BIENNIAL REPORT

2012 – 2013

Academy of Finland Centre of Excellence in Computational Inference (COIN)

Department of Information and Computer Science

Aalto University School of Science

P.O. Box 15400

FI–00076 Aalto, Finland

E. Oja & M. J. Mantere, editors

Otaniemi, April 2014

Unigrafia Oy
Helsinki 2014

Contents

Preface

The **Centre in Computational Inference Research** (COIN, laskennallisen päättelyn tutkimusyksikkö) was nominated as one of the national Centres of Excellence (CoE) by the Academy of Finland for the period 2012 - 2017. It is financed by the Academy, Aalto University, University of Helsinki, and Nokia Co.

The present Biennial Report covers the activities of COIN during its first two years 2012 and 2013. It concentrates on the research projects, but also lists the degrees and awards given to the staff, as well as the activities and international mobility of the CoE.

COIN is operating within three departments: The Department of Information and Computer Science (ICS), belonging to the School of Science of Aalto University, and the Departments of Computer Science (CS) and Mathematics and Statistics (MS), belonging to the University of Helsinki (UH). Aalto Distinguished Professor Erkki Oja is the director of COIN, and Professor Samuel Kaski is the vice-director, with Professors Ilkka Niemelä, Erik Aurell (jointly between Aalto and the Royal Institute of Technology in Stockholm, Sweden), Petri Myllymäki and Jukka Corander, as well as senior researcher Jorma Laaksonen, leading the CoE teams. In addition, 49 post-doctoral researchers, ca. 55 full-time graduate students, and a number of undergraduate students were working in the COIN projects.

To briefly list the main numerical outputs of COIN during the period 2012 - 2013, the Centre produced 11 D.Sc. (Tech.) or PhD degrees and 40 M.Sc. degrees. The number of scientific publications appearing during the period was 232, of which 78 were journal papers.

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, Aalto University, the University of Helsinki, and Nokia Corporation, COIN researchers managed to obtain a fairly large number of external projects. Many of these are going on still in 2014. The research staff were active in international organizations, editorial boards of journals, and conference committees, including the conferences and workshops "Statistical Mechanics of Unsatisfiability and Glasses, Mariehamn, 2012", "6th Int. Workshop on Stemmatology, Helsinki, 2012", "13th Scandinavian Symposium and Workshops on Algorithm Theory (SWAT), Helsinki, 2012", "3rd Permafrost Workshop on Modeling Bacterial and Viral Evolution, Cepina, 2012", "5th Workshop on Information Theoretic Methods in Science and Engineering (WITMSE), Amsterdam, 2012", "4th Permafrost Workshop on Modeling Bacterial and Viral Evolution, Cepina, 2013", "6th Workshop on Information Theoretic Methods in Science and Engineering (WITMSE), Tokyo, 2013", "18th Scandinavian Conference on Image Analysis (SCIA), Espoo, 2013", "16th Int. Conference on Theory and Applications of Satisfiability Testing (SAT), Helsinki, 2013", and "Statistical Mechanics of Biological Cooperativity, Mariehamn, 2013", which all were organized by the COIN staff and chaired by COIN 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 first meeting of the Scientific Advisory Board of COIN, consisting of Professors Adnan Darwiche, Dan Geiger and Roderick Murray-Smith, was held on November 1 - 2, 2012. The next meeting will be on May 8, 2014.

Erkki Oja

Distinguished Professor
Director, Centre of
Computational Inference
Research

Samuel Kaski

Professor
Vice-Director, Centre of
Computational Inference
Research

Personnel

Group of professor Erkki Oja, Director of COIN, Aalto-ICS

Oja, Erkki; D.Sc. (Tech.) Aalto Distinguished Professor
Dikmen, Onur; Ph.D. Postdoctoral researcher
Ilin, Alexander; Ph.D. Postdoctoral researcher
Mozeika, Alexander; Ph.D. Postdoctoral researcher
Raiko, Tapani; D.Sc. (Tech.) Postdoctoral researcher
Yang, Zhirong; D.Sc. (Tech.) Postdoctoral researcher
Cho, KyungHyung; M.Sc. Doctoral student
Luttinen, Jaakko; M.Sc. (Tech.) Doctoral student

Group of professor Samuel Kaski, Vice-director of COIN, HIIT/Aalto-ICS

Kaski, Samuel; D.Sc. (Tech.) Director of HIIT; professor
Honkela, Antti; D.Sc. (Tech.) Senior Researcher
Peltonen, Jaakko; D.Sc. (Tech.) Academy research fellow
Cheng, Lu; Ph.D. Postdoctoral researcher
Bunte, Kerstin; Ph.D. Postdoctoral researcher
Dutta, Ritabrata; Ph.D. Postdoctoral researcher
Eugster, Manuel; Ph.D. Postdoctoral researcher
Georgii, Elizabeth; Ph.D. Postdoctoral researcher
Gönen, Mehmet; Ph.D. Postdoctoral researcher
Klami, Arto; D.Sc. (Tech.) Postdoctoral researcher
Marttinen, Pekka; D.Sc. (Tech.) Postdoctoral researcher
Mononen, Tommi; D.Sc. (Tech.) Postdoctoral researcher
Rajala, Tuomas; D.Sc. (Tech.) Postdoctoral researcher
Zhao, Xuran; Ph.D. Postdoctoral researcher
Martinez, Ana; Ph.D. Postdoctoral researcher
Ajanki, Antti; M.Sc. (Tech.) Doctoral student
Ammad Ud Din, Muhammad; M.Sc. Doctoral student
Caldas, José; M.Sc. Doctoral student
Faisal, Ali; M.Sc. Doctoral student
Georgatzis, Konstantin; M.Sc. Doctoral student
Gillberg, Jussi; M.Sc. (Tech.) Doctoral student

Huopaniemi, Ilkka; M.Sc. (Tech.) Doctoral student
Kandemir, Melih; M.Sc. Doctoral student
Khan, Suleiman; M.Sc. Doctoral student
Leppäaho, Eemeli; M.Sc. (Tech.) Doctoral student
Lin, Ziyuan; Ph.D. Doctoral student
Nybo, Kristian; M.Sc. (Tech.) Doctoral student
Parkkinen, Juuso; M.Sc. (Tech.) Doctoral student
Pootchi, Hanieh; M.Sc. Doctoral student
Suvitaival, Tommi; M.Sc. (Tech.) Doctoral student
Topa, Hande; M.Sc. Doctoral student
Viinikanoja, Jaakko; M.Sc. (Tech.) Doctoral student
Virtanen, Seppo; M.Sc. (Tech.) Doctoral student

Group of professor Erik Aurell, Aalto-ICS

Aurell, Erik; D.Sc. (Tech.) FiDiPro professor
Chatterjee, Saikat; Ph.D. Senior researcher
Mozeika, Alexander; Ph.D. Postdoctoral researcher
Skwark, Martin; Ph.D. Posdoctoral researcher
Del Ferraro, Gino; M.Sc. Doctoral student
Zheng, HongLi; M.Sc. Doctoral student
Innocenti, Nicolas; M.Sc. Doctoral student
Lemoy, Remí; M.Sc. Doctoral student

Group of professor Jukka Corander, UH-MS

Corander, Jukka; Ph.D. Professor
Guttman, Michael; Ph.D. Postdoctoral reseracher
Martino, Luca; Ph.D. Postdoctoral researcher
Sirén, Jukka; Ph.D. Doctoral student, postdoctoral researcher
Wei, Lu; Ph.D. Postdoctoral researcher
Cheng, Lu; M.Sc. Doctoral student
Cui, Yaqiong; M.Sc. Doctoral student
Jääskinen, Väinö; M.Sc. Doctoral student
Kohonen, Jukka; M.Sc. Doctoral student
Miettinen, Minna; M.Sc. Doctoral student
Numminen, Elina; M.Sc. Doctoral student
Pessia, Alberto; M.Sc. Doctoral student
Roto, Elina; M.Sc. Doctoral student
Shubin, Mikhail; M.Sc. Doctoral student
Xiong, Jie; M.Sc. Doctoral student

Group of Dr. Jorma Laaksonen, Aalto-ICS

Laaksonen, Jorma; D.Sc. (Tech.) Teaching researcher
Kurimo, Mikko; D.Sc. (Tech.) Senior researcher
Koskela, Markus; D.Sc. (Tech.) Senior researcher
Palomäki, Kalle; D.Sc. (Tech.) Academy research fellow
Dhananjaya, Gowda; Ph.D. Postdoctoral researcher
Gonzalez Caro, Cristina; Ph.D. Postdoctoral researcher
Mesaros, Annamaria; Ph.D. Postdoctoral researcher
Chen, Xi; M.Sc. Doctoral student
Enarvi, Seppo; M.Sc. (Tech.) Doctoral student
Ishikawa, Satoru; M.Sc. Doctoral student
Kallasjoki,Heikki; M.Sc. (Tech.) Doctoral student
Karhila, Reima; M.Sc. (Tech.) Doctoral student
Keronen, Sami; M.Sc. (Tech.) Doctoral student
Kohonen, Oskar; M.Sc. (Tech.) Doctoral student
Mansikkaniemi, Andre; M.Sc. (Tech.) Doctoral student
Nieminen, Ilari; M.Sc. (Tech.) Doctoral student
Remes, Ulpu; M.Sc. (Tech.) Doctoral student
Ruokolainen, Teemu; M.Sc. (Tech.) Doctoral student
Sjöberg, Mats; M.Sc. (Tech.) Doctoral student
Smit, Peter; M.Sc. Doctoral student
Turunen, Ville; M.Sc. (Tech.) Doctoral student; Postdoctoral researcher
Varjokallio, Matti; M.Sc. (Tech.) Doctoral student
Viitaniemi, Ville; M.Sc. (Tech.) Postdoctoral researcher
Virpioja, Sami; M.Sc. (Tech.) Doctoral student

Group of Professor Petri Myllymäki, UH-CS

Myllymäki, Petri; Ph.D. Professor
Rissanen, Jorma; Ph.D. Senior researcher
Roos, Teemu; Ph.D. Assistant professor
Głowacka, Dorota; Ph.D. Postdoctoral researcher
Järvisalo, Matti; D.Sc. (Tech.) Postdoctoral researcher
Klami, Arto; D.Sc. (Tech.) Postdoctoral researcher
Kontkanen, Petri; Ph.D. Postdoctoral researcher
Malone, Brandon; Ph.D. Postdoctoral researcher
Tasoulis, Sotiris; Ph.D. Postdoctoral researcher
Määttä, Jussi; M.Sc. Doctoral student
Perkiö, Jukka; M.Sc. Doctoral student
Pulkkinen, Teemu; M.Sc. Doctoral student
Wettig, Hannes; M.Sc. Doctoral student
Zou, Yuan; M.Sc. Doctoral student

Group of Professor Ilkka Niemelä, Aalto-ICS

Niemelä, Ilkka; D.Sc. (Tech.) Vice-president, professor
Janhunen, Tomi; D.Sc. (Tech.) Professor (fixed term)
Junttila, Tommi; D.Sc. (Tech.) Teaching researcher
Rintanen, Jussi; D.Sc. (Tech.) Senior researcher
Gebser, Martin; Ph.D. Postdoctoral researcher
Hyvärinen, Antti; D.Sc. (Tech.) Postdoctoral researcher
Liu, Guohua; Ph.D. Postdoctoral researcher
Oikarinen, Emilia; D.Sc. (Tech.) Postdoctoral researcher
Kindermann, Roland; M.Sc. Doctoral student
Laitinen, Tero; M.Sc. (Tech.) Doctoral student

Support staff, Aalto University

Bingham, Ella Research Coordinator, Aalto
Ehrstedt, Stefan HR coordinator, Aalto
Jusslin, Anu Academic Coordinator, Aalto
Kauppila, Minna Secretary, Aalto
Koivisto, Leila Controller, Aalto
Mantere, Maarit Research Coordinator, Aalto
Pihamaa, Tarja Secretary, Aalto
Ranta, Markku Works Engineer, Aalto
Sirola, Miki Laboratory Engineer, Aalto

Support staff, University of Helsinki

Kuuppelomäki, Päivi Planning Officer, HIIT/UH
Meldo, Satu-Maija Secretary, UH
Moen, Pirjo Research Coordinator, UH
Nikunen, Martti IT support, UH

Awards and activities

Prizes and academic awards received by personnel of the unit

Professor Erkki Oja

- Appointment as Aalto Distinguished Professor (as of 1st of August 2013)

Professor Samuel Kaski

- Winner of Drug Sensitivity Prediction Challenge 2012, National Cancer Institute NCI (USA) & DREAM, USA.
- Best paper award, ACM International Conference on Intelligent User Interfaces, 2013, USA.

