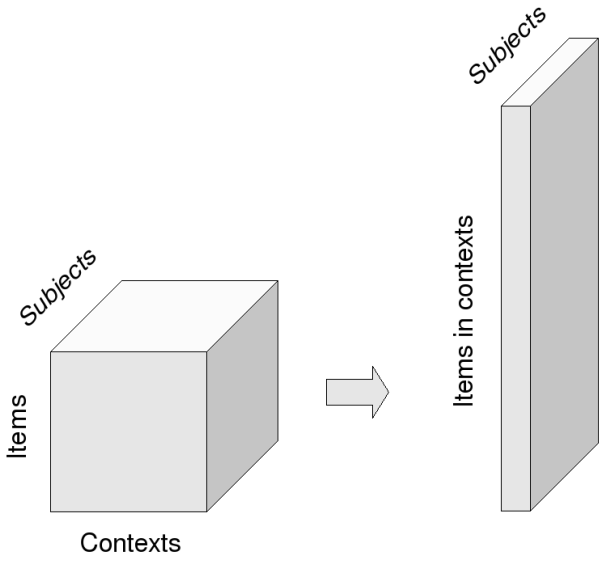


Figure 11.7: The $O \times C \times S$-element subjectivity cube flattened into a matrix in which each column corresponds to a subject and each row to a unique combination of an item and a context. The number of rows in this matrix is $O \times C$ and the number of columns is S. A transpose of this matrix gives rise to a map of persons that is the way-3 matricization of a GICA data tensor.

Conceptual survey of subjectivity

An essential step in the method is to collect a) a number of objects for which epistemological subjectivity is likely to take place, as well as b) a number of relevant contexts towards which the previously collected objects can be reflected. The context items can be short textual descriptions, longer stories, or even multimodal items such as physical objects, images or videos. The underlying idea is that between the objects and the contexts there is some kind of potential link of a varying degree. It is important to choose the contexts in such a manner that they are as clear and unambiguous as possible. The differences in the interpretations of the objects is best revealed if the "reflection surface" of the contexts is as shared as possible among the participants. Therefore, the contexts can include richer descriptions and even multimodal grounding.

The participants are then asked to fill in a data matrix which typically consists of the objects as rows and the contexts as columns. Each individual's task is to determine how strongly an object is associated with a context. A graded scale can be considered beneficial.

The data collected is analyzed using some suitable data analysis method. The essential aspect is to be able to present the rich data in a compact and understandable manner so that the conceptual differences are highlighted.

Text mining of subjectivity

Conducting a conceptual survey requires considerable amount of resources and therefore alternative means for obtaining subjectivity data are useful. As an alternative approach,

text mining can be used in this task. The basic idea is to analyze a number of documents stemming from different persons and to compare the use of a set of words or phrases by them. The comparison is based on analyzing the contexts in which each person has used each word. The more similar the contextual patterns between two persons for a word, the closer the conceptions are considered to be. The accuracy of the result is, of course, dependent on how much relevant data is available.

The method is illustrated by analyzing the State of the Union addresses by US presidents. These speeches have been given since president George Washington in 1790. For the detailed analysis, we selected all speeches between 1980 and 2011 given by Jimmy Carter, Ronald Reagan, George Bush, Bill Clinton, George W. Bush and Barack Obama. In this text mining case, populating the matrix takes place by calculating the frequencies on how often a subject uses an object word in the context of a context word. A specific feature in this study was that each president has given the State of the Union Address several times. The basic approach would be to merge all the talks by a particular president together. However, a further extension can be used, i.e., each year can be considered separately so that each president is "split" into as many subjects as the number of talks he has given (e.g. Reagan_{1984}, Reagan_{1985}, ...). This is a sensible option because it provides a chance to analyze the development of the conceptions over time. In our case, the vicinity was defined as 30 words preceding the object word. This contextual window cannot be the whole document because all the objects in a speech would obtain a similar status. On the other hand, a too short window would emphasize the syntactic role of the words. Fig. 11.8 shows a detailed view on the health area of the GICA map. Two specific conclusions can be made. First, a general tendency is that the handling of the health theme forms two clusters, the democrats of the left and the republicans on the right. However, the second conclusion is that in Barack Obama's speeches in 2010 and 2011, he has used the term in such a way that resembles the republican usage.

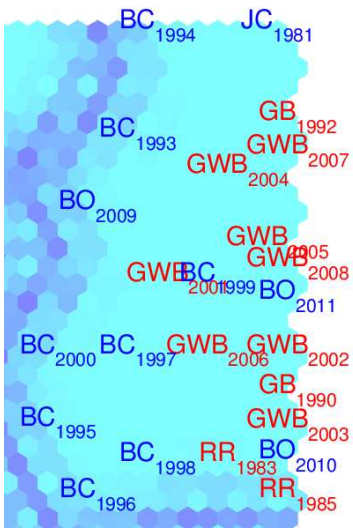


Figure 11.8: A zoomed view into the health area of the GICA map of the State of the Union Addresses.

More detailed information on the GICA method and examples of its use is available in [1, 3]. Future plans include developing the analysis, e.g., by applying different tensor analysis methods.

References

[1] T. Honkela, N. Janasik, K. Lagus, T. Lindh-Knuutila, M. Pantzar, and J. Raitio. GICA: Grounded intersubjective concept analysis – a method for enhancing mutual understanding and participation. TECHREP TKK-ICS-R41, AALTO-ICS, ESPOO, Dec. 2010.

[2] Harri Jäppinen, Timo Honkela, Heikki Hyötyniemi, and Aarno Lehtola. A multilevel natural language processing model. In *Nordic Journal of Linguistics*, 11:69–82, 1988.

[3] Timo Honkela, Juha Raitio, Krista Lagus, Ilari T. Nieminen, Nina Honkela, and Mika Pantzar. Subjects on Objects in Contexts: Using GICA Method to Quantify Epistemological Subjectivity. In *Proceedings of IJCNN 2012, International Joint Conference on Neural Networks*, 2012.

Adaptive Informatics Applications

Chapter 12

Intelligent data engineering

Miki Sirola, Mika Sulkava, Jukka Parviainen, Jaakko Talonen, Eimontas Augilius, David Ott, Kimmo Raivio, Antti Klapuri, Olli Simula

12.1 Data analysis in monitoring and decision making

Our objective has been to develop methodologies for failure management in nuclear power plant, and apply them in practice in solving typical problems on this application area [1]. Data-analysis with nuclear power plant data has been one important basic research methodology that we have used. We have developed and applied different case-based data-analysis methods and visualizations to help to detect and analyze various failure cases. Early detection of faults with data-based methods has been an important focus area. We have also used such methods as prediction and modelling to help our research [2]. To study the information value of SOM Maps (Self-Organizing Map method) for the operator with comparisons to other visualization methods is also our interest.

The fault dynamics and dependencies of power plant elements and variables have been inspected to open the way for modelling and creating useful statistics to detect process faults [3]. We have succeeded in using data mining to learn from industrial processes and finding out dependencies between variables by Principal Component Analysis (PCA) and Self-Organizing Map (SOM). In addition to industrial data also methods were developed with the voting advice application data of the Parliamentary elections [4].

In Finnish Metals and Engineering Competence Cluster (FIMECC), we have participated in Energy and Life Cycle Cost Efficient Machines (EFFIMA) Programme. The programme target has been to develop new technology and solutions that enable new machines, devices and systems with dramatically lower life cycle costs — and especially lower energy consumption — than what is the international state-of-the-art of today. Our contribution has been to develop and apply machine learning methods in measurement data analysis in order to achieve the above EFFIMA goals. As a result, a toolbox of adaptive data analysis methods have been developed for the industrial partners.

References

[1] Sirola, M., Talonen, J., Parviainen, J., and Lampi, G.. Decision support with data-analysis methods in a nuclear power plant. TKK Reports in Information and Computer Science (TKK-ICS-R29). Espoo, 2010. 23 p.

[2] Talonen, J., and Sirola, M. Modelling hypothetical wage equation by neural networks. International Conference on Artificial Neural Networks (ICANN 2011). Espoo, Finland, 2011.

[3] Talonen, J., Sirola, M., and Augilius, E. Modelling power output at nuclear power plant by neural networks. International Conference on Artificial Neural Networks (ICANN 2010). Thessaloniki, Greece, 2010.

[4] Talonen, J., and Sulkava, M. Analyzing Parliamentary Elections Based on Voting Advice Application Data. In The Tenth International Symposium on Intelligent Data Analysis (IDA), 2011.

Chapter 13

Time series prediction

Amaury Lendasse, Timo Honkela, Federico Pouzols, Antti Sorjamaa, Yoan Miche, Qi Yu, Eric Severin, Mark van Heeswijk, Erkki Oja, Francesco Corona, Elia Liitiäinen, Zhanxing Zhu, Laura Kainulainen, Emil Eirola, Olli Simula

13.1 Introduction

Amaury Lendasse

The Environmental and Industrial Machine Learning (EIML) group is a sub-group of the Adaptive Informatics Applications (AIA) group. It is part of both the Department of Information and Computer Science and the Adaptive Informatics Research Centre, Centre of Excellence of the Academy of Finland.

The EIML group is based on the former Time Series Prediction and Chemoinformatics group, and is developing new Machine Learning techniques: 1) to model environment (using e.g. time series prediction, variable selection and ensemble modeling); 2) to solve industrial problems (for example in the fields of chemometrics, electricity production and distribution, bankruptcy prediction and information security. The EIML group has been created and is lead by Dr. Amaury Lendasse, Docent. The Industrial Machine Learning is under the responsibility of Dr. Francesco Corona, Docent. The information security is under the responsibility of Dr. Yoan Miche.

13.2 Environmental Modeling and Related Tools

Amaury Lendasse,Timo Honkela, Federico Pouzols, Olli Simula and Antti Sorjamaa

Research Environmental Modeling and Time Series prediction are the main research areas of the EIML group.