Dr. Dorota Głowacka

- Best paper award for 'Directing exploratory search: Reinforcement learning from user interactions with keyword', co-authored by T. Ruotsalo, K. Konyushkova, K. Athukorala, S. Kaski, and G. Jacucci, in the 18th International Conference on Intelligent User Interfaces (IUI 2013).

Dr. Matti Järvisalo

- 13th International Conference on Principles of Knowledge Representation and Reasoning, Distinguished Student Paper Prize for the paper "Complexity-Sensitive Decision Procedures" co-authored with Wolfgang Dvorak, Johannes Peter Wallner, and Stefan Woltran. Rome, Italy, 2012.

Dr. Arto Klami

- Dissertation Award of the Finnish Artificial Intelligence Society (2008-2011).
- Best paper award, Asian conference on Machine learning, 2012, Singapore.

Dr. Teemu Roos

- Best Paper Honorable Mention Award, ACM SIGCHI Conference on Human Factors in Computing Systems, CHI-2013, 30.04.2013.

Dr. Jukka Sirén

- The Finnish society for Bioinformatics doctoral thesis award 2012.
- The Finnish Statistical Society Doctoral Thesis Award 2009-2012.

M.Sc. (Tech.) Tommi Suvitaival

- Foundation for the promotion of technological advances, Incentive grant 2012.

Important international positions of academic service held by personnel of the unit

Professor Erik Aurell:

- Program committee chair:
  Statistical Mechanics of Unsatisfiability and Glasses, Mariehamn, May 23-26, 2012.
  Statistical Mechanics of Biological Cooperativity, Mariehamn, May 22-25, 2013.
- Expert statement for filling a professorship, Beräkningsbiologi och biologisk fysik, dnr PA 2012/122, Lund University, Sweden.
- AERES, Member of evaluation committee of CNRS/Paris-Sud laboratory, LPTMS UMR 8626, France.
- Opponent at the doctoral dissertation of Matteo Figliuzzi, University of Rome 'La Sapienza', Italy, 2013.

Professor Jukka Corander:

- Organizing Committee Chair:
  3rd Permafrost Workshop on modeling bacterial and viral evolution, Cepina, Italy, 2.–6.3.2012.
  4th Permafrost Workshop on modeling bacterial and viral evolution, Cepina, Italy, 1.–5.2.2013.
- Member of the management committee for European Science Foundation Forward Looks[1] on personalized medicine 2011-12, representing the standing committee for physical and engineering sciences (PEN).
- Norwegian Research Council review panel for mathematics and statistics, October 2012.
- Opponent at the doctoral defense of Carl Nettelblad, Uppsala University, Sweden, 2012.

Professor Tomi Janhunen:

- Co-organizer of SAT/SMT Summer School 2013, Espoo, Finland, 2013.
- Program Committee Member:
  The 20th European Conference on Artificial Intelligence (ECAI'12) Montpellier, France, August 27–31, 2012.
  The 13th European Conference on Logics in Artificial Intelligence (JELIA'12) Tolouse, France, September 26–28, 2012.
  The 13th International Conference on Principles of Knowledge Representation and Reasoning (KR'12) Rome, Italy, June 10–14, 2012.
  The 26th Workshop on (Constraint) Logic Programming (WLP'12) Bonn, Germany, September 24–25, 2012.
  The 12th International Conference on Logic Programming and Nonmonotonic Reasoning Spain, 2013.
  The 7th International Workshop on Modular Ontologies, Spain, 2013.

[1] http://www.esf.org/iPM)

- Session Chairman:
  The 12th International Conference on Logic Programming and Nonmonotonic Reasoning, Spain, 2013.

Professor Samuel Kaski:

- Program Committee Member:
  The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), Bristol, UK, 2012.
  The 29th International Conference on Machine Learning (ICML), Edinburgh, UK, 2012
  The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), Prague, Czech Republic, 2013.
  The 30th International Conference on Machine Learning (ICML), Atlanta, USA, 2013.
  The Seventh International Workshop on Machine Learning in Systems Biology (MLSB), Berlin, Germany, 2013.

- Invited talk in ICML2012 Workshop on Machine Learning in Genetics and Genomics UK, 'Data translation in genomics with multi-view machine learning'.

- Session Chairman in ECML-PKDD 2012, UK.

- PASCAL Network of Excellence, Member of Steering Committee, UK.

- Associate Editor:
  International Journal of Knowledge Discovery in Bioinformatics, Singapore.
  Journal of Machine Learning Research, USA.

- Editorial Board Member, Cognitive Neurodynamics, Germany.

- Opponent at the doctoral dissertation of Trine Abrahamsen, DTU, Denmark, 2013.

Prof. Mikko Kurimo:

- Program Committee member:
  Interspeech 2012, Portland OR, USA, 10–13.9.2012.
  Sigmorphon 2012, Montreal Canada, 7.–6.2012.

- Session Chairman:
  Interspeech 2012, Portland OR, USA, 10–13.9.2012.

- Session Chairman:
  Simple4All, End of Year Meeting, 2012.
  Simple4All, End of Year Meeting, Cluj-Napoca, 15.-20.10.2013.

- Editorial Board Member, ACM transactions on speech and language processing, USA.

- Opponent at the doctoral dissertation of Burcu Can, University of York, UK, 2012.

Professor Petri Myllymäki:

- European Research Council, External referee, 2013.

- Program Committee Co-Chair:
  The Sixth International Workshop on Stemmatology, University of Helsinki, 27–30 June 2012.
  The Fifth Workshop on Information Theoretic Methods in Science and Engineering, WITMSE 2012, Amsterdam, the Netherlands, 27.–30.8.2012.
  The Sixth Workshop on Information Theoretic Methods in Science and Engineering, WITMSE 2013, Tokyo, Japan, 26.–29.08.2013.

- Senior Program Committee member:
  International Joint Conference on Artificial Intelligence (IJCAI-2013), 2013.
  The 28th Conference on Uncertainty in Artificial Intelligence (UAI-2012), 2012.
  The 29th Conference on Uncertainty in Artificial Intelligence (UAI-2013), 2013.

- Program Committee Member:
  The Sixth International Workshop on Stemmatology, University of Helsinki, 27–30 June 2012.
  The Sixth European Workshop on Probabilistic Graphical Models (PGM-2012), 19-21 September 2012, Granada, Spain.

- Invited talk at the Industry Track of the ECML/PKDD-2012 conference, 'Somebody Needs Your Algorithm - Cloud'N'Sci.fi', Bristol, UK, 27.9.2012.

- Guest Editor of International Journal of Approximate Reasoning, Special Issue on Selected Papers from PGM-2010, 2012.

- Assesment for a professorship, University of Electro-Communications, Japan, 2013.

- Opponent at the doctoral defense of Juan Diego Rodriguez, University of the Basque Contry, 2013.

Professor Ilkka Niemelä:

- European Science Foundation, Steering Committee Member of the ESF Research Networking Program on Games for Design and Verification, France.

- European Science Foundation, Member of the Management Committee of Action IC0901: Rich-Model Toolkit - An Infrastructure for Reliable Computer Systems, France.

- Associate Editor, Theory and Practice of Logic Programming, UK.

Professor Erkki Oja:

- International Neural Network Society, INNS, College of Fellows -steering committee member, USA.

- IEEE Computational Intelligence Society, Fellowship Committee: chairman; Awards Committee: member, USA.

- European Commission, FET Flagship - programme External Advisory Group: member, Belgium.

- European Commission, FET Flagship ERANET - program evaluator, Belgium.

- Euroopan Commission, ERC Starting Grant - evaluation panel member, field of computer science, Belgium.

- European Commission, FET Open-project Brain-i-Nets EU-evaluator, Belgium.
- Halmstad Universitet, Research evaluation panel member, Sweden.
- Danmarks Tekniske Universitet, DTU Compute research evaluation panel chairman, Denmark.
- Elected Member of the Academia Europaea.
- Chairman of the Program Committee: 18th Scandinavian Conference on Image Analysis (SCIA), June 17-20, 2013, Espoo.
- Editorial Board Member:
  Natural Computing - An International Journal, the Netherlands
  Neural Computation, USA.
  International Journal of Pattern Recognition and Artificial Intelligence, Singapore.

Doc. Tommi Junttila:

- Program Committee Member:
  JELIA 2012 - 13th European Conference on Logics in Artificial Intelligence.
  SAT 2013 - 16th International Conference on Theory and Applications of Satisfiability Testing.
  10th International Workshop on the Implementation of Logics, 2013.
- Member of the PhD jury of Maximilien Colange, l'Université Pierre & Marie Curie, Paris, 2013.

Dr. Matti Järvisalo:

- Co-organizer of SAT-SMT Summer School 2013, Espoo, Finland, 2013.
- Program Committee Co-Chair: The Sixth International Workshop on Stemmatology, 27–30 June 2012, University of Helsinki.
- Chair of Local Organizing Committee & Co-Chair of the Program Committee: the 16th International Conference on Theory and Applications of Satisfiability Testing (SAT 2013), Helsinki, Finland, 2013.
- Local Organizer of 13th Scandinavian Symposium and Workshops on Algorithm Theory (SWAT 2012), Helsinki, Finland, 2012.
- Organizer of International SAT Competition 2013 & SAT Challenge 2012.
- Program Committee Member:
  the 14th International Workshop on Non-Monotonic Reasoning, Rome, Italy, 2012.
  the Sixth International Workshop on Stemmatology, 27–30 June 2012, University of Helsinki.
  he 1st Workshop on Combining Constraint Solving with Mining and Learning (CoCoMile 2012), Montpellier, France, 2012.
  2012 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2012), Bristol, UK, 2012.
  the 26th AAAI Conference on Artificial Intelligence, Toronto, Canada, 2012.
  the 3rd Workshop on Logics for Component Configuration, Budapest, Hungary, 2012.

the 3rd Workshop on Pragmatics of SAT (PoS 2012), Trento, Italy, 2012.
the 20th RCRA International Workshop on Experimental Evaluation of Algorithms for Solving Problems with Combinatorial Explosion (RCRA 2013), Rome, Italy.
the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013), Beijing, China.
the 2nd Workshop on Combining Constraint Solving with Mining and Learning (CoCoMiLe 2013), Bellevue, USA.
the 9th ICLP Doctoral Consortium (ICLP-DC 2013), Turkey.

- Steering Committee Member of SAT: International Conferences on Theory and Applications of Satisfiability Testing, 2013.

- Peer review of manuscripts:
  Reviewer for the 12th International Conference on Formal Methods in Computer-Aided Design (FMCAD 2012), UK, 2012.
  Reviewer for the 15th International Conference on Theory and Applications of Satisfiability Testing (SAT 2012), Italy, 2012.
  Reviewer for 18th International Conference on Logic for Programming, Artificial Intelligence and Reasoning (LPAR-18 / 2012), Bolivarian Republic of Venezuela, 2012.
  Reviewer for 18th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS 2012), 2012.
  Reviewer for 28th IEEE International Conference on Data Engineering (ICDE 2012), USA, 2012.
  Reviewer for 30th Symposium on Theoretical Aspects of Computer Science (STACS 2013), Germany, 2013.
  Artificial Intelligence Journal, 2012–2013.
  Journal on Satisfiability, Boolean Modeling and Computation, 2012.
  Journal of Experimental and Theoretical Artificial Intelligence, 2013.

Dr. Markus Koskela:

- Program Committee Member, SCIA 2013, Espoo, 17.-20.6.2013.

Doc. Jorma Laaksonen:

- Editorial Board Member, Pattern Recognition Letters, the Netherlands.

Dr. Brandon Malone

- Program Committee Member of International Joint Conference on Artificial Intelligence, 2013.

- Peer review of manuscripts:
  AAAI Conference, 2013.
  AI Communications, 2012.
  International Journal of Approximate Reasoning, 2013.
  Journal of Artificial Intelligence Research, 2012.
  Journal of Machine Learning Research, 2012.

Dr. Emilia Oikarinen:

- Program Committee Member:
  13th International Conference on Principles of Knowledge Representation and Reasoning (KR 2012), Rome, Italy, June 10-14, 2012.

5th Workshop on Answer Set Programming and Other Computing Paradigms (ASPOCP 2012), Budapest, Hungary, September 4, 2012.
6th Workshop on Answer Set Programming and Other Computing Paradigms (ASPOCP 2013), Istanbul, Turkey, August 25, 2013.

Doc. Kalle Palomäki:

- Program Committee Member: The 2nd 'CHiME' Speech Separation and Recognition Challenge, Vancouver, Canada, 1.6.2013.

Dr. Jaakko Peltonen:

- Program committee member:
  CIP 2012, 3rd International Workshop on Cognitive Information Processing, Parador de Baiona, Spain, 28.5.2012-30.5.2012.
  MLSP 2012, IEEE International Workshop on Machine Learning for Signal Processing, Santander, Spain, 23.9.2012-26.9.2012.
  ICANN 2012, International Conference on Artificial Neural Networks, Lausanne, Switzerland, 11.9.2012-14.9.2012.
  ISNN 2012, 10th International Symposium on Neural Networks, Dalian, China, 4.7.2012-6.7.2012.
  ICANN 2013, International Conference on Artificial Neural Networks, Sofia, Bulgaria, 10-13 September, 2013.
  NC2, New Challenges in Neural Computation workshop, Saarbrucken, Germany, 3 September, 2013.

- Session Chairman:
  Dagstuhl seminar on Information Visualization, Visual Data Mining and Machine Learning, Dagstuhl, Germany, 19.2.2012-24.2.2012.
  ACML 2013, Fifth Asian Conference on Machine Learning, Canberra, Australia, 13-15 November, 2013.

- Editorial Board Membership, Neural Processing Letters, Germany.

Doc. Tapani Raiko:

- Invited talk, Panelist in the Deep Learning panel dicussion, International Conference on Neural Information Processing, South Korea, 2013.

Dr. Jussi Rintanen:

- Program Committee Member:
  AAAI Conference on Artificial Intelligence, 2012.
  International Conference on Principles of Knowledge Representation and Reasoning, 2012.
  International Conference on Theory and Applications of Satisfiability Testing, 2012.
  European Conference on Logic in Artificial Intelligence JELIA, 2012.
  International Conference on Automated Planning and Scheduling, Italy, 2013.
  International Conference on Theory and Applications of Satisfiability Testing, 2013.

- Member of senior program committee, Australasian Joint Conference of Artificial Intelligence, 2012.

- International Conference on Automated Planning and Scheduling, member of council, USA.

Dr. Teemu Roos:

- Evaluation of applications for the Austrian Science Fund, Austrian Science Fund (FWF), 2013.
- Program Committee Co-Chair:
  6th International Workshop on Stemmatology, University of Helsinki, 27.6.-30.6.2012.
  Sixth Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-2013), Tokyo, Japan, 26.-29.8.2013.
- Program Committee Member:
  28th Conference on Uncertainty in Artificial Intelligence (UAI2012), USA, 15.-17.8.2012.
  European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, USA, 24.-28.9.2012.
  29th Conference on Uncertainty in Artificial Intelligence (UAI2013), USA, 2013.
  23rd International Joint Conference on Artificial Intelligence (IJCAI-13) China, 2013.
- Scientific Committee Member: Sixth Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-2013), Tokyo, Japan, 26.-29.8.2013.
- Keynote talk:
  Alan Turing Centenary Reception, Brittish Embassy, Helsinki, 12.11.2012.
  Phylomemetic and phylogenetic approaches in the humanities, Bern, 2012.
- Invited talk: 2012 Information Theory and Applications Workshop, San Diego
- Peer review of manuscripts:
  IEEE Communication Letters, 2013.
  IEEE Transactions on Information Theory, 2013.
  International Journal of Approximate Reasoning, 2013.
  PLOS One, 2012.
- Guest Editor of International Journal of Approximate Reasoning, Special Issue on Selected Papers from PGM-2010, 2012.
- Opponent at doctoral defense of Thomas Toftkjaer, Aarhus University, Denmark, 2012.

Dr. Zhirong Yang:

- Program Committee Member:
  IEEE/WIC/ACM International Conferences on Web Intelligence, Macau, China, 2012.
  International Conference on Artificial Neural Networks, Bulgaria, 2013.
  IEEE/WIC/ACM International Conference on Web Intelligence USA, 2013.

Important domestic positions of academic service by personnel of the unit

Professor Erik Aurell:

- Opponent at the doctoral dissertation of Hannes Wettig, University of Helsinki, 2013.

Professor Jukka Corander:

- Board member of the Finnish Doctoral Programme in Population Genetics.
- Board member of the Finnish Doctoral Programme in Stocastics and Statistics.
- Opponent at the doctoral defense of Markku Kohonen, University of Helsinki, 2013.
- Pre-examiner of a doctoral thesis
  Karri Seppä, University of Oulu, 2012.
  Marika Kaakinen, University of Oulu, 2012.

Professor Samuel Kaski:

- Opponent at the doctoral dissertation of Anna-Maria Lahesmaa-Korpinen, University of Helsinki.

Professor Mikko Kurimo:

- Opponent at the doctoral dissertation of Jani Nurminen, Tampere University of Technology.

Professor Petri Myllymäki:

- Director of the Doctoral Programme in Computer Science (DoCS), 2013–
- Domain Expert for the 2012 Millennium Award
- Advisory Committee of the Helsinki Doctoral Training Centre of the EIT ICT Labs Doctoral School, Board member.
- The Finnish Club in Helsinki, member 2012–
- The Finnish Academy of Technology, member 2013–

Professor Erkki Oja:

- Academy of Finland, Chairman of the Research Council for Natural Sciences and Engineering
- Academy of Finland, Chairman of the Steering Group for evaluation of physics in Finland
- Academy of Finland, Chairman of the Committee on national infrastructures
- Academy of Finland, Member of the Board
- Evaluating candidates for the professor's chair at the University of Vaasa, Finland
- Opponent: University of Oulu, Yimo Guo

Dr. Markus Koskela:

- President, Suomen Hahmontunnistuksen seura ry, Pattern Recognition Society of Finland, 2012.

Doc. Kalle Palomäki:

- Opponent at the doctoral dissertation of Toni Mäkinen, Tampere University of Technology.

Dr. Jaakko Peltonen:

- Helsinki Graduate School on Computational Science and Engineering, Advisor and board member.
- Opponent at the doctoral dissertation of Panu Luosto, University of Helsinki.

Doc. Tapani Raiko:

- Suomen Tekoälyseura (STeS), Finnish Artificial Intelligence Society, Board member.

Dr. Teemu Roos:

- Pre-examiner of the doctoral thesis of Tapio Manninen, Tampere University of Technology, 2013.

Research visits abroad by personnel of the unit

- Jukka Corander: Center for Communicable Disease Dynamics, Harvard School of Public Health, 1 month.
- Samuel Kaski: University College London, UK, 3 months.
- Mikko Kurimo: International Computer Science Institute, USA, 3 months.
- Tommi Mononen: Norwegian University of Science and Technology (NTU), 3 weeks.
- Elina Numminen: Sanger Institute, Cambridge, UK, 3 months.
- Kalle Palomäki: International Computer Science Instition, Berkeley, California, USA, 6 months.
- Jaakko Peltonen: Universite Catholique de Louvain, Belgium, 4 weeks.
- Alberto Pessia: Sanger Institute, Cambridge, 6 months.
- Tapani Raiko: University of Toronto, Canada, 1 month.
- Teemu Roos:
  Corpus Christi College, University of Cambridge, UK, 3.5 months.
  Finnish Institute in Rome, Italy, 2 months.
- Tommi Suvitaival: Doctoral studies in the University of Glasgow, 3 months.
- Seppo Virtanen: International Computer Science Institution, USA, 3 months.
- Jie Xiong: Department of Statistics, University of Oxford, Oxford, UK, 5 weeks.
- Zhirong Yang:
  University of Alberta, Canada, 5 months
  The Chinese University of Hong Kong, 4 weeks

Research visits by foreign researchers to the unit

- David Balding, professor, University College London, UK, 3 days. Research visit, 2012.
- Andrew Barron, professor, Yale University, USA, 1 month, 2013.
- Peter Brusilowsky, professor, University of Pittsburgh, USA, 3 months. Research visit, 2013.
- Colin Campbell, Dr, University of Bristol, 3 days. Research visit, 2012.
- Changyou Chen, MSc, NICTA, Australia, 8 days. Research visit, 2013.
- James Cussens, University of York, UK, 1 week. Research visit, 2013
- Ioannis Dimopoulos, PhD, University of Cyprus, Cyprus, 4 days. Research visit, 2013.
- Ralf Eggeling, MSc, Martin-Luther-Universität, Germany
  5.5 months. Research visit, 2012.
  5 months. Research visit, 2013.
- Dhananjaya Gowda, PhD, IIIT Hyderabad, India, 12 months. Research and teaching, 2013.
- Timo Koski, PhD, Kungliga tekniska h?gskolan, Sweden,
  2 months. Research visit, 2012.
  2 weeks. Research visit, 2013.
- Hideo Makino, professor, Niigata University, Japan, 3.5 months. Research visit, 2012-2013.
- Erik McDermott, PhD, Google, USA, 4 days. Research and teaching, 2013.
- Jose Moreno, BSc, Univ. Politécnica de Madrid, Spain, 3 months. Research and teaching, 2013.
- Brian Roark, Oregon Health and Science Univeristy, USA, 4 days. Research and teaching, 2013
- Simon Rogers, PhD, University of Glasgow, UK, 1 month. Visiting professor, 2012.
- Arturo Romero, BSc, Univ. Politécnica de Madrid, Spain, 3 months. Research and teaching, 2013.
- Johan Rung, PhD, EMBL-EBI European Bioinformatics Institute, UK, 3 days. Research and teaching, 2012.
- Murat Saraçlar, Boğaziçi University, Turkey, 4 days. Research and teaching, 2013
- Hannes Schulz, MSc, University of Bonn, Germany, 2 weeks. Research visit, 2012.
- Alan Smeaton, professor, Dublin City University, Ireland, 3 days. Research visit 2012.

- Kazuho Watanabe, PhD, Nara Institute of Science and Technology, Japan,
  1 month. Research visit, 2012.
  2 weeks. Research visit, 2013.
- Oliver Watts, University of Edinburgh, UK, 1 day. Research and teaching, 2013.
- Patric Wira, PhD, University of Haute Alsace, France, 1 day. Research visit, 2013.
- Eric Xing, professor, Carnegie Mellon University, 3 days. Research visit, 2012.
- Lan Yueheng, PhD, Tsinghua University, China, 2 months. Teaching visit, 2012.

Doctoral dissertations

Graphical models for biclustering and information retrieval in gene expression data

José Caldas

Doctoral dissertation for the degree of Doctor of Science in Technology on the 20th of April 2012

External examiners:
Guido Sanguinetti, D.Phil.
Sampsa Hautaniemi, D. Sc. (Tech.)
Opponent:
Eric Xing, PhD

Abstract:
The cell coordinates its biological response to the environment partly via the selective synthesis of thousands of unique RNA and protein molecules. Understanding the molecular biology of the cell is thus essential to the advancement of areas such as health care, agriculture, and energy production, but requires the ability to simultaneously acquire information about thousands of molecules in a sample. Recent high-throughput measurement technologies address this concern. While being useful, they generate a high volume of data and bring in methodological challenges, effectively shifting the bottleneck in molecular biology research from data acquisition to data analysis. In particular, an important challenge is the genome-wide analysis of how RNA is transcribed under different conditions, organisms, and tissues, a process known as gene expression.