Environmental Sciences Environmental Sciences have seen a great deal of development and attention over the last few decades, fostered by an impressive improvement in observational capabilities and measurement procedures. The fields of environmental modeling and analysis seek to better understand phenomena ranging from Earth-Sun interactions to ecological changes caused by climatic factors. Traditional environmental modeling and analysis approaches emphasize deterministic models and standard statistical analyses, respectively. However, the application of further developed data-driven analysis methods has shown the great value of computational analysis in environmental monitoring research. Furthermore, these analyses have provided evidence for the feasibility of predicting environmental changes. Thus, linear and nonlinear methods and tools for the analysis and predictive modeling of environmental phenomena are sought. In interpreting biological monitoring data, there is an even stronger need to develop new modeling techniques, because biota does not respond in a linear manner to environmental changes. In addition, a time lag between a stimulus and a response is common, e.g., a change in nutrient concentration and subsequent changes in algal growth, or turbidity of water. Hence, there is a demand for predictive and causal models. In this context, we follow a multidisciplinary approach, involving diverse areas of machine learning. These include time series prediction, ensemble modeling, feature selection and dimension reduction. Our activity concentrates specially on developing new methods and tools motivated by real-world needs in close cooperation with experts in the field. Our current research spans a number of areas, including long-term prediction, spatial-temporal analysis, missing data, irregular and incomplete sampling, and time-frequency analysis. Among a number of application areas, we focus on Marine Biology. Applications of our research have a direct relevance for the Baltic Sea countries. Sophisticated environmental models are needed and directly or indirectly requested by policy makers, industry and citizens. This is of special relevance in the context of regulations such as the EU Water Framework Directive, among others. Our research aims to contribute to the scientific and technological challenges posed by such regulations as well as general challenges in Environmental Sciences worldwide. Time Series Prediction

What is Time series prediction? Time series forecasting is a challenge in many fields. In finance, one forecasts stock exchange or stock market indices; data processing specialists forecast the flow of information on their networks; producers of electricity forecast the electric load and hydrologists forecast river floods. The common point to their problems is the following: how can one analyze and use the past to predict the future? In general, these methods try to build a model of the process. The model is then used on the last values of the series to predict future values. A new challenge in time series prediction is the long-term prediction also known as multiple step-ahead prediction. Many methods designed for time series forecasting perform well (depending on the complexity of the problem) on a rather short-term horizon but are rather poor on a longer-term one. This is due to the fact that these methods are usually designed to optimize the performance at short term,

their use at longer term being not optimized. Furthermore, they generally carry out the prediction of a single value while the real problem sometimes requires predicting a vector of future values in one step. One particular problem of long-term prediction is studied: the prediction of the electric load. This problem is very complex and is more and more crucial because of the liberalization of the electricity market. Electricity producers and network companies are looking for models to predict not only their needs for the next hours but also for next days and next weeks.

Our main results can be found in [1, 2, 3, 4, 5, 6, 7].

13.3 Extreme Learning Machine

Yoan Miche, Qi Yu, Eric Severin, Antti Sorjamaa, Mark van Heeswijk, Erkki Oja, Federico Pouzols, Olli Simula and Amaury Lendasse

The amount of information is increasing rapidly in many fields of science. It creates new challenges for storing the massive amounts of data as well as to the methods, which are used in the data mining process. In many cases, when the amount of data grows, the computational complexity of the used methodology also increases.
Feed-forward neural networks are often found to be rather slow to build, especially on important datasets related to the data mining problems of the industry. For this reason, the nonlinear models tend not to be used as widely as they could, even considering their overall good performances. The slow building of the networks comes from a few simple reasons; many parameters have to be tuned, by slow algorithms, and the training phase has to be repeated many times to make sure the model is proper and to be able to perform model structure selection (number of hidden neurons in the network, regularization parameters tuning. . .).
Guang-Bin Huang et al. propose an original algorithm for the determination of the weights of the hidden neurons called Extreme Learning Machine (ELM). This algorithm decreases the computational time required for training and model structure selection of the network by hundreds. Furthermore, the algorithm is rather simplistic, which makes the implementation easy.
In our research, a methodology called Optimally-Pruned ELM (OP-ELM), based on the original ELM, is proposed. The OP-ELM methodology is compared using several experiments and two well-known methods, the Least-Squares Support Vector Machine (LS-SVM) and the Multilayer Perceptron (MLP).

Our main results can be found in [8, 9, 10, 11, 12].

13.4 Process Informatics

Francesco Corona, Elia Liitiäinen, Olli Simula, Zhanxing Zhu and Amaury Lendasse

Process Informatics investigates the development and application of modeling methods from adaptive informatics on measurements from process industry. The methods aim at representing complex chemical and physical processes with models directly derived from the data collected by the automation systems present in the process plants, without an explicit regard to the first principles. We concentrate on algorithmic methods satistying properties like accuracy, robustness, computational efficiency and understandability. Accuracy, robustness and efficiency favor on-line implementations of the models in full-scale applications, whereas understandability permits the interpretation of the models from the aprioristic knowledge of the underlying phenomena. Such an approach to process modeling provides tools that can be used in real-time analysis and supervision of the processes and can be embedded in advanced model -based control strategies and optimization. Specific application domains are chemometrics, spectroscopy, chromatography and on-line analytical technologies in process and power industry. On the algorithmic side we concentrate on methods for nonlinear dimensionality reduction, variable selection, functional and regularized regression.

Our main results can be found in [16, 17, 18, 19, 20].

13.5 Bankruptcy prediction

Yu Qi, Laura Kainulainen, Eric Severin, Olli Simula, Yoan Miche, Emil Eirola and Amaury Lendasse

Bankruptcies are not only financial but also individual crises which affect many lives. Although unpredictable things may happen, bankruptcies can be predicted to some extent.

This is important for both the banks and the investors that analyze the companies, and for the companies themselves. The aim of our research is to see, whether new machine learning models combined with variable selection perform better than traditional models: Linear Discriminant Analysis, Least Squares Support Vector Machines and Gaussian Processes. They form a good basis for comparison, since LDA is a widely spread technique in the financial tradition of bankruptcy prediction, LSSVM is an example of Support Vector Machine classifiers and Gaussian Processes is a relatively new Machine Learning method.

Since all the possible combinations of the variables cannot be evaluated due to time constraints, forward selection may offer a fast and accurate solution for finding suitable variables.

Our main results can be found in [13, 14, 15].

References

[1] A. Ventela, T. Kirkkala, A. Lendasse, M. Tarvainen, H. Helminen and J. Sarvala. Climate-related challenges in long-term management of Sakylan Pyhajarvi (SW Finland) In Hydrobiologia, volume 660, pages 49–58. 2011.

[2] A. Lendasse, T. Honkela and O. Simula. European Symposium on Times Series Prediction In Neurocomputing, volume 73, pages 1919–1922. June, 2010.

[3] A. Guillen, L. Herrera, G. Rubio, A. Lendasse and H. Pomares. New method for instance or prototype selection using mutual information in time series prediction In Neurocomputing, volume 73, pages 2030–2038. June, 2010.

[4] P. Merlin, A. Sorjamaa, B. Maillet and A. Lendasse. X-SOM and L-SOM: A Double Classification Approach for Missing Value Imputation In Neurocomputing, volume 73, pages 1103-1108. March, 2010.

[5] A. Sorjamaa, A. Lendasse, Y. Cornet and E. Deleersnijder. An improved methodology for filling missing values in spatiotemporal climate data set In Computational Geosciences, volume 14, pages 55-64. January, 2010.

[6] F. Pouzols and A. Barros. Automatic Clustering-Based Identification of Autoregressive Fuzzy Inference Models for Time Series In Journal of Multivariate Analysis, volume 101, pages 811–823 . April, 2010.

[7] F. Pouzols, A. Lendasse and A. Barros. Autoregressive Time Series Prediction by Means of Fuzzy Inference Systems Using Nonparametric Residual Variance Estimation In Fuzzy Sets and Systems, volume 161, pages 471–497. February, 2010.

[8] M. Heeswijk, Y.Miche, E. and A. Lendasse. GPU-Accelerated and Parallelized ELM Ensembles for Large-scale Regression In Neurocomputing, volume 74, pages 2430-2437. September, 2011.

[9] Y. Miche, M.Heeswijk, P. Bas, O. Simula and A. Lendasse. TROP-ELM: a Double-Regularized ELM using LARS and Tikhonov Regularization In Neurocomputing, volume 74, pages 2413-2421. September, 2011.

[10] F. Pouzols and A. Lendasse. Evolving fuzzy optimally pruned extreme learning machine for regression problems In Evolving Systems, volume 1, pages 43–58. August, 2010.

[11] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten and A. Lendasse. OP-ELM: Optimally-Pruned Extreme Learning Machine In IEEE Transactions on Neural Networks, volume 21, pages 158–162. January, 2010.

[12] Q. Yu, Y. Miche, A. Sorjamaa, A. Guillen, A. Lendasse and E. Severin. OP-KNN: Method and Applications In Advances in Artificial Neural Systems, volume 2010, pages 6 pages. February, 2010.

[13] L. Kainulainen, Y. Miche, E. Eirola, Q. Yu, B. Frenay, E. Severin and A. Lendasse. Ensembles of Local Linear Models for Bankruptcy Analysis and Prediction In Case Studies in Business, Industry and Government Statistics (CSBIGS), volume 4. November, 2011.

[14] Q. Yu, Y. Miche, E. Severin and A. Lendasse. Bankruptcy Prediction with Missing Data In Proceedings of the 2011 International Conference on Data Mining, pages 279-285. July, 2011.

[15] L. Kainulainen, Q. Yu, Y. Miche, E. Eirola, E. Severin and A. Lendasse. Ensembles of Locally Linear Models: Application to Bankruptcy Prediction In Proceedings of the 2010 International Conference on Data Mining, pages 280–286. July, 2010.

[16] Z. Zhu, F. Corona, A. Lendasse, R. Baratti and J. Romagnoli. Local linear regression for soft-sensor design with application to an industrial deethanizer In 18th World Congress of the International Federation of Automatic Control (IFAC). August, 2011.

[17] F. Corona, A. Lendasse and E. Liitiainen. A boundary corrected expansion of the moments of nearest neighbor distributions In Random Structures and Algorithms, volume 37, pages 223–247. September, 2010.

[18] M. Toiviainen, F. Corona, J. Paaso and P. Teppola. Blind source separation in diffuse reflectance NIR spectroscopy using independent component analysis In Journal of Chemometrics, volume 24, pages 514–522. May, 2010.

[19] E. Liitiainen, F. Corona and A. Lendasse. On the Curse of Dimensionality in Supervised Learning of Smooth Regression Functions In Neural Processing Letters, volume 34, pages 133–154. 2011.

[20] E. Liitiainen, A. Lendasse and F. Corona. Residual variance estimation using a nearest neighbor statistic In Journal of Multivariate Analysis, volume 101, pages 811–823 . April, 2010.

Publications of the Adaptive Informatics Research Centre 2010-2011

[1] A. Ajanki, M. Billinghurst, H. Gamper, T. Järvenpää, M. Kandemir, S. Kaski, M. Koskela, M. Kurimo, J. Laaksonen, K. Puolamäki, T. Ruokolainen, and T. Tossavainen. Contextual information access with augmented reality. In *Proceedings of MLSP 2010, IEEE International Workshop on Machine Learning for Signal Processing*, pages 95–100. IEEE, August 2010.

[2] A. Ajanki, M. Billinghurst, H. Gamper, T. Järvenpää, M. Kandemir, S. Kaski, M. Koskela, M. Kurimo, J. Laaksonen, K. Puolamäki, T. Ruokolainen, and T. Tossavainen. An augmented reality interface to contextual information. *Virtual Reality*, 15(2-3):161–173, 2011.

[3] A. Ajanki, M. Billinghurst, M. Kandemir, S. Kaski, M. Koskela, M. Kurimo, J. Laaksonen, K. Puolamäki, and T. Tossavainen. Ubiquitous contextual information access with proactive retrieval and augmentation. In *The fourth International Workshop on Ubiquitous Virtual Reality (IWUVR2010)*, 2010.

[4] A. Ajanki and S. Kaski. Probabilistic proactive timeline browser. In T. Honkela, W. Duch, M. A. Girolami, and S. Kaski, editors, *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN), Part II*, Lecture Notes in Computer Science, pages 357–364, Berlin, 2011. Springer.

[5] M. Almeida, J. Bioucas-Dias, and R. Vigário. Independent phase analysis: separating phase-locked subspaces. In *Proc. 9th Int. Conf. on Latent Variable Analysis and Source Separation (LVA/ICA'2010)*, pages 427–434, St. Malo, France, 2010.

[6] M. Almeida, J. Schleimer, J. Bioucas-Dicas, and R. Vigário. Source Separation of Phase-Locked Signals. *IEEE transactions on Neural Networks*, 22:1419–1434, 2011.

[7] M. Almeida, R. Vigário, and J. Bioucas-Dias. Phase locked matrix factorization. In *Proc. 19th European Signal Processing Conference (EUSIPCO'2011)*, Barcelona, Spain, 2011.

[8] T. Alumäe and M. Kurimo. Domain adaptation of maximum entropy language models. In *Proceedings of the ACL 2010*. ACL, July 2010.

[9] T. Alumäe and M. Kurimo. Efficient estimation of maximum entropy language models with n-gram features: an SRILM extension. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*. ISCA, September 2010.

[10] P. Auer, Z. Hussain, S. Kaski, A. Klami, J. Kujala, J. Laaksonen, A. P. Leung, K. Pasupa, and J. Shawe-Taylor. Pinview: Implicit feedback in content-based image retrieval. In *Proceedings of ICML Workshop on Reinforcement Learning and Search in Very Large Spaces*, Haifa, Israel, June 2010.

[11] P. Auer, Z. Hussain, S. Kaski, A. Klami, J. Kujala, J. Laaksonen, A. P. Leung, K. Pasupa, and J. Shawe-Taylor. Pinview: Implicit feedback in content-based image retrieval. In T. Diethe, N. Cristianini, and J. Shawe-Taylor, editors, *Proceedings of Workshop on Applications of Pattern Analysis*, volume 11 of *JMLR Workshop and Conference Proceedings*, pages 51–57, 2010.

[12] I. Borze, M. Guled, S. Musse, A. Raunio, E. Elonen, U. Saarinen-Pihkala, M.-L. Karjalainen-Lindsberg, L. Lahti, and S. Knuutila. MicroRNA microarrays on archive bone marrow core biopsies of leukemias - method validation. *Leukemia Research*, 35(2):188–195, 2011.

[13] J. Caldas and S. Kaski. Hierarchical generative biclustering for microRNA expression analysis. In B. Berger, editor, *Research in Computational Molecular Biology, Proceedings of 14th Annual International Conference RECOMB 2010, Lisbon, Portugal, April 25-28*, pages 65–79, Berlin, 2010. Springer.

[14] J. Caldas and S. Kaski. Hierarchical generative biclustering for microRNA expression analysis. *Journal of Computational Biology*, 18:251–261, 2011.

[15] X. Chen and M. Koskela. Mobile visual search from dynamic image databases. In *Proceedings of Scandinavian Conference on Image Analysis (SCIA 2011)*, Ystad Saltsjöbad, Sweden, May 2011.

[16] X. Chen, M. Koskela, and J. Hyväkkä. Image based information access for mobile phones. In *Proceedings of 8th International Workshop on Content-Based Multimedia Indexing*, Grenoble, France, June 2010.

[17] K. Cho. Improved Learning Algorithms for Restricted Boltzmann Machines. Master's thesis, Aalto University School of Science, 2011.

[18] K. Cho, A. Ilin, and T. Raiko. Improved learning of gaussian-bernoulli restricted boltzmann machines. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN 2011)*, pages 10–17, Espoo, Finland, June 2011.

[19] K. Cho, T. Raiko, and A. Ilin. Parallel tempering is efficient for learning restricted boltzmann machines. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2010)*, pages 3246 – 3253, Barcelona, Spain, July 2010.

[20] K. Cho, T. Raiko, and A. Ilin. Enhanced gradient and adaptive learning rate for training restricted boltzmann machines. In *Proceedings of the International Conference on Machine Learning (ICML 2011)*, Bellevue, Washington, USA, June 2011.

[21] K. Cho, T. Raiko, and A. Ilin. Gaussian-bernoulli deep boltzmann machine. In *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, Sierra Nevada, Spain, December 2011.

[22] F. Corona, A. Lendasse, and E. Liitiäinen. A boundary corrected expansion of the moments of nearest neighbor distributions. *Random Structures and Algorithms*, 37(2):223–247, September 2010.

[23] F. Corona, E. Liitiäinen, A. Lendasse, R. Baratti, and L. Sassu. A continuous regression function for the delaunay calibration method. In *Proceedings of IFAC/DYCOPS 2010 9th International Symposium on Dynamics and Control of Process Systems, Leuven (Belgium)*, pages 180–185, July 5-7 2010.

[24] F. Corona, M. Mulas, R. Baratti, and J. Romagnoli. On the topological modeling and analysis of industrial process data using the SOM. *Computers and Chemical Engineering*, 34(12):2022–2032, 2010.

[25] Y. Deville, C. Jutten, and R. Vigário. Handbook of blind source separation: Independent component analysis and applications. In *Overview of source separation applications*, pages 639–681. Academic Press, 2010.

[26] R. S. Doddipatla and M. Kurimo. A study on combining vtln and sat to improve the performance of automatic speech recognition. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011*. ISCA, August 2011.

[27] A. Faisal, F. Dondelinger, D. Husmeier, and C. M. Beale. Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods. *Ecological Informatics*, 5:451–464, 2010.

[28] A. Faisal, A. Karilainen, M. Li, P. Ramkumar, P. Salminen, M. Turunen, and O. Pitkänen. Creating a flourishing innovation climate. In Y. Neuvo and S. Ylonen, editors, *Bit Bang II - Engergising Innovation, Innovating Energy*, pages 8–31. Helsinki University Print, 2010.

[29] J. F. Gemmeke, B. Cranen, and U. Remes. Sparse imputation for large vocabulary noise robust ASR. *Computer Speech and Language*, 25:462–479, April 2011.

[30] J. F. Gemmeke, U. Remes, and K. J. Palomäki. Observation uncertainty measures for sparse imputation. In *Proc. INTERSPEECH*, pages 2262–2265, Makuhari, Japan, September 26–30 2010.

[31] M. Gibson, T. Hirsimaki, R. Karhila, M. Kurimo, and W. Byrne. Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, March 2010.

[32] J. Gillberg. Targeted learning by imposing asymmetric sparsity. Master's thesis, Aalto University, Department of Information and Computer Science, June 2011.

[33] M. Gönen, M. Kandemir, and S. Kaski. Multitask learning using regularized multiple kernel learning. In B.-L. Lu, L. Zhang, and J. Kwok, editors, *Proceedings of 18th International Conference on Neural Information Processing (ICONIP)*, volume 7063 of *Lecture Notes in Computer Science*, pages 500–509, Berlin / Heidelberg, 2011. Springer.

[34] A. Guillén, L. Herrera, G. Rubio, A. Lendasse, and H. Pomares. New method for instance or prototype selection using mutual information in time series prediction. *Neurocomputing*, 73(10–12):2030–2038, June 2010.

[35] A. Guillén, F. M. Pouzols, A. B. Barros, L. J. Herrera, J. Gonzalez, H. Pomares, and I. Rojas. Identifying fuzzy inference models by means of possibilistic clustering: Socio-economic applications. In *Computational Methods for Modelling and Learning in Social and Human Sciences (MASHS)*, June 2010.

[36] A. Guillén, M. van Heeswijk, D. Sovilj, M. G. Arenas, L. J. Herrera, H. Pomares, and I. Rojas. Variable selection in a GPU cluster using delta test. In *IWANN (1)*, pages 393–400, 2011.