When developing computational methods for biological data analysis tasks, probabilistic frameworks constitute promising approaches due to their flexibility, soundness, and ability to handle noisy data. In this thesis, the contributions are in the development of probabilistic methods for two relevant tasks in genome-wide gene expression analysis, namely biclustering and information retrieval.

Biclustering concerns the simultaneous grouping of objects, e.g., genes, and conditions. The first contribution is the development of a Bayesian extension to an existing biclustering model. The second contribution is a novel probabilistic method that allows deriving a hierarchical organization of microarrays in a gene expression data set and at the same time indicate the genes that characterize the hierarchy. Finally, the third contribution is a general probabilistic biclustering framework that easily lends itself to different data types and model assumptions.

Information retrieval in gene expression data is needed because of the increasing amount of available data stored in public databases. Two probabilistic methods for information retrieval are proposed. The models are used in a series of biological case studies that show how the proposed approaches have the potential to accelerate biological research by jointly analyzing data from different studies. In particular, several connections between biological conditions found by the models either correspond to existing biological knowledge or were used in a confirmatory follow-up study to obtain novel biological findings.

Visual category detection: an experimental perspective

Ville Viitaniemi

Doctoral dissertation for the degree of Doctor of Science in Technology on the 9th of May 2012

External examiners:
Dr. Gabriela Csurka, Xerox Research Centre Europe, France
Prof. Serkan Kiranyaz, Tampere University of Technology, Finland
Opponent:
Prof. Alan Smeaton, Dublin City University, Ireland

Abstract:
Nowadays huge volumes of digital visual data are constantly being produced and archived. Automatically distilling useful information from such information masses requires one to somehow answer the grand long-standing question of computer vision: how to make computers understand images?

In this thesis the visual content analysis problem is looked at as a category detection problem. In the category detection formulation, the general image content understanding task is partitioned into a number of small binary decision tasks. In each of the sub-tasks, one decides whether an image belongs to some pre-defined category. A category could be defined, for example, to consist of images taken indoors. By defining an appropriate set of categories, the visual content of an image can be described on a desired level of granularity by determining the image's membership in each one of the categories.

This thesis discusses a framework for visual category detection that consists of three major components: feature extraction, feature-wise detection and fusion of the detection results. The point of view in the discussion is empirical: the framework is validated by the good levels of performance systems implementing it have demonstrated in various benchmark tasks of visual analysis. A body of experiments is described that compare various technological alternatives for implementing the three major components of the framework. In addition to comparing implementation techniques, the experiments demonstrate that the discussed generic category detection architecture is very versatile: a set of diverse visual analysis problems can be addressed using the same visual category detection system as a backbone component by equipping the system with a task-specific front-end.

From the experiments and discussion in the thesis, one can conclude that the category detection formulation is a useful way of approaching the general image content understanding problem. In category detection, performances reaching the state-of-the-art can be realised using the presented fusion-based system architecture and implementation technologies of the system components.

Morph-based speech retrieval: Indexing methods and evaluations of unsupervised morphological analysis

Ville Turunen

Doctoral dissertation for the degree of Doctor of Science in Technology on the 24th of August 2012

External examiners:
Dr. Gareth J. F. Jones, Dublin City University, Ireland
Prof. Kalervo Järvelin, University of Tampere, Finland
Opponent:
Prof. Murat Salaçlar, Boğaziçi University, Turkey

Abstract:
Speech retrieval enables users to find information in collections of spoken material. Automatic speech recognition (ASR) is used to transform the spoken words into text, and information retrieval (IR) methods are used for searching. Traditional ASR systems have a predefined vocabulary of words, and any word that is out-of-vocabulary (OOV) can not be recognized. Typically, rare words are excluded, which is problematic for retrieval, because query words are often rare words such as proper names. The limited vocabulary is especially problematic for languages such as Finnish that have a very large number of distinct word forms.

In this thesis, morpheme-like subword units are used for speech recognition and retrieval. The subword units, referred to as morphs, are discovered using a data driven method that learns morphological structure from text data. Using this approach, it is possible to recognize any word in speech, even a word that was not in the training data, as a sequence of morphs. A rule-based morphological analyzer could be used to find base forms of the recognized words for indexing. However, the vocabulary of the analyzer is also limited, and recognition errors cause further problems for the analyzer. Instead, in this work, morphs are used as index terms as well.

In Finnish speech retrieval experiments, the morph-based approach is compared to using word-based language models in ASR, and to using base forms in retrieval. Also, morphs are compared for story segmentation of speech. The results show that morph-based language models clearly outperform word-based models in retrieval performance. As index terms, using morphs is about as efficient as using base forms, but combining the two approaches is better than either alone, especially when there are a high proportion of unseen words in the queries. The effect of unoptimal morph segmentations is reduced by using alternative morph segmentations of query words and by using latent semantic indexing.

Even if the morph deemed most likely by the ASR is incorrect, it is possible that the correct one is among the candidates the ASR considers. Utilizing the candidates in retrieval can improve performance. In this thesis, a representation of ASR hypotheses called confusion network is used for extracting alternative recognition results. A rank-based weighting of index terms is proposed, and found to outperform posterior probability based weighting.

This thesis also studies evaluation metrics for unsupervised morphological analysis methods. Application evaluations such as speech retrieval are time consuming and cannot be used during method development. Different linguistic evaluation metrics have been proposed and are compared in this thesis by e.g. correlating the metrics to the results of application performance.

Multivariate multi-way modelling of multiple high-dimensional data sources

Ilkka Huopaniemi

Doctoral dissertation for the degree of Doctor of Science in Technology on the 12th of October 2012

External examiners:
Prof. Antti Penttinen, University of Jyväskylä
Dr. Simon Rogers, University of Glasgow, United Kingdom
Opponent:
Dr. Colin Campbell, University of Bristol, United Kingdom

Abstract:
A widely employed strategy in current biomedical research is to study samples from patients using high-throughput measurement techniques, such as transcriptomics, proteomics, and metabolomics. In contrast to the static information obtained from the DNA sequence, these techniques deliver a "dynamic fingerprint" describing the phenotypic status of the patient in the form of absolute or relative concentrations of hundreds, or even tens of thousands of molecules: mRNA, proteins, metabolites and lipids. The huge number of variables measured opens up new possibilities for biomedical research; harnessing the information contained in such 'omics' data requires advanced data analysis methods.

The standard setup in biomedical research is comparing case (diseased) and control (healthy) samples and determining differentially expressed molecules that are then considered potential bio-markers for disease. In modern biomedical experiments, more complicated research questions are common. For instance, diet or drug treatments, gender and age play central roles in many case-control experiments and the measurements are often in the form of a time-series. Due to these additional covariates, the experimental setting becomes a multi-way experimental design, but few tools for proper data-analysis of high-dimensional data with such a design exist. Moreover, the task of integrating multiple data sources with different variables is nowadays often encountered in two classes of biomedical experiments: (i) Multiple omics types or samples from several tissues are measured from each patient (paired samples), (ii) Translating biomarkers between human studies and model organisms (no paired samples). These data integration tasks usually additionally involve a multi-way experimental design.

In this dissertation, a novel Bayesian machine learning model for multi-way modelling of data from such multi-way, single-source or multi-source setups is presented, covering the majority of situations commonly encountered in statistical analysis of omics data coming from current biomedical research. The problem of high dimensionality is solved by assuming that the data can be described as highly correlated groups of variables. The Bayesian modelling approach involves training a single, unified, interpretable model to explain all the data. This approach can overcome the main difficulties in omics analysis: small sample-size and high dimensionality, multicollinearity of data, and the problem of multiple testing. This approach also enables rigorous uncertainty estimation, dimensionality reduction and easy interpretability of results from a complex setup involving multiple covariates and multiple data sources.

Learning constructions of natural language: Statistical models and evaluations

Sami Virpioja

Doctoral dissertation for the degree of Doctor of Science in Technology on the 10th of December 2012

External examiners:
Doc. Krister Lindén, University of Helsinki, Finland
Prof. Richard Wicentowski, Swarthmore College, USA
Opponents:
Doc. Krister Lindén, University of Helsinki, Finland
Prof. Brian Roark, Oregon Health & Science University, USA

Abstract:
The modern, statistical approach to natural language processing relies on using machine learning techniques on the increasing amount of text and speech data in electronic format. Typical applications for statistical methods include information retrieval, speech recognition, and machine translation. Many problems encountered in the applications can be solved without language-dependent resources, such as annotated data sets, by the means of unsupervised learning. This thesis focuses on one such problem: the selection of lexical units. It is the first step in processing text data, preceding, for example, the estimation of language models or extraction of vectorial representations. While the lexical units are often selected using simple heuristics or grammatical rule-based methods, this thesis proposes the use of unsupervised and semi-supervised machine learning. Advantages of the data-driven unit selection include greater flexibility and independence from the linguistic resources that exist for a particular language and domain.

Statistically learned lexical units do not always fit to the categories in traditional linguistic theories. In this thesis, they are called constructions according to construction grammars, a family of usage-based, cognitive theories of grammar. For learning constructions of a language, the thesis builds on Morfessor, an unsupervised statistical method for morphological segmentation. Morfessor is successfully extended to the tasks of learning allomorphs, semi-supervised learning of morphological segmentation, and learning phrasal constructions of sentences. The results are competitive especially for the morphology induction problems. The thesis also includes new techniques for using the sub-word constructions learned by Morfessor in statistical language modeling and machine translation. In addition to its usefulness in the applications, Morfessor is shown to have psycholinguistic competence: its probability estimates have high correlations with human reaction times in a lexical decision task.

Furthermore, direct evaluation methods for the unit selection and other learning problems are considered. Direct evaluations, such as comparing the output of the algorithm to existing linguistic annotations, are often quicker and simpler than indirect evaluation via the end-user applications. However, with unsupervised algorithms, the comparison to the reference data is not always straightforward. In this thesis, direct evaluation methods are developed for two unsupervised tasks, morphology induction and learning semantic vector representations of documents. In both cases, the challenge is to find relationships between the pairs of features in multidimensional data. The proposed methods are quick to use and they can accurately predict the performance in different applications.

Probabilistic, information-theoretic models for etymological alignment

Hannes Wettig

Doctoral dissertation for the degree of Doctor of Philosophy on the 9th of February 2013

External examiners:
Steven de Rooij, Centrum Wiskunde & Informatica (CWI), Amsterdam, Netherlands
Timo Honkela, Aalto University School of Science, Finland
Opponent:
Erik Aurell, Skolan för dataveteskap och kommunikation (KTH), Sweden and Aalto University, Finland

Abstract:
This thesis starts out by reviewing Bayesian reasoning and Bayesian network models. We present results related to discriminative learning of Bayesian network parameters. Along the way, we explicitly identify a number of problems arising in Bayesian model class selection. This leads us to information theory and, more speci cally, the minimum description length (MDL) principle. We look at its theoretic foundations and practical implications. The MDL approach provides elegant solutions for the problem of model class selection and enables us to objectively compare any set of models, regardless of their parametric structure. Finally, we apply these methods to problems arising in computational etymology. We develop model families for the task of sound-by-sound alignment across kindred languages. Fed with linguistic data in the form of cognate sets, our methods provide information about the correspondence of sounds, as well as the history and ancestral structure of a language family. As a running example we take the family of Uralic languages.

Towards efficient and robust automatic speech recognition: Decoding techniques and discriminative training

Janne Pylkkönen

Doctoral dissertation for the degree of Doctor of Science in Technology on the 22nd of March 2013.

External examiners:
Dr. Thomas Hain, University of Sheffield, UK
Doc. Tuomas Virtanen, Tampere University of Technology, Finland

Opponent:
Dr. Erik McDermott, Google, Inc., USA

Abstract:
Automatic speech recognition has been widely studied and is already being applied in everyday use. Nevertheless, the recognition performance is still a bottleneck in many practical applications of large vocabulary continuous speech recognition. Either the recognition speed is not sufficient, or the errors in the recognition result limit the applications. This thesis studies two aspects of speech recognition, decoding and training of acoustic models, to improve speech recognition performance in different conditions.