[37] B. Hanseeuw, K. V. Leemput, M. Kavec, C. Grandin, X. Seron, and A. Ivanoiu. Mild cognitive impairment: differential atrophy in the hippocampal subfields. *American Journal of NeuroRadiology*, 32(9):1658–1661, 2011.

[38] J. Hegedüs, Y. Miche, A. Ilin, and A. Lendasse. Methodology for behavioral-based malware analysis and detection using random projections and k-nearest neighbors classifiers. In *7th International Conference on Computational Intelligence and Security (CIS2011) TEST*, pages 1016 – 1023, Sanya, China, December 2011.

[39] J. Hegedüs, Y. Miche, A. Ilin, and A. Lendasse. Random projection method for scalable malware classification. In *14th International Symposium on Recent Advances in Intrusion Detection*, California, USA, September 2011.

[40] H. Hino, N. Reyhani, and N. Murata. Multiple kernel learning by conditional entropy minimization. In *ICMLA*, pages 223–228, 2010.

[41] A. Honkela, P. Gao, J. Ropponen, M. Rattray, and N. D. Lawrence. tigre: Transcription factor Inference through Gaussian process Reconstruction of Expression. BioConductor 2.6, April 2010. Computer program.

[42] A. Honkela, C. Girardot, E. H. Gustafson, Y.-H. Liu, E. E. M. Furlong, N. D. Lawrence, and M. Rattray. Model-based method for transcription factor target identification with limited data. *Proc Natl Acad Sci U S A*, 107(17):7793–7798, Apr 2010.

[43] A. Honkela, M. Milo, M. Holley, M. Rattray, and N. D. Lawrence. Ranking of gene regulators through differential equations and Gaussian processes. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010)*, pages 154–159, Kittilä, Finland, 2010.

[44] A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen. Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *Journal of Machine Learning Research*, 11:3235–3268, Nov 2010.

[45] T. Honkela. Directions for e-science and science 2.0 in human and social sciences. In *Proceedings of MASHS 2010, Computational Methods for Modeling and Learning in Social and Human Sciences*, pages 119–134. Multiprint, 2010.

[46] T. Honkela. Kunnioitusta erilaisuutta kohtaan. *Tieteessä tapahtuu*, 29(4-5):38–39, 2011.

[47] T. Honkela, W. Duch, M. Girolami, and S. Kaski, editors. *Artificial Neural Networks and Machine Learning Research - ICANN 2011, Part I*, Berlin, 2011. Springer.

[48] T. Honkela, W. Duch, M. Girolami, and S. Kaski, editors. *Artificial Neural Networks and Machine Learning Research - ICANN 2011, Part II*, Berlin, 2011. Springer.

[49] T. Honkela, W. Duch, M. A. Girolami, and S. Kaski, editors. *Artificial Neural Networks and Machine Learning - Proceedings of ICANN 2011 - 21st International Conference on Artificial Neural Networks, Part I.* Springer, 2011.

[50] T. Honkela, W. Duch, M. A. Girolami, and S. Kaski, editors. *Artificial Neural Networks and Machine Learning - Proceedings of ICANN 2011 - 21st International Conference on Artificial Neural Networks, Part II.* Springer, 2011.

[51] T. Honkela, A. Hyvärinen, and J. Väyrynen. WordICA - Emergence of linguistic representations for words by independent component analysis. *Natural Language Engineering*, 16(3):277–308, 2010.

[52] T. Honkela, N. Janasik, K. Lagus, T. Lindh-Knuutila, M. Pantzar, and J. Raitio. GICA: Grounded intersubjective concept analysis – a method for enhancing mutual understanding and participation. TECHREP TKK-ICS-R41, AALTO-ICS, ESPOO, Dec. 2010.

[53] T. Honkela, J. Laaksonen, H. Törrö, and J. Tenhunen. Media map: A multilingual document map with a design interface. In *Proceedings of 8th International Workshop on Self-Organizing Maps (WSOM 2011)*, Espoo, Finland, 2011.

[54] T. Honkela, J. Laaksonen, H. Törrö, and J. Tenhunen. Media map: A multilingual document map with a design interface. In *Advances in Self-Organizing Maps - Proceedings of WSOM 2011, 8th International Workshop*, pages 247–256, 2011.

[55] T. Honkela, T. Lindh-Knuutila, and K. Lagus. Measuring Adjective Spaces. *Proceedings of ICANN 2010, Artificial Neural Networks*, pages 351–355, 2010.

[56] T. Honkela, J. Väyrynen, and M. Pöllä. Cognitive systems blog, year 2011. Blogspot blog, 2011. Blog.

[57] I. Huopaniemi, T. Suvitaival, J. Nikkilä, M. Orešič, and S. Kaski. Multivariate multi-way analysis of multi-source data. *Bioinformatics*, 26:i391–i398, 2010. (ISMB 2010).

[58] I. Huopaniemi, T. Suvitaival, M. Orešič, and S. Kaski. Graphical multi-way models. In J. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, editors, *Machine Learning and Knowledge Discovery in Databases. Proceedings of European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010*, volume I, pages 538–553, Berlin, 2010. Springer.

[59] O.-P. Huovilainen and L. Lahti. pint: Pairwise integration of functional genomics data. BioConductor, April 2010. Computer program.

[60] K. E. H. II, M. Kurimo, and V. D. Calhoun. The sixth annual MLSP competition, 2010. In *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Kittilä, Finland, August 2010.

[61] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research (JMLR)*, 11:1957–2000, July 2010.

[62] T. Jantunen, M. Koskela, J. Laaksonen, and P. Rainò. Towards automated visualization and analysis of signed language motion: Method and linguistic issues. In *Proceedings of 5th International Conference on Speech Prosody*, Chicago, Ill. (USA), May 2010.

[63] H. Järvinen, P. Räisänen, M. Laine, J. Tamminen, A. Ilin, E. Oja, A. Solonen, and H. Haario. Estimation of ECHAM5 climate model closure parameters with adaptive MCMC. *Atmospheric Chemistry and Physics Discussion*, 10:11951–11973, 2010.

[64] C. Jutten, M. Babaie-Zadeh, and J. Karhunen. Chapter 14: Nonlinear mixtures. In C. Jutten and P. Comon, editors, *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, pages 549–592. Academic Press, 2010.

[65] A. A. Kalaitzis, P. Gao, A. Honkela, and N. D. Lawrence. gptk: Gaussian Processes Tool-Kit. CRAN, December 2010. Computer program.

[66] H. Kallasjoki, S. Keronen, G. J. Brown, J. F. Gemmeke, U. Remes, and K. Palomï¿½ki. Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments. In *International Workshop on Machine Listening in Multisource Environments*, Florence, Italy, September 1 2011.

[67] H. Kallasjoki, U. Remes, J. F. Gemmeke, T. Virtanen, and K. Palomï¿½ki. Uncertainty measures for improving exemplar-based source separation. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011*, pages 469–472, Florence, Italy, August 28-31 2011. ISCA.

[68] M. Kandemir, V.-M. Saarinen, and S. Kaski. Inferring object relevance from gaze in dynamic scenes. In *Proceedings of ETRA 2010, ACM Symposium on Eye Tracking Research & Applications, Austin, TX, USA, March 22-24*, pages 105–108, New York, NY, 2010. ACM.

[69] R. Karhila. Cross-lingual acoustic model adaptation for speaker-independent speech recognition. Master's thesis, Aalto University, 2010.

[70] R. Karhila and M. Kurimo. Unsupervised cross-lingual speaker adaptation for accented speech recognition. In *Proceedings of 2010 IEEE Workshop on Spoken Language Technology*, Berkeley, CA, December 2010.

[71] R. Karhila and M. Wester. Rapid adaptation of foreign-accented HMM-based speech synthesis. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association, INTERSPEECH*, Florence, August 2011.

[72] M. Karppa, T. Jantunen, M. Koskela, J. Laaksonen, and V. Viitaniemi. Method for visualisation and analysis of hand and head movements in sign language video. In *Proceedings of the 2nd Gesture and Speech in Interaction conference (GESPIN 2011)*, Bielefeld, Germany, 2011.

[73] S. Kaski. Self-organizing maps. In C. Sammut and G. I. Webb, editors, *Encyclopedia of Machine Learning*, pages 886–888. Springer, Berlin, 2010.

[74] S. Kaski. Three paths to relevance. In A. Hanazawa, T. Miki, and K. Horio, editors, *Brain-Inspired Information Technology*, pages 11–13. Springer, Berlin Heidelberg, 2010.

[75] S. Kaski, D. J. Miller, E. Oja, and A. Honkela, editors. *Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010)*. IEEE, Piscataway, NJ, 2010.

[76] S. Kaski and J. Peltonen. Dimensionality reduction for data visualization. *IEEE Signal Processing Magazine*, 28(2):100–104, 2011.

[77] S. Keronen, J. Pohjalainen, P. Alku, and M. Kurimo. Noise robust feature extraction based on extended weighted linear prediction in LVCSR. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011*. ISCA, August 2011.

[78] S. Keronen, U. Remes, K. Palomäki, T. Virtanen, and M. Kurimo. Comparison of noise robust methods in large vocabulary speech recognition. In *Proceedings of the 18th European Signal Processing Conference, EUSIPCO 2010*, Aalborg, Denmark, August 2010.

[79] I. Kivimäki, K. Lagus, I. T. Nieminen, J. J. Väyrynen, and T. Honkela. Using correlation dimension for analysing text data. In *Proceedings of the 20th International Conference on Artificial Neural Networks (ICANN 2010)*, volume 6352 of *Lecture Notes in Computer Science*, pages 368–373. Springer Academic Publishers, 2010.

[80] R. Kivisaari, P. Rapeli, K. V. Leemput, S. Kähkönen, V. Puuskari, O. Jokela, and T. Autti. Cerebral measurements and their correlation with the onset age and the duration of opioid abuse. *Journal of Opioid Management*, 6(6):423–429, November/December 2010.

[81] A. Klami. Inferring task-relevant image regions from gaze data. In S. Kaski, D. J. Miller, E. Oja, and A. Honkela, editors, *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 101–106. IEEE, 2010.

[82] A. Klami, editor. *Proceedings of ICANN/PASCAL2 Challenge: MEG Mind Reading*, Aalto University Publication series SCIENCE + TECHNOLOGY 29/2011, Espoo, 2011.

[83] A. Klami, P. Ramkumar, S. Virtanen, L. Parkkonen, R. Hari, and S. Kaski. ICANN/PASCAL2 challenge: MEG mind reading – overview and results. In A. Klami, editor, *Proceedings of ICANN/PASCAL2 Challenge: MEG Mind Reading*, Aalto University Publication series SCIENCE + TECHNOLOGY 29/2011, pages 3–19, Espoo, 2011.

[84] A. Klami, S. Virtanen, and S. Kaski. Bayesian exponential family projections for coupled data sources. In P. Grunwald and P. Spirtes, editors, *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (2010)*, pages 286–293, Corvallis, Oregon, 2010. AUAI Press.

[85] O. Kohonen, S. Virpioja, and K. Lagus. Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 78–86, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[86] O. Kohonen, S. Virpioja, L. Leppänen, and K. Lagus. Semi-supervised extensions to morfessor baseline. In *Proceedings of the Morpho Challenge 2010 Workshop*. Aalto University School of Science and Technology Faculty of Information and Natural Sciences Department of Information and Computer Science, Espoo, Finland, September 2010.

[87] T. Kohonen. Contextually Self-Organized Maps of Chinese words. TECHREP TKK-ICS-R30, AALTO-ICS, ESPOO, Apr. 2010.

[88] T. Kohonen. Contextually Self-Organized Maps of Chinese words, Part II. TECHREP TKK-ICS-R35, AALTO-ICS, ESPOO, Aug. 2010.

[89] T. Kohonen. New developments of nonlinear projections for the visualization of structures in nonvectorial data sets. TECHREP Aalto-ST 8/2011, Aalto University School od Science, ESPOO, May 2011.

[90] C. M. Krause, A. Alafuzoff, M. Laine, and R. Vigário. The critical nature of between- and within-subject variation in event-related brain oscillatory EEG responses. *Journal of Psychophysiology*, 77:305–305, 2010.

[91] M. Kurimo, W. Byrne, J. Dines, P. N. Garner, Matthew, Y. Guan, T. Hirsimäki, R. Karhila, S. King, H. Liang, K. Oura, L. Saheer, M. Shannon, S. Shiota, J. Tian, K. Tokuda, M. Wester, Y.-J. Wu, and J. Yamagishi. Personalising speech-to-speech translation in the EMIME project. In *Proceedings of the ACL 2010 System Demonstrations*, pages 48–53. ACL, July 2010.

[92] M. Kurimo, S. Virpioja, and V. T. T. (Editors). Proceedings of the Morpho Challenge 2010 Workshop. TECHREP TKK-ICS-R37, AALTO-ICS, ESPOO, Sept. 2010.

[93] M. Kurimo, S. Virpioja, V. Turunen, and K. Lagus. Morpho Challenge 2005-2010: Evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 87–95. ACL, July 2010.

[94] M. Kurimo, S. Virpioja, and V. T. Turunen. Overview and results of Morpho Challenge 2010. In *Proceedings of the Morpho Challenge 2010 Workshop*, Espoo, Finland, September 2010.

[95] M. Kurimo, S. Virpioja, V. T. Turunen, G. W. Blackwood, and W. Byrne. Overview of Morpho Challenge 2009. In *Multilingual Information Access Evaluation Vol. I-II, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*, Lecture Notes in Computer Science. Springer, September 2010.

[96] M. Kuusela, J. W. Lämsä, E. Malmi, P. Mehtälä, and R. Orava. Multivariate techniques for identifying diffractive interactions at the LHC. *International Journal of Modern Physics A*, 25(8):1615–1647, 2010.

[97] M. Kuusela, E. Malmi, R. Orava, and T. Vatanen. Title: Soft classification of diffractive interactions at the LHC. In *Proceedings of Diffraction 2010*. American Institute of Physics, 2010.

[98] M. Kuusela, E. Malmi, R. Orava, and T. Vatanen. Soft classification of diffractive interactions at the LHC. *AIP Conference Proceedings*, 1350(1):111–114, 2011.

[99] M. Kuusela, E. Malmi, T. Vatanen, R. Orava, T. Aaltonen, and Y. Nagai. Detection of new physics using density estimation based anomaly search. CDF/DOC/EXOTIC/CDFR/10227 (Internal note), 2010.

[100] J. Laaksonen and T. Honkela, editors. *Proceedings of 8th International Workshop on Self-Organizing Maps (WSOM 2011)*, volume LNCS 6731 of *Lecture Notes in Computer Science*. Springer, June 2011.

[101] L. Lahti. opencomp blog. Wordpress blog, 2010. Blog.

[102] L. Lahti. *Probabilistic analysis of the human transcriptome with side information*. PhD thesis, Aalto University School of Science and Technology, Faculty of Information and Natural Sciences, Department of Information and Computer Science, Espoo, December 2010. The LaTeX sources and the pdf version are freely available under cc-by license at http://www.iki.fi/Leo.Lahti.

[103] L. Lahti, L. L. Elo, T. Aittokallio, and S. Kaski. Probabilistic analysis of probe reliability in differential gene expression studies with short oligonucleotide arrays. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8:217–225, 2011.

[104] L. Lahti, A. Gusmao, and O.-P. Huovilainen. Netresponse (functional network analysis). BioConductor, October 2010. Computer program.

[105] L. Lahti and O.-P. Huovilainen. Dependency modeling toolkit. ICML workshop, June 2010. Computer program.

[106] L. Lahti, J. E. A. Knuuttila, and S. Kaski. Global modeling of transcriptional responses in interaction networks. *Bioinformatics*, 26:2713–2720, 2010. Implementations in R and Matlab available at http://netpro.r-forge.r-project.org/.

[107] G. Leen, J. Peltonen, and S. Kaski. Focused multi-task learning using Gaussian processes. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases (proceedings of ECML PKDD 2011), Part II*, pages 310–325. Springer Berlin / Heidelberg, 2011. Best Paper Award in Machine Learning.

[108] A. Lendasse, T. Honkela, and O. Simula. European symposium on times series prediction. *Neurocomputing*, 73(10–12):1919–1922, June 2010.

[109] J. Letosa and T. Honkela. Elementary Logical Reasoning in the SOM Output Space. In *Proceedings of ICANN 2010, Artificial Neural Networks*, pages 432–437. Springer, 2010.

[110] R. Louhimo, V. Aittomäki, A. Faisal, M. Laakso, P. Chen, K. Ovaska, E. Valo, L. Lahti, V. Rogojin, S. Kaski, and S. Hautaniemi. Systematic use of computational methods allows stratifying treatment responders in glioblastoma multiforme. In *Proceedings of CAMDA 2011 conference, Critical Assessment of Massive Data Analysis*, 2011. Source code available at http://csbi.ltdk.helsinki.fi/camda/.

[111] J. Luttinen and A. Ilin. Transformations in variational Bayesian factor analysis to speed up learning. *Neurocomputing*, 73:1093–1102, 2010.

[112] F. Mateo, D. Sovilj, and R. Gadea. Approximate k-NN delta test minimization method using genetic algorithms: Application to time series. *Neurocomputing*, 73(10-12):2017–2029, June 2010.

[113] M.-P. Matikainen and A. Honkela. tigreBrowser: Gene expression model browser for results from tigre R package, September 2010. Computer program.

[114] N. Matsuda, J. Laaksonen, F. Tajima, N. Miyatake, and H. Sato. Fundus image analysis using subspace classifier and its performance. In *Proceedings of 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on Advanced Intelligent Systems*, Dec. 2010.

[115] Y. Mehmood, M. Abbas, X. Chen, and T. Honkela. Self-organizing maps of nutrition, lifestyle and health situation in the world. In *Advances in Self-Organizing Maps - Proceedings of WSOM 2011, 8th International Workshop*, pages 160–167. Springer, 2011.

[116] B. Menze, K. V. Leemput, A. Honkela, E. Konukoglu, M.-A. Weber, N. Ayache, and P. Golland. A generative model for the image-based modeling of tumor growth. In *Lecture Notes in Computer Science*, pages 735–742, 2011. Proceedings of IPMI 2011, July 3-8, 2011, Irsee, Germany.

[117] B. Menze, K. V. Leemput, D. Lashkari, M. Weber, N. Ayache, and P. Golland. A generative model for brain tumor segmentation in multi-modal images. In *Lecture Notes in Computer Science*, pages 151–159, 2010. Proceedings of MICCAI2010, September 20-24, 2010, Beijing, China.

[118] P. Merlin, A. Sorjamaa, B. Maillet, and A. Lendasse. X-SOM and l-SOM: A double classification approach for missing value imputation. *Neurocomputing*, 73(7-9):1103–1108, March 2010.

[119] Y. Miche. *Developing Fast Machine Learning Techniques with Applications to Steganalysis Problems*. Doctoral dissertation, TKK Dissertations in Information and Computer Science TKK-ICS-D20, Aalto University School of Science and Technology (Espoo, Finland) and INPG (Grenoble, France), November 2010.

[120] Y. Miche, P. Bas, and A. Lendasse. Using multiple re-embeddings for quantitative steganalysis and image reliability estimation. TECHREP TKK-ICS-R34, AALTO-ICS, ESPOO, June 2010.

[121] Y. Miche, E. Eirola, P. Bas, O. Simula, C. Jutten, A. Lendasse, and M. Verleysen. Ensemble modeling with a constrained linear system of leave-one-out outputs. In M. Verleysen, editor, *ESANN2010: 18th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 19–24, Bruges, Belgium, April 28–30 2010. d-side Publications.

[122] Y. Miche, B. Schrauwen, and A. Lendasse. Machine learning techniques based on random projections. In M. Verleysen, editor, *ESANN2010: 18th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 295–302, Bruges, Belgium, April 28–30 2010. d-side Publications.