A major part of this thesis studies discriminative training of acoustic models. The emphasis is on the most popular algorithm for discriminative model estimation, the extended Baum-Welch algorithm. The thesis points out theoretical connections of the algorithm to general constrained optimization. It also proposes new control methods for the algorithm, which are shown to improve the robustness of the acoustic models in several large vocabulary speech recognition tasks. Discriminative training methods are widely applied in the state-of-the-art speech recognizers which utilize the prevalent hidden Markov models for acoustic modeling. Therefore the proposed methods have many immediate practical applications.

The speech recognition system developed at the Aalto university was utilized and significantly improved during the research of this thesis. The thesis gives an overview of that system and describes the decoder of the system in more detail. In speech recognition systems, the decoder combines the information from the statistical models of acoustics and language to implement the search for the word sequence which best matches the input speech. The thesis proposes new methods for improving the speed of this search, without incurring losses to the recognition accuracy.

Statistical models for inferring the structure and history of populations from genetic data

Jukka Sirén

Doctoral dissertation for the degree of Doctor of Philosophy on the 23rd of April 2013

External examiners:
Professor Pekka Pamilo, Department of Biosciences, University of Helsinki, Finland
Professor Ziheng Yang, Department of Genetics, Evolution and Environment, University College London, UK
Opponent:
Professor David Balding, UCL Genetics Institute, University College London, UK

Abstract:
Population genetics has enjoyed a long and rich tradition of applying mathematical, computational and statistical methods. The connection between these fields has deepened in the last few decades as advances in genotyping technology have led to an exponential increase in the amount of genetic data allowing fundamental questions involving the nature of genetic variation to be asked. The massive quantities of data have necessitated the development of new mathematical and statistical models along with computational techniques to provide answers to these questions.

In this work we address two problems in population genetics by constructing statistical models and analyzing their performance with simulated and real data. The first one concerns the identification of genetic structure in natural populations from molecular data, which is an important aspect in many fields of applied science, including genetic association mapping and conservation biology. We frame it as a problem of clustering and classification and utilize background information to achieve a higher accuracy, when the genetic data is sparse. We develop a computationally efficient method for taking advantage of geographical sampling locations of the individuals. The method is based on the assumption that the spatial structure of the populations correlates strongly with the genetic structure, which has been proven reasonable for human populations.

In the assignment of individuals into known populations, we also show how improvements in the efficiency of the inference can be obtained by considering all of the individuals jointly. The result is derived in the context of classification, which is major field of study in machine learning and statistics, making it applicable in a wide range of situations outside population genetics.

The other problem involves the reconstruction of evolutionary processes that have resulted in the structure present in current populations. The genetic variation between populations is caused to large extent by genetic drift, which corresponds to random fluctuations in the distribution of a genetic type due to demographic processes. Depending on the genetic marker under study, mutation has only a minor or even negligible role, in contrast with traditional phylogenetic methods, where mutational processes dominate as the time scales are longer. We follow the change in the relative frequencies of different genetic types in populations by deriving approximations to widely used models in population genetics. The direct modeling of population level properties allows the method to be applied data sets harboring thousands of samples, as demonstrated by the analysis of global population structure of Streptococcus pneumoniae.

Learning mental states from biosignals

Melih Kandemir

Doctoral dissertation for the degree of Doctor of Science in Technology on the 4th of May 2013

External examiners:
Dr. Päivi Majaranta, University of Tampere, Finland
Dr. David Roi Hardoon, SAS, Singapore
Opponent:
Dr. Cristina Conati, University of British Columbia, Canada

Abstract:
As computing technology evolves, users perform more complex tasks with computers. Hence, users expect from user interfaces to be more proactive than reactive. A proactive interface should anticipate the user's intentions and take the right action without requiring a user command. The crucial first step for such an interface is to infer the user's mental state, which gives important cues about user intentions. This thesis consists of several case studies on inferring mental states of computer users. Biosensing technology provides a variety of hardware tools for measuring several aspects of human physiology, which is correlated with emotions and mental processes. However, signals gathered with biosensors are notoriously noisy. The mainstream approach to overcome this noise is either to increase the signal precision by expensive and stationary sensors or to control the experiment setups more heavily. Both of these solutions undermine the usability of the developed methods in real-life user interfaces.

In this thesis, machine learning is used as an alternative strategy for handling the biosignal noise in mental state inference. Computer users have been monitored under loosely controlled experiment setups by cheap and inaccurate biosensors, and novel machine learning models that infer mental states such as affective state, mental workload, relevance of a real-world object, and auditory attention are built.

The methodological contributions of the thesis are mainly on multi-view learning and multitask learning. Multi-view learning is used for integrating signals of multiple biosensors and the stimuli. Multitask learning is used for inferring multiple mental states at once, and for exploiting the inter-subject similarities for higher prediction accuracy. A novel multitask learning algorithm that transfers knowledge across multi-view learning tasks is introduced. Another novelty is a Bayesian factor analyzer with a time-dependent latent space that captures the dynamic nature of biosignals better than methods that assume independent samples. The overall outcome of the thesis is that it is feasible to predict mental states from unobtrusive biosensors with reasonable accuracy using state-of-the-art machine learning models.

Inference of relevance for proactive information retrieval

Antti Ajanki

Doctoral dissertation for the degree of Doctor of Science in Technology on the 27th of September 2013

External examiners:
Prof. Markku Tukiainen (University of Eastern Finland)
Dr. Emilio Parrado-Hernandez (Univ. Carlos III de Madrid, Spain)
Opponent:
Prof. Kari-Jouko Räihä, University of Tampere

Abstract:
Search engines have become very important as the amount of digital data has grown dramatically. The most common search interfaces require one to describe an information need using a small number of search terms, but that is not feasible in all situations. Expressing a complex query as precise search terms is often difficult. In the future, better search engines can anticipate user's goals and provide relevant results automatically, without the need to specify search queries in detail.

Machine learning methods are important building blocks in constructing more intelligent search engines. Methods can be trained to predict which documents are relevant for the searcher. The prediction is based on recorded feedback or observations of how the user interacts with the search engine and result documents. If the relevance can be estimated reliably, interesting documents can be retrieved and displayed automatically.

This thesis studies machine learning methods for information retrieval and new kinds of applications enabled by them. The thesis introduces relevance inference methods for estimating query terms from eye movement patterns during reading and for combining relevance feedback given on multiple connected data domains, such as images and their captions. Furthermore, a novel retrieval application for accessing contextually relevant information in the real world surroundings through augmented reality data glasses is presented, and a search interface that provides browsing cues by making potentially relevant items more salient is introduced.

Prototype versions of the proposed methods and applications have been implemented and tested in simulation and user studies. The tests show that these methods often help the searcher to locate the right items faster than traditional keyword search interfaces would.

The experimental results demonstrate that, by developing custom machine learning methods, it is possible to infer intent from feedback and retrieve relevant material proactively. In the future, applications based on similar methods have the potential to make finding relevant information easier in many application areas.

Bayesian methods in bacterial population genomics

Lu Cheng

Doctoral dissertation for the degree of Doctor of Philosophy on the 4th of October 2013

External examiners:
Professor Daniel Thorburn, Stockholm University, Sweden
Professor Tanel Tenson, University of Tartu, Estonia
Opponent:
Associate Professor Zhaohui Steve Qin, Emory University, USA

Abstract:
Vast amounts of molecular data are being generated every day. However, how to properly harness the data remains often a challenge for many biologists. Firstly, due to the typical large dimension of the molecular data, analyses can either require exhaustive amounts of computer memory or be very time-consuming, or both. Secondly, biological problems often have their own special features, which put demand on specially designed software to obtain meaningful results from statistical analyses without imposing too much requirements on the available computing resources. Finally, the general complexity of many biological research questions necessitates joint use of many different methods, which requires a considerable expertise in properly understanding the possibilities and limitations of the analysis tools.

In the first part of this thesis, we discuss three general Bayesian classification/clustering frameworks, which in the considered applications are targeted towards clustering of DNA sequence data, in particular in the context of bacterial population genomics and evolutionary epidemiology. Based on more generic Bayesian concepts, we have developed several statistical tools for analyzing DNA sequence data in bacterial metagenomics and population genomics.

In the second part of this thesis, we focus on discussing how to reconstruct bacterial evolutionary history from a combination of whole genome sequences and a number of core genes for which a large set of samples are available. A major problem is that for many bacterial species horizontal gene transfer of DNA, which is often termed as recombination, is relatively frequent and the recombined fragments within genome sequences have a tendency to severely distort the phylogenetic inferences. To obtain computationally viable solutions in practice for a majority of currently emerging genome data sets, it is necessary to divide the problem into parts and use different approaches in combination to perform the whole analysis. We demonstrate this strategy by application to two challenging data sets in the context of evolutionary epidemiology and show that biologically significant conclusions can be drawn by shedding light into the complex patterns of relatedness among strains of bacteria. Both studied organisms (*Escherichia coli* and *Campylobacter jejuni*) are major pathogens of humans and understanding the mechanisms behind the evolution of their populations is of vital importance for human health.

Theses

Master of Science in Technology

2012

Ammad-ud-din, Mohammad
Modelling drug response in cancer cells by combining multiple drug and cell line views

Hyyrynen, Lasse
Towards multilingual speech recognition; Decreasing foreign word error rate

Luostarinen, Tapio
Content-based recommender system exploiting topic models

Nguyen, Mai
Preferential Optimization of University Students' Timetables

Murphy, David
Mobile devices as platforms for sensor based AR

Nallamothu, Kranthi
Decoding of dynamic visual scenes from gaze features

Noeva, Polina
Sampling methods for missing value reconstruction

Padilla Arias, José David
Coupling acoustic and language models in speech recognition in noisy conditions

Rajaraman, Sitaram
Computational analysis of heart transcriptome in genetically modified rats

Sem, Federico
Automatic segmentation of nucleus accumbens

Sun, Tao
Parameter selection in translation-based approaches to answer set solving

Turunen, Tuomas
Utilizing three-dimensional models in operation and maintenance of a powerplant

Uziela, Karolis
Making microarray and RNA-seq gene expression data comparable

Vatanen, Tommi

Missing Value Imputation Using Subspace Methods with Applications on Survey Data

Xu, Jiang
Modeling dynamics of influenza with antigenic change

2013

Cao, Yang
Multi-variate analysis of brain images in seasonal affective disorder

Dong, Siyuan
A time dependent adaptive learning process for estimating drug exposure from register data - applied to insulin and its analogues

Golumbeanu, Monica
Statistical analysis of PAR-CLIP data

Gulzar, Kashif
Wireless evaluation tools; A simulator for testing wireless sensor network systems

Judin, Jussi
A data encoding approach to video communication system verification

Kanto, Jaakko
Data mining of product data

Klapuri, Jussa
Collaborative filtering methods on a very sparse reddit recommendation dataset

Lebre Magalhães Pereira, João Pedro
Supervised learning for relationship extraction from textual documents

Leppäaho, Eemeli
Transfer Learning with Group Factor Analysis

Malmi, Eric
Human mobility prediction; A probabilistic transfer learning approach

Outamaa, Ville-Petteri
Visualization of preregistered multimodal medical images; Implementing a viewing environment

Remes, Sami
Extending group factor analysis to model events in MEG data

Sandholm, Max
Information Retrieval Perspective to Interactive Data Visualization

Tirunagari, Santosh
Mining causal relations from maritime accident investigation reports

Master of Science thesis

2012

Debarshi, Ray
Data Gathering in Digital Homes

Gauriot, Romain
Statistical challenges in the quantification of gunshot residue evidence

Hiltunen, Suvi
Minimum Description Length Modeling of Etymological Data

Kosunen, Ilkka
Clustering Psychophysiological Data with Mixtures of Generalized Linear Models

Lübbers, Henning
Effects of dynamic parameter discovery on lossless compression algorithms

2013

Benner, Christian
Bayesian confirmatory factor analysis for detection of differential gene expression

Huttunen, Jyri-Petteri
Neuroevolution methods for real-time combining of expert neural networks

Konyushkova, Ksenia
ImSe: Instant Interactive Image Retrieval System with Exploration/Exploitation trade-off

Kuosmanen, Anna
Comparison of RNA-seq data analysis software

Wang, Ziran
Applying component models on theme-based news tracking and detection

Xu, Lilu
WikiMage: a web game based on Wikipedia

Licentiate Theses

2012

Enarvi, Seppo
Finnish Language Speech Recognition for Dental Health Care

Siltanen, Sanni
Theory and applications of marker-based augmented reality

2013

Järvinen, Paula
A data model based approach for visual analytics of monitoring data

Research Projects

Introduction

Erkki Oja, Director of COIN

The Finnish Centre of Excellence in Computational Inference Research (COIN) started in January 2012 as a joint effort between the Department of Information and Computer Science at Aalto University, which is acting as the coordinator, and two units at the University of Helsinki: the Department of Computer Science and the Department of Mathematics and Statistics.

COIN has the overall objective to forge and deliver methods and tools which can transform large quantities of raw data from many kinds of sources into useful information, in batch mode and in on-line mode. The core of the solution is large-scale data-intensive computational modeling and inference: how to model the data to infer what is relevant in the vast data masses.

The last decade has seen an unprecedented acceleration in the amount and richness of data collected, stored, analyzed, and acted upon in all spheres of human activity. In addition to text-based information, there is a rapid increase of the accessible data stored in video, audio, or other numerical formats, which calls for efficient indexing and retrieval techniques to exploit these potentially extremely valuable information sources. The central aim of such work is to transform vast quantities of numbers into information which is comprehensible to humans and can be usefully employed.

COIN addresses two main application problems. The first one is intelligent information access through proactive user interfaces, in which a user gets seamless access to relevant information in a real-world context. The other application field is computational molecular biology and medicine, in which massive databases of experiments and knowledge are used to enhance diagnosis, personalized treatment, and discovery of biological mechanisms. We have chosen these two application fields because of our proven expertise in their core problems, because of their strategic importance in the near future, and because of their methodological interrelations. In each of these two fields, a *Flagship project* has been built within COIN, that directs and consolidates the methodological research and facilitates collaborations with external domain experts and enterprises.

The central research problem in both our Flagship projects is how to best access and use relevant information buried in massive databanks and input data streams. The information content varies, from personal images, videos and files in contextual information retrieval scenarios to genome-wide biomedical measurements in computational biology and medicine. Both applications need models of the data to define what is relevant: which sets, samples, and variables. This setup brings up four *Core methodological grand challenges*.

The first one, in which we have a particularly strong track-record, is how to learn effective models from massive data. The second one, motivated by the need to analyse massive set of datasets in the flagships, is multi-source machine learning. The third challenge is how to do the modeling when the models are very complex in the sense of being structured by prior information and constraints. Fourth, in particular in contextual information retrieval it is crucially important to get access to the information instantly and all the time, requiring on-line learning of the models and extremely rapid inference of what is relevant in the current context.

In addition to the two flagship domains, the usability of the methodological research results are tested in a spectrum of carefully chosen collaborative projects in other domains. The partners are other national Centres of Excellence or equivalent top-level groups in Finland or abroad, as well as industry.

The research is mainly theoretical, but heavy experimentation for testing the algorithms is a central part of the work, especially in the two Flagships. The data used in the experiments is obtained from various sources: from our own user experiments, from public-domain sources, from the research and industrial collaborators of COIN, and from the previous and ongoing scientific projects of COIN's research groups. The data is stored permanently and made available for COIN's researchers from the computer systems of the two participating universities that are equipped with enough disk capacity to hold the required massive data sets and also serve the parts of data made public to the outsiders. Both universities have recently-installed, massive computing clusters available for the project: the Department of Computer Science at University of Helsinki has the third most powerful supercomputer cluster in Finland with 1940 cores, and the Faculty of Information and Natural Sciences at Aalto University has a similar cluster with 1344 cores.

Organizationally, the COIN Centre of Excellence consists of six interrelated research groups: the two Flagships F1 and F2, and the four Core challenges C1, C2, C3, and C4 according to Figure 1. In the matrix, '○' means that the PI (row) is in charge of the coordination of the challenge or flagship (column), while '×' signifies active collaboration. Thus, coherence between the groups comes naturally from the synergy of the research topics, fostered by active informal discussions.

	COIN	C1	C2	C3	C4	F1	F2
E. Oja	○	×				×	
E. Aurell	×	×			×		○
J. Corander	×	×	×	○			×
S. Kaski	×	×	○	×		×	×
J. Laaksonen	×	×	×			○	
P. Myllymäki	×	○	×		×	×	
I. Niemelä	×	×		×	○	×	

Figure 1: *The organization of the COIN Centre of Excellence*

Each group in 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 2012 - 2013 details the individual research projects of the six groups during the first two years of the six-year period of COIN. Additional information including demos etc. is available from our Web pages, `www.research.ics.aalto.fi/coin/`.

Chapter 1

C1: Learning models from massive data

Petri Myllymäki, Matti Järvisalo, Samuel Kaski, Markus Koskela, Jorma Laaksonen, Erkki Oja, Jaakko Peltonen, Jorma Rissanen, Teemu Roos, Jeremias Berg, Kerstin Bunte, Xi Chen, Onur Dikmen, Dorota Głowacka, Mehmet Gönen, Tele Hao, Satoru Ishikawa, Ziyuan Lin, Zhiyun Lu, Brandon Malone, Mats Sjöberg, Zhirong Yang, He Zhang, Zhanxing Zhu

1.1 Introduction

As noted in the research plan, automated collection of data takes nowadays routinely place in increasingly many problem domains, and machine learning techniques have a lot of new potential application areas. The goal of challenge area C1 is to address the methodological problems arising in particular "Big Data" type of settings, and develop methods for scaling up data-driven machine learning. The "scaling up" problem can be understood and approached in various ways: one approach is to take standard machine learning settings, e.g., the problem of learning the structure of graphical models, and study how to solve the problem *optimally* (in a sense that is defined in the setting) as efficiently as possible. In this setting we have combined the expertise of several COIN groups and made good progress in applying state-of-the-art constraint reasoning techniques for solving the stated machine learning problem (results of this research area are discussed in Section 4.4). We have also developed methods in this setting that utilize standard *heuristic search* methods that still provably produce the optimal solution (see Section 1.2).

In addition to model structure learning, two other popular machine learning tasks we have focused on are data clustering and data visualization, and we summarize our progress in these areas in Sections 1.3 and 1.4, respectively. In both settings, most our approaches are based on approximations that lead to theoretically justifiable, yet computationally efficient methods that extend the scope of domains where this type of methods can be used. On the other hand, we have also considered both in the clustering and in the visualization case exact methods, based on constraint reasoning techniques, that lead to provably optional solutions. The approaches in these cases are based on conceptually similar ideas as the constraint satisfaction techniques of Section 4.4, developed for graphical model structure learning.

Another issue raised in the research plan is that in many new problem domains no prior information about the modeled phenomena can be found, in which case one has to resort to non-informative, purely data-driven machine learning methods that can construct models based on observational data alone. Section 1.5 briefly describes some of our new results in this area. Most of the approaches are based on approximations of theoretically optimal, but computationally infeasible theoretical approaches, and some of the theoretical frameworks developed are sequential/incremental in nature, which enables methods that can handle streaming data and scale machine learning up in this sense.

Another viewpoint in scaling up machine learning is to take into account the specific needs of certain important application areas, and here we should of course primarily consider our Flagship Areas F1 and F2. The machine learning tasks discussed above — structure learning, clustering, visualization — are already all well motivated by the needs of these Flagship Areas, but in Section 1.6 we take a more specific view and summarize our progress in methodological modeling work that targets the specific needs of Flagship Area F1: the data in this context is typically massive in nature, consisting of e.g. huge volumes of media data, and what is more, the learning settings are often dynamic in nature, meaning that new data is not fixed but keeps streaming in, so that we need to develop online algorithms that can process the data continuously, and traditional batch type of machine learning methods are not directly applicable. Section 1.6 discusses briefly our methodological solutions that have been motivated by the specific nature of data sets emerging in F1, while later in Chapter 5 we take a broader view and a holistic approach to discuss our progress in Flagship Area F1 as a whole.

1.2 Speeding up learning

In the heuristic search formulation for learning Bayesian network structures, the learning problem is cast as a graph search problem. The costs of paths correspond to scores of networks, and the shortest path gives the globally optimal Bayesian network. In the worst case, the size of the state space is exponential; however, a *heuristic function* is used to bound the quality of paths. A search strategy, such as best-first, depth-first branch and bound, etc., is then guided by the heuristic function to explore the space. Unpromising regions of the space are either implicitly or explicitly pruned based on the heuristic and search strategy.

A recent journal paper [11] collected algorithms and results from previous conference papers [12, 10] which gave the basic shortest-path formulation, heuristic functions and a best-first search algorithm for learning optimal structures. Other studies have investigated efficient algorithms for depth-first search [6], parallel search [5] and anytime, best-first variants with bounded optimality guarantees [7]. In particular, the anytime, best-first algorithms are often able to find optimal solutions very quickly (within a few seconds) and spend the rest of the time proving optimality. In ongoing research, we have further increased the scalability of learning by constructing much tighter heuristics [3] and exploiting implicit information in the learning problem input [2] .

In addition to improving the heuristic search solver for learning Bayesian networks, we have also developed machine learning techniques to predict the running time of several solvers (namely, heuristic search and integer linear programming [4]) on a particular learning problem instance [1]. The best solver for the given instance is then used. Empirical results have shown that the solver performances are often orthogonal, so selecting the best algorithm on a per-instance basis leads to large performance gains compared to using only a single solver on all instances. We have also applied similar techniques for selecting dynamic loop scheduling algorithms in large, heterogeneous computing environments [8, 9]. Similar performance improvements were observed. At the same time we are also exploring more conventional ways to parallelize our machine algorithms, either in a computing cluster or a GPU processor (see also Section 1.6).

References

[1] B. Malone, J. K. Kangas, M. Järvisalo, M. Koivisto, and P. Myllymäki. Predicting the Hardness of Learning Bayesian Networks. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*, AAAI Press, 2014 (to appear).

[2] X. Fan, B. Malone and C. Yuan. There IS a free lunch: constraints for learning Bayesian networks. Unpublished manuscript, 2014.

[3] X. Fan, C. Yuan and B. Malone. Tightening bounds for Bayesian network structure learning. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2014 (to appear).

[4] M. Bartlett and J. Cussens. Advances in Bayesian network learning using integer programming. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, 2013.

[5] B. Malone and C. Yuan. A bounded error, anytime parallel algorithm for exact Bayesian network structure learning. In *Proceedings of the 6th European Workshop on Probabilistic Graphical Models*, 2012.

[6] B. Malone and C. Yuan. A depth-first branch and bound algorithm for learning optimal Bayesian networks. In *3rd International Workshop on Graph Structures for Knowledge Representation and Reasoning*, 2013.

[7] B. Malone and C. Yuan. Evaluating anytime algorithms for learning optimal Bayesian networks. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, 2013.

[8] S. Srivastava, B. Malone, N. Sukhija, I. Banicescu, and F. M. Ciorba. Predicting the flexibility of dynamic loop scheduling using an artificial neural network. In *Proceedings of the 12th International Symposium on Parallel and Distributed Computing*, 2013.

[9] N. Sukhija, B. Malone, S. Srivastava, I. Banicescu, and F. M. Ciorba. Portfolio-based selection of robust dynamic loop scheduling algorithms using machine learning. In *Proceedings of the 3rd International Workshop on Parallel and Distributed Computing for Large Scale Machine Learning and Big Data Analytics*, 2014.

[10] C. Yuan and B. Malone. An improved admissible heuristic for finding optimal Bayesian networks. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2012.

[11] C. Yuan and B. Malone. Learning optimal Bayesian networks: A shortest path perspective. *Journal of Artificial Intelligence Research*, 48:23–65, 2013.

[12] C. Yuan, B. Malone, and X. Wu. Learning optimal Bayesian networks using A* search. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 2011.

1.3 Clustering

In data-driven machine learning, there are basically two kinds of tasks: supervised and unsupervised learning. The central approach in unsupervised learning is cluster analysis: to divide the available data into clusters, such that the data items within a given cluster are similar and those between the different clusters are dissimilar. Cluster analysis is basically a combinatorial optimization problem: each of the data items must be placed in exactly one of the clusters. In [1] we presented for certain type of clustering problems a MaxSAT-based algorithm that provably leads to the global optimum. To simplify the computations further, various relaxation techniques have been proposed. The technique of Nonnegative Matrix Factorization (NMF) has earlier proven to be an effective method for finding part-based representation of data, i.e. grouping features [2]. However, with the data matrix rotated, the grouping can be done on the data samples, i.e. giving answers for cluster analysis. NMF relaxes the hard-clustering problem which is often NP-hard.