[123] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse. OP-ELM: Optimally-pruned extreme learning machine. *IEEE Transactions on Neural Networks*, 21(1):158–162, January 2010.

[124] Y. Miche, M. van Heeswijk, P. Bas, O. Simula, and A. Lendasse. TROP-ELM: a double-regularized ELM using LARS and tikhonov regularization. *Neurocomputing*, 74(16):2413–2421, September 2011.

[125] N. Mosakhani, M. Guled, L. Lahti, I. Borze, M. Forsman, J. Ryhänen, and S. Knuutila. Unique microRNA profile in Dupuytren's contracture supports deregulation of β-catenin pathway. *Modern Pathology*, 23:1544–1522, 2010.

[126] M. Nevala. Discovering functional gene-microRNA modules with probabilistic methods. Master's thesis, Aalto University School of Science and Technology, Department of Information and Computer Science, 2010.

[127] I. Nieminen. Tag recommendation in folksonomies. Master's thesis, Aalto University School of Science and Technology, Faculty of Information and Natural Sciences, Department of Information and Computer Science, Espoo, Finland, June 2010.

[128] T. Niini, L. Lahti, F. Michelacci, S. Ninomiya, C. M. Hattinger, M. Guled, T. Böhling, P. Picci, M. Serra, and S. Knuutila. Array comparative genomic hybridization reveals frequent alterations of G1/S checkpoint genes in undifferentiated pleomorphic sarcoma of bone. *Genes, Chromosomes and Cancer*, 50(5):291–306, 2011.

[129] K. Nyberg, T. Raiko, T. Tiinanen, and E. Hyvönen. Document classification utilising ontologies and relations between documents. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs (MLG 2010)*, Washington DC, USA, July 2010.

[130] P. Nymark, M. Guled, I. Borze, A. Faisal, L. Lahti, K. Salmenkivi, E. Kettunen, S. Anttila, and S. Knuutila. Integrative analysis of microRNA, mRNA and aCGH data reveals asbestos- and histology-related changes in lung cancer. *Genes, Chromosomes and Cancer*, 50(8):585–597, 2011.

[131] T. Ogawa, H. Hino, N. Reyhani, N. Murata, and T. Kobayashi. Speaker recognition using multiple kernel learning based on conditional entropy minimization. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, pages 2204–2207. Institute of Electrical and Electronics Engineers (IEEE), 2011.

[132] E. Oja, A. Ilin, J. Luttinen, and Z. Yang. Linear expansions with nonlinear cost functions: modeling, representation, and partitioning. In *WCCI 2010, Barcelona: Plenary and Invited Lectures*, pages 115 – 134. IEEE Press, 2010.

[133] E. Oja and Z. Yang. Orthogonal nonnegative learning for sparse feature extraction and approximate combinatorial optimization. *Frontiers of Electrical and Electronic Engineering in China*, 5(3):261–273, 2010.

[134] T. Pahikkala, J. Väyrynen, J. Kortela, and A. Airola, editors. *Proceedings of the 14th Finnish Artificial Intelligence Conference, STeP 2010*, number 25 in Publications of the Finnish Artificial Intelligence Society. Finnish Artificial Intelligence Society, 2010.

[135] J. Pajarinen and J. Peltonen. Efficient Planning for Factored Infinite-Horizon DEC-POMDPs. In *Proceedings of IJCAI-11, the 22nd International Joint Conference on Artificial Intelligence*, pages 325–331. AAAI Press, July 2011.

[136] J. Pajarinen and J. Peltonen. Periodic Finite State Controllers for Efficient POMDP and DEC-POMDP Planning. In *Proceedings of NIPS 2011, the 25th Annual Conference on Neural Information Processing Systems*, Dec. 2011.

[137] J. Pajarinen, J. Peltonen, and M. A. Uusitalo. Fault tolerant machine learning for nanoscale cognitive radio. *Neurocomputing*, 74(5):753 – 764, 2011.

[138] J. Pajarinen, J. Peltonen, and M. A. Uusitalo. Fault tolerant machine learning for nanoscale cognitive radio. *Neurocomputing*, 74(5):753 – 764, 2011.

[139] P. Paju, E. Malmi, and T. Honkela. Text mining and qualitative analysis of an it history interview collection. In *History of Nordic Computing*, IFIP Publications, pages 433–443. Springer, 2011.

[140] J. Parkkinen and S. Kaski. Searching for functional gene modules with interaction component models. *BMC Systems Biology*, 4:4, 2010.

[141] J. Parkkinen, K. Nybo, J. Peltonen, and S. Kaski. Graph visualization with latent variable models. In *Proceedings of MLG-2010, the Eighth Workshop on Mining and Learning with Graphs*, pages 94–101, New York, NY, USA, 2010. ACM. DOI: http://doi.acm.org/10.1145/1830252.1830265.

[142] E. Parviainen, J. Riihimäki, Y. Miche, and A. Lendasse. Interpreting extreme learning machine as an approximation to an infinite neural network. In *KDIR 2010: Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, Valencia, Spain, October 2010.

[143] M. Paukkeri, I. Kivimäki, S. Tirunagari, E. Oja, and T. Honkela. Effect of dimensionality reduction on different distance measures in document clustering. In B.-L. Lu, L. Zhang, and J. Kwok, editors, *ICONIP 2011, Part III*, number 7064 in LNCS, pages 167–176. Springer–Verlag Berlin Heidelberg, Shanghai, China, November 2011.

[144] M.-S. Paukkeri, A. P. García-Plaza, S. Pessala, and T. Honkela. Learning taxonomic relations from a set of text documents. In *Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT 2010)*, pages 105–112, Wisla, Poland, October 2010. IEEE.

[145] M.-S. Paukkeri and T. Honkela. Likey: Unsupervised Language-Independent Keyphrase Extraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 162–165, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[146] V. Peltola and A. Honkela. Variational inference and learning for non-linear state-space models with state-dependent observation noise. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010)*, pages 190–195, Kittilä, Finland, 2010.

[147] J. Peltonen, H. Aidos, N. Gehlenborg, A. Brazma, and S. Kaski. An information retrieval perspective on visualization of gene expression data with ontological annotation. In *Proceedings of ICASSP 2010, IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2178–2181, Piscataway, NJ, 2010. IEEE.

[148] J. Peltonen and S. Kaski. Generative modeling for maximizing precision and recall in information visualization. TECHREP TKK-ICS-R38, AALTO-ICS, ESPOO, Nov. 2010.

[149] J. Peltonen and S. Kaski. Generative modeling for maximizing precision and recall in information visualization. In G. Gordon, D. Dunson, and M. Dudik, editors,

Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, volume 15 of *JMLR W&CP*, pages 597–587. JMLR, 2011.

[150] J. Peltonen, Y. Yaslan, and S. Kaski. Relevant subtask learning by constrained mixture models. *Intelligent Data Analysis*, 14:641–662, 2010.

[151] C. Peters, G. D. Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Penas, and G. R. (Eds.). *Multilingual Information Access Evaluation I - Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Part I.* Lecture Notes in Computer Science, Information Systems and Applications, incl. Internet/Web, and HCI, Vol. 6241. Springer, Berlin, September 2010.

[152] M. Pöllä and T. Honkela. Negative selection of written language using character multiset statistics. *Journal of Computer Science and Technology*, 25(6):1256–1266, November 2010.

[153] F. M. Pouzols and A. B. Barros. Automatic clustering-based identification of autoregressive fuzzy inference models for time series. *Neurocomputing*, 73(10):1937–1949, August 2010.

[154] F. M. Pouzols and A. Lendasse. Effect of different detrending approaches on computational intelligence models of time series. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1729–1736, Barcelona, Spain, July 2010.

[155] F. M. Pouzols and A. Lendasse. Evolving fuzzy optimally pruned extreme learning machine: A comparative analysis. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1339–1346, Barcelona, Spain, July 2010.

[156] F. M. Pouzols and A. Lendasse. Evolving fuzzy optimally pruned extreme learning machine for regression problems. *Evolving Systems*, 1(1):43–58, August 2010.

[157] F. M. Pouzols, A. Lendasse, and A. B. Barros. Autoregressive time series prediction by means of fuzzy inference systems using nonparametric residual variance estimation. *Fuzzy Sets and Systems*, 161(4):471–497, February 2010.

[158] F. M. Pouzols, D. R. Lopez, and A. B. Barros. *Mining and Control of Network Traffic by Computational Intelligence*, volume 342 of *Studies in Computational Intelligence*. Springer, January 2011. To Appear.

[159] H. Pulakka, U. Remes, M. Kurimo, K. Palomï¿½ki, and P. Alku. Low-frequency bandwidth extension of telephone speech using sinusoidal synthesis and Gaussian mixture model. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011*. ISCA, August 2011.

[160] H. Pulakka, U. Remes, K. Palomï¿½ki, M. Kurimo, and P. Alku. Speech bandwidth extension using Gaussian mixture model-based estimation of the highband mel spectrum. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[161] J. Pylkkï¿½nen. Mallikompleksisuuden vaikutus puheentunnistimen tarkkuuteen diskriminatiivisten opetusmenetelmien kanssa. In T. K. S. Werner, editor, *XXVI Fonetiikan pï¿½ivï¿½t 2010*, pages 6–10, 2011.

[162] T. Raiko, K. Cho, and A. Ilin. Derivations of the enhanced gradient for the Boltzmann machine. Technical Report TKK-ICS-R37, Aalto University, TKK Reports in Information and Computer Science, Espoo, Finland, 2011.

[163] T. Raiko, K. Cho, and A. Ilin. Enhanced gradient for learning boltzmann machines (abstract). In *The Learning Workshop*, Fort Lauderdale, Florida, April 2011.

[164] T. Raiko and H. Valpola. Oscillatory neural network for image segmentation with biased competition for attention. In *Proceedings of the Brain Inspired Cognitive Systems (BICS 2010) symposium*, Madrid, Spain, July 2010.

[165] T. Raiko and H. Valpola. Chapter 7: Oscillatory neural network for image segmentation with biased competition for attention. In *From Brains to Systems: Brain-Inspired Cognitive Systems 2010*, volume 718 of *Advances in Experimental Medicine and Biology*, pages 75–86. Springer New York, 2011.

[166] T. Raiko, H. Valpola, and Y. LeCun. Deep learning made easier by linear transformations in perceptrons. In *Proceedings of the NIPS workshop on Deep Learning and Unsupervised Feature Learning*, Sierra Nevada, Spain, December 2011.

[167] K. Raivio. Neuraalilaskenta. In *Matemaattinen mallinnus*, pages 115–123. WSOY, 2010.

[168] J. Rajasekharan, U. Scharfenberger, N.Concalves, and R. Vigário. Image approach towards document mining in neuroscientific publications. In *Proc. 9th Int. Symposium in Advances in Intelligent Data Analysis (IDA'2010)*, pages 147–158, Tucson, Arizona, 2010.

[169] U. Remes, Y. Nankaku, and K. Tokuda. GMM-based missing-feature reconstruction on multi-frame windows. In *Proc. INTERSPEECH*, pages 1665–1668, Florence, Italy, September 2011.

[170] U. Remes, K. J. Palomäki, T. Raiko, A. Honkela, and M. Kurimo. Missing-feature reconstruction with a bounded nonlinear state-space model. *IEEE Signal Processing Letters*, 18:563–566, October 2011.

[171] N. Reyhani, H. Hino, and R. Vigario. New probabilistic bounds on eigenvalues and eigenvectors of random kernel matrices. In *Proceedings of the Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 627–634, Corvallis, Oregon, 2011. AUAI Press.

[172] N. Reyhani and E. Oja. Non-gaussian component analysis using density gradient covariance matrix. In *The 2011 International Joint Conference on Neural Networks(IJCNN 2011)*, pages 966–972. Institute of Electrical and Electronics Engineers (IEEE), 2011.

[173] T. Riklin-Raviv, K. V. Leemput, B. Menze, W. Wells, and P. Golland. Segmentation of image ensembles via latent atlases. *Medical Image Analysis*, 14(5):654–665, October 2010.

[174] S. Rogers, A. Klami, J. Sinkkonen, M. Girolami, and S. Kaski. Infinite factorization of multiple non-parametric views. *Machine Learning*, 79(1-2):201–226, 2010.

[175] M. Sabuncu and K. V. Leemput. The relevance voxel machine (RVoxM): A bayesian method for image-based prediction. In *Lecture Notes in Computer Science*, pages 99–106, 2011. Proceedings of MICCAI 2011, September 18-22, 2009, Toronto, Canada.

[176] M. Sabuncu, B. Yeo, K. V. Leemput, B. Fischl, and P. Golland. A generative model for image segmentation based on label fusion. *IEEE Transactions on Medical Imaging*, 29(10):1714–1729, October 2010.

[177] T. Saukkonen, S. Heikkinen, A. Hakkarainen, A. Häkkinen, K. V. Leemput, M. Lipsanen-Nyman, and N. Lundbom. Association of intramyocellular, intraperitoneal and liver fat with glucose tolerance in severely obese adolescents. *European Journal of Endocrinology*, 163(3):413–419, September 2010.

[178] E. Savia. *Mutual Dependency-Based Modeling of Relevance in Co-Occurrence Data*. PhD thesis, Aalto University School of Science and Technology, Faculty of Information and Natural Sciences, Department of Information and Computer Science, Espoo, June 2010.

[179] S. Savola, A. Klami, S. Myllykangas, C. Manara, K. Scotlandi, P. Ricci, S. Knuutila, and J. Vakkila. High expression of complement component 5 (C5) at tumor site associates with superior survival in Ewing's sarcoma family of tumour patients. *ISRN Oncology*, 2011, 2011.

[180] M. Sirola and J. Talonen. New visualization techniques and their assessment. In *Proceedings of the 6th IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*, 2011.

[181] M. Sirola and J. Talonen. Self-organizing map in process visualization. In *15th Int. Conf. on Knowledge-Based and Intelligent Information and Engineering Systems: KES2011*, 2011.

[182] M. Sirola, J. Talonen, J. Parviainen, and G. Lampi. Decision support with data-analysis methods in a nuclear power plant. TECHREP TKK-ICS-R29, AALTO-ICS, ESPOO, Mar. 2010.

[183] M. Sjöberg, S. Ishikawa, M. Koskela, J. Laaksonen, and E. Oja. PicSOM experiments in TRECVID 2011. In *Proceedings of the TRECVID 2011 Workshop*, Gaithersburg, MD, USA, December 2011.

[184] M. Sjöberg, M. Koskela, M. Chechev, and J. Laaksonen. PicSOM experiments in TRECVID 2010. In *Proceedings of the TRECVID 2010 Workshop*, Gaithersburg, MD, USA, November 2010.

[185] M. Sjöberg, M. Koskela, V. Viitaniemi, and J. Laaksonen. Indoor location recognition using fusion of SVM-based visual classifiers. In *Proceedings of 2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 343–348, Kittilä, Finland, August-September 2010.

[186] M. Sjöberg, M. Koskela, V. Viitaniemi, and J. Laaksonen. PicSOM experiments in ImageCLEF RobotVision. In D. Ünay, Z. Çataltepe, and S. Aksoy, editors, *Recognizing Patterns in Signals, Speech, Images and Videos*, volume 6388 of *Lecture Notes in Computer Science*, pages 190–199, Istanbul, Turkey, August 2010. Springer Berlin / Heidelberg.

[187] M. Sjöberg and J. Laaksonen. Analysing the structure of semantic concepts in visual databases. In J. Laaksonen and T. Honkela, editors, *Advances in Self-Organizing Maps*, volume 6731 of *Lecture Notes in Computer Science*, pages 338–347. Springer, 2011.

[188] P. Smit. Stacked transformations for foreign accented speech recognition. Master's thesis, Aalto University School of Science, May 2011.

[189] P. Smit and M. Kurimo. Using stacked transformations for recognizing foreign accented speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[190] P. Smit and M. Kurimo. Using stacked transformations for recognizing foreign accented speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011.

[191] A. Sorjamaa. *Methodologies for Time Series Prediction and Missing Value Imputation.* PhD thesis, Aalto University School of Science and Technology, November 2010.

[192] A. Sorjamaa and A. Lendasse. Fast missing value imputation using ensemble of SOMs. TECHREP TKK-ICS-R33, AALTO-ICS, ESPOO, June 2010.

[193] A. Sorjamaa, A. Lendasse, Y. Cornet, and E. Deleersnijder. An improved methodology for filling missing values in spatiotemporal climate data set. *Computational Geosciences*, 14:55–64, January 2010.

[194] A. Sorjamaa, A. Lendasse, and E. Séverin. Combination of SOMs for fast missing value imputation. In *Proceedings of MASHS 2010, Modèles et Apprentissage en Sciences Humaines et Sociale, Lille (France).* Models and learnings in Human and social Sciences, June 2010.

[195] D. Sovilj. Multistart strategy using delta test for variable selection. In T. Honkela, W. Duch, M. Girolami, and S. Kaski, editors, *ICANN 2011, Part II*, volume 6792 of *Lecture Notes in Computer Science*, pages 413–420. Springer, June 2011.

[196] D. Sovilj, T. Raiko, and E. Oja. Extending self-organizing maps with uncertainty information of probabilistic pca. In *IJCNN 2010*, pages 1915–1921, Barcelona, Spain, July 18-23 2010.

[197] D. Sovilj, A. Sorjamaa, Q. Yu, Y. Miche, and E. Séverin. OP-ELM and OP-KNN in long-term prediction of time series using projected input data. *Neurocomputing*, 73(10-12):1976–1986, June 2010.

[198] T. Suvitaival, I. Huopaniemi, M. Orešič, and S. Kaski. Cross-species translation of multi-way biomarkers. In T. Honkela, W. Duch, M. Girolami, and S. Kaski, editors, *Artificial Neural Networks and Machine Learning - ICANN 2011*, volume 6791 of *Lecture Notes in Computer Science*, pages 209–216. Springer Berlin / Heidelberg, 2011.

[199] M. Sysi-Aho, A. Ermolov, P. V. Gopalacharyulu, A. Tripathi, T. Seppänen-Laakso, J. Maukonen, I. Mattila, S. T. Ruohonen, L. Vähätalo, L. Yetukuri, T. Härkönen, E. Lindfors, J. Nikkilä, J. Ilonen, O. Simell, M. Saarela, M. Knip, S. Kaski, E. Savontaus, and M. Orešič. Metabolic regulation in progression to autoimmune diabetes. *PLoS Computational Biology*, 7:e1002257, 2011.

[200] S. B. Taieb, A. Sorjamaa, and G. Bontempi. Multiple-output modelling for multi-step-ahead time series forecasting. *Neurocomputing*, 73(10-12):1950–1957, June 2010.

[201] J. Talonen and M. Sirola. Modelling hypothetical wage equation by neural networks. In T. Honkela, W. Duch, M. Girolami, and S. Kaski, editors, *Artificial Neural Networks and Machine Learning – ICANN 2011*, volume 6792 of *Lecture Notes in Computer Science*, pages 381–388. Springer, 2011.

[202] J. Talonen, M. Sirola, and E. Augilius. Modelling power output at nuclear power plant by neural networks. In K. Diamantaras, W. Duch, and L. S. Iliadis, editors, *Artificial Neural Networks – ICANN 2010*, volume 6352 of *Lecture Notes in Computer Science*, pages 46–49. Springer, 2010.

[203] J. Talonen and M. Sulkava. Analyzing Parliamentary Elections Based on Voting Advice Application Data. In *The Tenth International Symposium on Intelligent Data Analysis (IDA)*, 2011.

[204] M. Toiviainen, F. Corona, J. Paaso, and P. Teppola. Blind source separation in diffuse reflectance NIR spectroscopy using independent component analysis. *Journal of Chemometrics*, 24(7–8):514–522, May 2010.