Recently we have presented two types of clustering methods based on nonnegative matrix decomposition. First, we have proposed the Quadratic Nonnegative Matrix Factorization (QNMF) [6], where some of the factorizing matrices can appear twice in the approximation. An example is to approximately factorize the similarity matrix S into the product two low-rank nonnegative matrices: $S \approx WW^T$, where W indicates the cluster assignments. Conventionally the approximation is measured by Least Square Error (LSE). However, we have shown that LSE is not suitable for the original similarity matrix S and we should replace S with its smoothed version, e.g. by graph random walk [4] (see Figure 1.1). Our new method, called NMFR, has significantly improved the state-of-the-art in terms of cluster purity, especially for large datasets that situate in curved manifolds.

Second, we have proposed a matrix decomposition beyond factorization which is based on the Data-Cluster-Data random walk (DCD) [5]. The DCD learning objective fulfills the following requirements: (1) approximation error measure that takes into account sparse similarities, (2) decomposition form of the approximating matrix, where the decomposing matrices should contain just enough parameters for clustering but not more, and (3) normalization of the approximating matrix, which ensures relatively balanced clusters and equal contribution of each data sample. We have empirically shown that DCD can achieve excellent cluster purity given suitable initializations [5]. We have also studied a generalized divergence family for clustering tasks [8].

The new clustering objectives demand appropriate optimization algorithms. We have presented a unified development method for a wide variety of QNMF problems [6]. For NMD with the stochasticity constraint, we have proposed two ways for developing multiplicative update rules, by reparameterization and by Lagrangian relaxation [10]. Moreover, we have also improved the efficiency [9] and scalability [7] of the multiplicative optimization algorithms.

Another important problem in NMD is to determine the approximation error measure, i.e. to select the best divergence in a parameterized family. We have presented pioneering work in this direction, by using approximative Tweedie distribution and score matching [3].

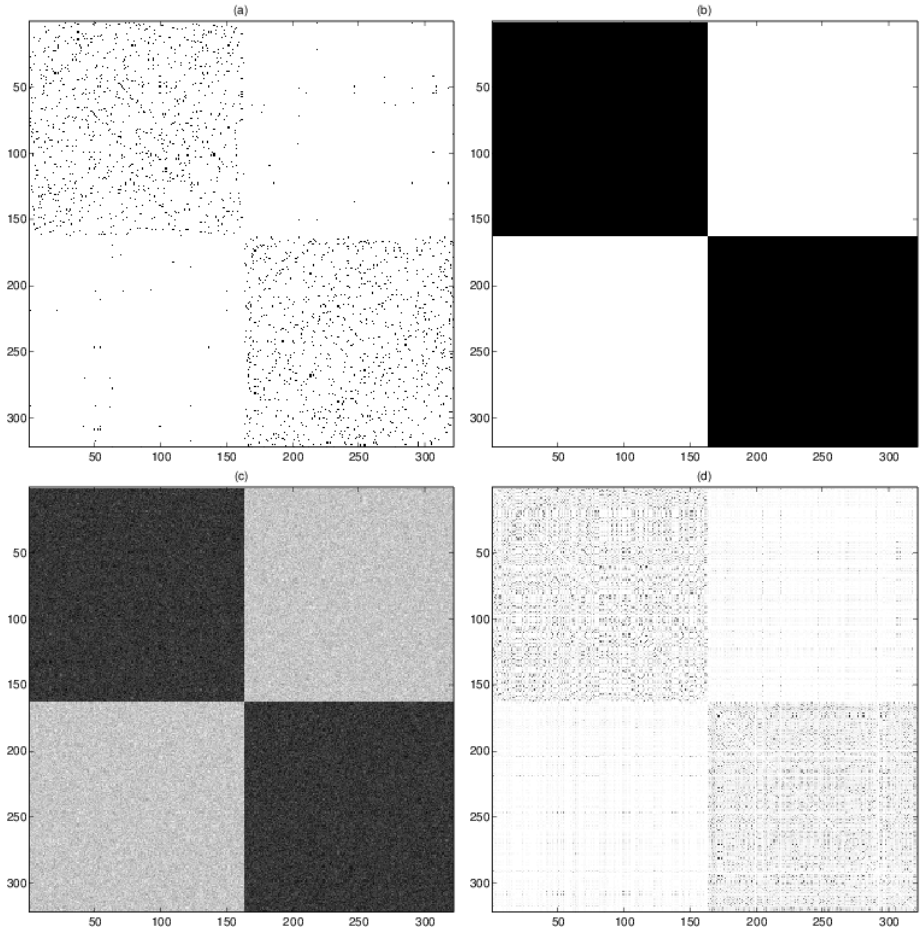


Figure 1.1: Illustration of clustering the *SEMEION* handwritten digit dataset by NMF based on LSE [4]: (a) the symmetrized 5-NN graph, (b) the correct clusters to be found, (c) the ideally assumed data that suits the least square error, (d) the smoothed input by using graph random walk. The matrix entries are visualized as image pixels. Darker pixels represent higher similarities. For clarity we show only the subset of digits "2" and "3". In this paper we show that because (d) is "closer" to (c) than (a), it is easier to find correct clusters using (d)≈(b) instead of (a)≈(b) by NMF with LSE.

References

[1] Jeremias Berg and Matti Järvisalo. Optimal Correlation Clustering via MaxSAT. In Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW 2013), IEEE Press, 2013.

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

[3] Zhiyun Lu, Zhirong Yang, and Erkki Oja. Selecting β-divergence for nonnegative matrix factorization by score matching. In *Proceedings of 22nd International Conference on Artificial Neural Networks (ICANN 2012)*, pages 419–426, Lausanne, Switzerland, 2012.

[4] Zhirong Yang, Tele Hao, Onur Dikmen, Xi Chen, and Erkki Oja. Clustering by nonnegative matrix factorization using graph random walk. In *Advances in Neural*

Information Processing Systems 25 (NIPS2012), pages 1088–1096, Lake Tahoe, USA, 2012.

[5] Zhirong Yang and Erkki Oja. Clustering by low-rank doubly stochastic matrix decomposition. In *Proceedings of 29th International Conference on Machine Learning (ICML2012)*, pages 831–838, Edinbourgh, United Kingdom, 2012.

[6] Zhirong Yang and Erkki Oja. Quadratic nonnegative matrix factorization. *Pattern Recognition*, 45(4):1500–1510, 2012.

[7] Zhirong Yang, He Zhang, and Erkki Oja. Online projective nonnegative matrix factorization for large datasets. In *Proceedings of 19th International Conference on Neural Information Processing (ICONIP 2012)*, pages 285–290, Doha, Qatar, 2012.

[8] He Zhang, Tele Hao, Zhirong Yang, and Erkki Oja. Pairwise clustering with t-PLSI. In *Proceedings of 22nd International Conference on Artificial Neural Networks (ICANN 2012)*, pages 411–418, Lausanne, Switzerland, 2012.

[9] He Zhang, Zhirong Yang, and Erkki Oja. Adaptive multiplicative updates for projective nonnegative matrix factorization. In *Proceedings of 19th International Conference on Neural Information Processing (ICONIP 2012)*, pages 277–284, Doha, Qatar, 2012.

[10] Zhanxing Zhu, Zhirong Yang, and Erkki Oja. Multiplicative updates for learning with stochastic matrices. In *Proceedings of the 18th conference Scandinavian Conferences on Image Analysis (SCIA 2013)*, pages 143–152, Espoo, Finland, 2013.

1.4 Data visualization

Visualization is a key part of data analysis especially in the first stages when strong hypotheses and models are not yet available. In visualization, we develop modeling-driven methods for looking at the data. We have focused in particular on our recent method NeRV [6] which formalizes data visualization as retrieval of relevant data points from a display, but most results are more general. We have developed methods for scaling up the visualizations to larger data sets [3, 4, 7], to more accurate visualizations [1], to principled combinations of linear dimensionality reduction and linear supervised learning [2], and to meta-visualization [5].

Neighbor embedding (NE) methods have found their use in data visualization but are limited in big data analysis tasks due to their quadratic complexity with respect to the number of data samples n. We have pursued three approaches towards fast visualization of large data sets. Firstly, in [3] we introduced an efficient learning algorithm where relationships between data are approximated through mixture modeling, yielding efficient computation with near-linear computational complexity with respect to the number of data. Secondly, in [4] we introduced a multiplicative update rule that converges much faster than the original additive gradient based optimization algorithm while yielding equally good results. Thirdly, in [7], we demonstrated that the obvious approach of subsampling produces inferior results, and proposed a generic approximated optimization technique that reduces the NE optimization cost to $O(n \log n)$. The technique is based on realizing that in visualization the embedding space is necessarily very low-dimensional (2D or 3D), and hence efficient approximations developed for n-body force calculations can be

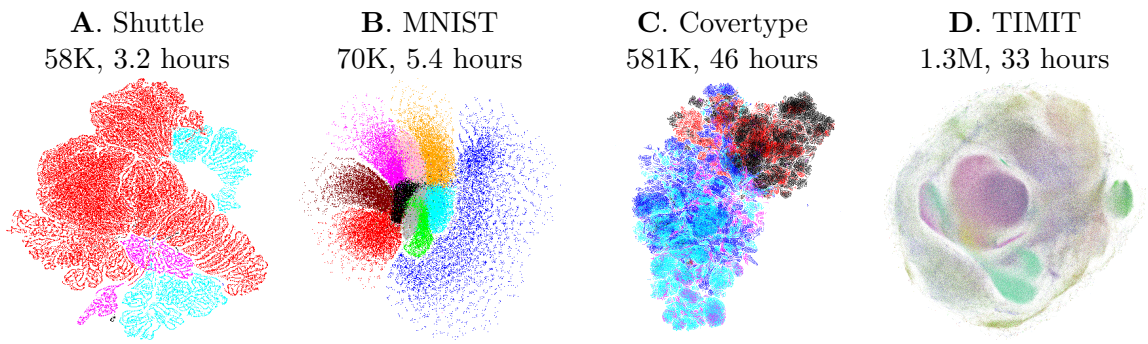


Figure 1.2: Visualization of large-scale datasets made feasible with the new approximations in [7]. A. UCI Shuttle using a new fast implementation of the SNE (Stochastic Neighbor Embedding) visualization method and spectral direction (SD) optimization; B. MNIST using the NeRV method and SD; C. UCI Covertype using SNE and SD; and D. TIMIT using the SNE method and momentum. Titles of subfigures show the dataset name, dataset size, and the learning time.

applied. In gradient-based NE algorithms the gradient for an individual point decomposes into "forces" exerted by the other points. The contributions of close-by points need to be computed individually but far-away points can be approximated by their "center of mass", rapidly computable by applying a recursive decomposition of the visualization space into quadrants. The new algorithm brings a significant speed-up for medium-size data, and brings big data within reach of visualization.

Current neighbor embedding methods need complicated nonlinear optimization approaches that reach only local optima of their cost functions, and thus may not yield as good visualizations as would have been possible. In an ongoing yet-unpublished work [1], we present a novel approach to low-dimensional neighbor embedding for visualization, based on formulating an information retrieval based neighborhood preservation cost function as *Maximum satisfiability* on a discretized output display. The method has a rigorous interpretation as optimal visualization based on the cost function. Unlike previous low-dimensional neighbor embedding methods, our formulation is guaranteed to yield globally optimal visualizations, and does so reasonably fast. Unlike previous manifold learning methods yielding global optima of their cost functions, our cost function and method are designed for low-dimensional visualization where evaluation and minimization of visualization errors are crucial. Our method performs well in experiments, yielding clean embeddings of datasets where a state-of-the-art comparison method yields poor arrangements. In a real-world case study for semi-supervised WLAN signal mapping in buildings we outperform state-of-the-art methods, as shown in Figure 1.3.