[205] A. Tripathi, A. Klami, M. Orešič, and S. Kaski. Matching samples of multiple views. *Data Mining and Knowledge Discovery*, 23:300–321, 2011.

[206] A. Tripathi, A. Klami, and S. Virpioja. Bilingual sentence matching using kernel CCA. In S. Kaski, D. J. Miller, E. Oja, and A. Honkela, editors, *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 130–135. IEEE, 2010.

[207] V. T. Turunen and M. Kurimo. Speech retrieval from unsegmented finnish audio using statistical morpheme-like units for segmentation, recognition, and retrieval. *ACM Trans. Speech Lang. Process.*, 8(1):1:1–1:25, October 2011.

[208] M. van Heeswijk, Y. Miche, E. Oja, and A. Lendasse. GPU-accelerated and parallelized ELM ensembles for large-scale regression. *Neurocomputing*, 74(16):2430–2437, 2010.

[209] M. van Heeswijk, Y. Miche, E. Oja, and A. Lendasse. Solving large regression problems using an ensemble of GPU-accelerated ELMs. In M. Verleysen, editor, *ESANN2010: 18th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 309–314, Bruges, Belgium, April 28–30 2010. d-side Publications.

[210] M. van Heeswijk, Y. Miche, E. Oja, and A. Lendasse. GPU-accelerated and parallelized ELM ensembles for large-scale regression. *Neurocomputing*, 74(16):2430–2437, September 2011.

[211] T. Vatanen, M. Kuusela, E. Malmi, T. Raiko, T. Aaltonen, and Y. Nagai. Fixed-background EM algorithm for semi-supervised anomaly detection. Technical report, Aalto University School of Science, 2011.

[212] T. Vatanen, J. J. Väyrynen, and S. Virpioja. Language identification of short text segments with n-gram models. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 3423–3430. European Language Resources Association (ELRA), 2010.

[213] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.

[214] A.-M. Ventelä, T. Kirkkala, A. Lendasse, M. Tarvainen, H. Helminen, and J. Sarvala. Climate-related challenges in long-term management of säkylän pyhäjärvi (SW finland). *Hydrobiologia*, Online First:1–10, 2010.

[215] J. Viinikanoja. Locally linear robust bayesian dependency modeling of co-occurrence data. Master's thesis, Aalto University, School of Science and Technology, Department of Information and Computer Science, August 2010.

[216] J. Viinikanoja, A. Klami, and S. Kaski. Variational Bayesian mixture of robust CCA models. In A. G. José Luis Balcázar, Francesco Bonchi and M. Sebag, editors, *Machine Learning and Knowledge Discovery in Databases. Proceedings of European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010*, volume III, pages 370–385, Berlin, 2010. Springer.

[217] V. Viitaniemi and J. Laaksonen. Region matching techniques for spatial bag of visual words based image category recognition. In *Proceedings of 20th International Conference on Artificial Neural Networks (ICANN 2010)*, volume 6352 of *Lecture Notes in Computer Science*, pages 531–540, Thessaloniki, Greece, Sept. 2010. Springer Verlag.

[218] V. Viitaniemi, M. Sjöberg, M. Koskela, and J. Laaksonen. Automatic video search using semantic concepts. In *Proceedings of 8th European Conference on Interactive TV and Video (EuroITV 2010)*, Tampere, Finland, June 2010.

[219] V. Viitaniemi, M. Sjöberg, M. Koskela, and J. Laaksonen. Concept-based video search with the PicSOM multimedia retrieval system. TECHREP TKK-ICS-R39, AALTO-ICS, ESPOO, Dec. 2010.

[220] S. Virpioja, O. Kohonen, and K. Lagus. Unsupervised morpheme analysis with Allomorfessor. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments: 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 – October 2, 2009, Revised Selected Papers*, volume 6241 of *Lecture Notes in Computer Science*, pages 609–616. Springer Berlin / Heidelberg, September 2010.

[221] S. Virpioja, O. Kohonen, and K. Lagus. Evaluating the effect of word frequencies in a probabilistic generative model of morphology. In B. S. Pedersen, G. Nešpore, and I. Skadina, editors, *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, volume 11 of *NEALT Proceedings Series*, pages 230–237. Northern European Association for Language Technology, Riga, Latvia, May 2011.

[222] S. Virpioja, M. Lehtonen, A. Hultén, R. Salmelin, and K. Lagus. Predicting reaction times in word recognition by unsupervised learning of morphology. In T. Honkela, W. Duch, M. Girolami, and S. Kaski, editors, *Artificial Neural Networks and Machine Learning — ICANN 2011*, volume 6791 of *Lecture Notes in Computer Science*, pages 275–282. Springer Berlin / Heidelberg, June 2011.

[223] S. Virpioja, A. Mansikkaniemi, J. Väyrynen, and M. Kurimo. Applying morphological decompositions to statistical machine translation. In *Proceedings of the Joint*

Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 201–206. Association for Computational Linguistics, July 2010.

[224] S. Virpioja, V. T. Turunen, S. Spiegler, O. Kohonen, and M. Kurimo. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2):45–90, 2011.

[225] S. Virtanen. Bayesian exponential family projections. Master's thesis, Aalto University School of Science and Technology, Department of Information and Computer Science, June 2010.

[226] S. Virtanen, A. Klami, and S. Kaski. Bayesian cca via group sparsity. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 457–464, New York, NY, USA, June 2011. ACM.

[227] S. Vishwanathan, S. Kaski, J. Neville, and S. Wrobel, editors. *Machine Learning, Special Issue on Learning and Mining with Graphs, 82(2)*, 2011.

[228] P. Wagner. On the stability of reinforcement learning under partial observability and generalizing representations. Master's thesis, Aalto University School of Science and Technology, Faculty of Information and Natural Sciences, Department of Information and Computer Science, Espoo, Finland, June 2010.

[229] M. Wester, J. Dines, M. Gibson, H. Liang, Y.-J. Wu, L. Saheer, S. King, K. Oura, P. N. Garner, W. Byrne, Y. Guan, T. Hirsimäki, R. Karhila, M. Kurimo, M. Shannon, S. Shiota, J. Tian, K. Tokuda, and J. Yamagishi. Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In *Proceedings of the 7th ISCA Speech Synthesis Workshop, SSW7*. ISCA, September 2010.

[230] M. Wester and R. Karhila. Speaker similarity evaluation of foreign-accented speech synthesis using hmm-based speaker adaptation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, May 2011.

[231] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo. Thousands of voices for HMM-based speech synthesis-analysis and application of TTS systems built on various ASR corpora. *IEEE Transactions on Audio, Speech and Language Processing*, 18(5):984–1004, July 2010.

[232] Z. Yang and E. Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 21(5):734–749, 2010.

[233] Z. Yang and E. Oja. Projective nonnegative matrix factorization based on alpha-divergence. *Journal of Artificial Intelligence and Soft Computing Research*, 1(1):7–16, 2011.

[234] Z. Yang and E. Oja. Unified development of multiplicative algorithms for linear and quadratic nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 22(12):1878–1891, 2011.

[235] Z. Yang, C. Wang, and E. Oja. Multiplicative updates for t-sne. In *Proceedings of the 20th IEEE International Workshop on Machine Learning For Signal Processing (MLSP2010)*, pages 19–23, Kittilä, August 2010.

[236] Z. Yang, H. Zhang, Z. Yuan, and E. Oja. Kullback-leibler divergence for nonnegative for nonnegative matrix factorization. In *Proceedings of 21st International Conference on Artificial Neural Networks*, pages 14–17, Espoo, Finland, 2011. Springer.

[237] Z. Yang, Z. Zhu, and E. Oja. Automatic rank determination in projective nonnegative matrix factorization. In *Proceedings of the 9th International Conference on Latent Variable Analysis and Signal Separation (LVA2010)*, volume 6365 of *Lecture Notes in Computer Science*, pages 514–521, Saint Malo, France, September 2010. Springer.

[238] L. Yetukuri, I. Huopaniemi, A. Koivuniemi, M. Maranghi, A. Hiukka, H. Nygren, S. Kaski, M.-R. Taskinen, I. Vattulainen, M. Jauhiainen, and M. Orešič. High density lipoprotein structural changes and drug response in lipidomic profiles following the long-term fenofibrate therapy in the FIELD substudy. *PLoS ONE*, 6(8):e23589, 2011.

[239] Q. Yu, Y. Miche, E. Eirola, M. van Heeswijk, E. Séverin, and A. Lendasse. Regularized extreme learning machine for regression with missing data. In *International Symposium on Extreme Learning Machines (ELM2011)*, Hangzhou, China, December 2011.

[240] Q. Yu, Y. Miche, E. Séverin, and A. Lendasse. Bankruptcy prediction with missing data. In *The International Conference on Data Mining*, pages 279–285, Las vegas, USA, July 2011.

[241] H. Zhang, E. Augilius, T. Honkela, J. Laaksonen, H. Gamper, and H. Alene. Analyzing emotional semantics of abstract art using low-level image features. In *Proceedings of 10th International Symposium on Intelligent Data Analysis (IDA 2011)*. Springer, 2011.

[242] H. Zhang, T. Ruokolainen, J. Laaksonen, C. Hochleitner, and R. Traunmüller. Gaze- and speech-enhanced content-based image retrieval in image tagging. In *Proceedings of 21st International Conference on Artificial Neural Networks (ICANN 2011)*, Espoo, Finland, 2011.

[243] H. Zhang, M. Sjöberg, J. Laaksonen, and E. Oja. A multimodal information collector for content-based image retrieval system. In *Proceedings of 18th International Conference on Neural Information Processing (ICONIP 2011)*. Springer, 2011.

[244] Z. Zhu, F. Corona, A. Lendasse, R. Baratti, and J. Romagnoli. Local linear regression for soft-sensor design with application to an industrial deethanizer. In *Proceedings of the 18th IFAC World Congress*, volume 18, pages 2839–2844, Milano, Italy, 2011.