The methods discussed above have been unsupervised. In supervised tasks dimensionality reduction has commonly been used as a preprocessing step before training a supervised learner; however, coupled training of dimensionality reduction and supervised learning steps may improve the prediction performance. In [2], we introduce a simple novel Bayesian *supervised dimensionality reduction* method that combines linear dimensionality reduction and linear supervised multiclass classification in a principled way. Classification experiments on three benchmark data sets show that the new model significantly outperforms seven baseline linear dimensionality reduction algorithms on very low dimensions in terms of generalization performance on test data. The proposed model also obtains the best

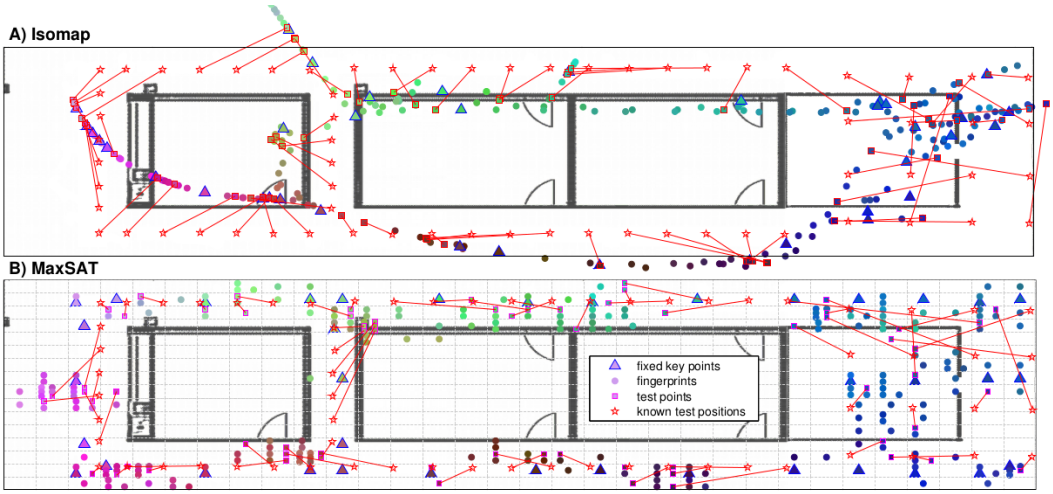


Figure 1.3: Visualizations of WLAN radio mapping, where the goal is to map ("visualize") radio signal vectors (RSSI) on a 2D plane. Most popular WLAN positioning solutions today perform this mapping task in a time-consuming manual fashion. We compare two methods that can use easy-to-collect pure radio signal data without manual data calibration: a recent automated manifold learning approach, 2-stage Isomap (A), and our proposed maximum satisfiability based neighborhood preserving visualization, MaxSAT on a 16×64 grid (B). Dots represent the 200 fingerprint vectors to be mapped (RSSI vectors whose true positions are unknown during training and testing), blue triangles represent the 38 given key point positions (RSSI vectors whose true positions are known during training), pink squares represent the 66 mapped test points. Similar RSSI vectors are colored similarly. Red stars are the recorded geographical positions of the test points; lines connect the mapped and recorded positions. Our method outperforms Isomap since our method maps the test points closer to their true locations.

results on an image recognition task in terms of classification and retrieval performances.

Lastly, in visual data exploration with scatter plots, no single plot is sufficient to analyze complicated high-dimensional data sets. In [5] we point out that given numerous visualizations created with different features or methods, *meta-visualization* is needed to analyze the visualizations together. We solve how to arrange numerous visualizations onto a meta-visualization display, so that their similarities and differences can be analyzed. We introduce a machine learning approach to optimize the meta-visualization, based on an information retrieval perspective: two visualizations are similar if the analyst would retrieve similar neighborhoods between data samples from either visualization. Based on the approach, we introduce a nonlinear embedding method for meta-visualization: it optimizes locations of visualizations on a display, so that visualizations giving similar information about data are close to each other.

References

[1] Kerstin Bunte, Matti Järvisalo, Jeremias Berg, Petri Myllymäki, Jaakko Peltonen, and Samuel Kaski. Optimal neighborhood preserving visualization by maximum satisfiability. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*, AAAI Press, 2014.

[2] Mehmet Gönen. Bayesian supervised dimensionality reduction. *IEEE Transactions on Cybernetics*, 43(6):2179–2189, 2013.

[3] Jaakko Peltonen and Konstantinos Georgatzis. Efficient optimization for data visualization as an information retrieval task. In Ignacio Santamaría, Jerónimo Arenas-García, Gustavo Camps-Valls, Deniz Erdogmus, Fernando Pérez-Cruz, and Jan Larsen, editors, *Proceedings of MLSP 2012, the 2012 IEEE International Workshop on Machine Learning for Signal Processing*, page electronic proceedings, Piscataway, NJ, 2012. IEEE.

[4] Jaakko Peltonen and Ziyuan Lin. Multiplicative update for fast optimization of information retrieval based neighbor embedding. In Saeid Sanei, Paris Smaragdis, Asoke Nandi, Anthony TS Ho, and Jan Larsen, editors, *Proceedings of MLSP 2013, the 2013 IEEE International Workshop on Machine Learning for Signal Processing*, page electronic proceedings, Piscataway, NJ, 2012. IEEE.

[5] Jaakko Peltonen and Ziyuan Lin. Information retrieval perspective to meta-visualization. In *Proceedings of ACML 2013, Fifth Asian Conference on Machine Learning*, JMLR W&CP, volume 29, pages 165–180, 2013. JMLR.

[6] 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.

[7] Zhirong Yang, Jaakko Peltonen, and Samuel Kaski. Scalable optimization of neighbor embedding for visualization. In *Proceedings of ICML 2013, the 30th International Conference on Machine Learning*, volume 28 of *JMLR W&CP*, pages 127–135. JMLR, 2013.

1.5 Data-driven machine learning

Bayesian techniques have become extremely popular in many areas owing in part to their ability to flexibly combine different sources of information, including prior knowledge. However, in some situations no reliable prior knowledge is available in a form that is amenable to efficient computation. In such cases a Bayesian procedure may be sensitive to arbitrary choice of priors, potentially leading to unsatisfactory outcomes. Therefore, it is of great utility to develop data-driven or priorless techniques that are robust in the sense that they can achieve nearly as good performance as the best available strategy with hind-sight (knowing the outcome in advance). The Minimum Description Lengh (MDL) principle is a powerful approach for developing such techniques. Modern MDL is based on the so called normalized maximum likelihood (NML) distribution, which involves no prior distributions. The theory of modern MDL is still going through major transformations, as summarized in the recent monograph [4].

The exact computation of NML is feasible only in a handful of models, and approximations are required in most situations. The important multinomial case is one of the few cases where an exact NML solution is available. In two papers [2, 3], we have shown that various histogram-like models can be reduced to the multinomial and other simpler cases

so that efficient exact solutions are applicable. Similar reductions to the multinomial case were shown to lead to approximative model selection criteria in [1]. These criteria demonstrate the robustness of the MDL-based criteria: in the experiments, they outperformed their Bayesian counterparts even when the parameters of the Bayesian methods were tuned for the best possible performance. Similar observations were made in an empirical evaluation of Bayesian network learning learning functions [8]. Yet another angle to the multinomial case is provided by studying nearly optimal approximations that, unlike to exactly optimal NML distribution, do not require that the total sample size is known in advance. In [5], we characterize the achievability of asymptotic minimax optimality without sample size dependence. Such considerations are important from the point of view of online learning and streaming data where predictions are needed before the length of the sequence is known.

On the other hand in [7], we construct Bayes-like mixture codes where the prior is automatically optimized so as to match the NML distribution exactly. Once such an 'NML prior' is constructed, conditional and marginal distributions required for prediction can be obtained in linear time compared to the earlier exponential-time algorithms. Moreover, a Bayes-like representation of NML provides an interesting contrast between MDL and Bayesian methods in certain cases where the NML prior involves *negative* weights.

Approximations should always be used with care. Even when the approximation error is guaranteed to asymptotically vanish, it may have practically devastating consequences. This was demonstrated to be the case concerning certain refined versions of the Bayesian information criterion (BIC), which completely break down for Markov chain models of moderate to high complexity [6]. Further study will extend these results to other practically relevant model classes and aim at providing approximations that are more safe to use.

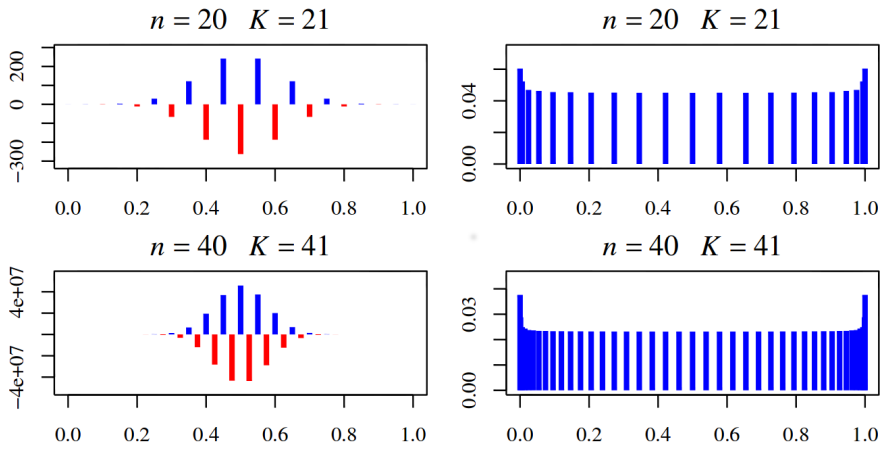


Figure 1.4: Discrete prior distributions for the Bernoulli model with sequence length $n = 20$ (top) and $n = 40$ (bottom) with $K = 21$ and $K = 41$ support points, respectively. Under these priors, the resulting mixture model agrees with the normalized maximum likelihood (NML) distribution. For uniformly spaced support points (left), the prior includes some negative weights whereas using a theoretically justified non-uniform spacing leads to all-positive weights. Image source: [7].

References

[1] R. Eggeling, T. Roos, P. Myllymäki, and I. Grosse. Robust learning of inhomogeneous PMMs. To appear in *Proc. 17th Conf on Artificial Intelligence and Statistics (AISTATS-2014)*, 2014.

[2] C.D. Giurcaneanu, P. Luosto and P. Kontkanen. On The Performance Of Histogram-Based Entropy Estimators. In *Proceedings of the 2012 IEEE International Workshop on Machine Learning for Signal Processing (MLSP'2012), September 2012*, Santander, Spain, 2012.

[3] P. Luosto, C.D. Giurcaneanu and P. Kontkanen. Construction of irregular histograms by penalized maximum likelihood: a comparative study. *IEEE Information Theory Workshop 2012 (ITW)*, Lausanne, Switzerland, 3-7 September 2012.

[4] J. Rissanen. *Optimal Estimation of Parameters*, Cambridge University Press, 2012.

[5] K. Watanabe, T. Roos, and P. Myllymäki. Achievability of asymptotic minimax regret in online and batch prediction. In *Proc. 5th Asian Conference on Machine Learning (ACML-2013)*, 2013.

[6] T. Roos and Y. Zou. Keep it simple stupid - On the effect of lower-order terms in BIC-like criteria. Invited paper in *Proc. 2013 Information Theory and Applications Workshop, (ITA-2013)*, 2013.

[7] A. Barron, T. Roos, and K. Watanabe. Bayesian properties of normalized maximum likelihood and its fast computation. Accepted to *IEEE International Symposium on Information Theory (ISIT-2014)*, arXiv:1401.7116.

[8] Z. Liu, B. Malone, and C. Yuan. Empirical evaluation of scoring functions for Bayesian network model selection. *BMC Bioinformatics*, 13(Suppl 15):S14, 2012.

1.6 Models for Intelligent Information Access

As discussed in the Introduction, the Flagship Area F1 poses several difficult problems from the modeling point of view. First of all, the data is typically big, and in particular digital audiovisual data represent the most rapidly increasing form of digital data today. Media data are created with mobile phones and often stored in huge cloud-based image services and social networks. Other sources of audiovisual data include streaming media such as broadcast radio and television programs. In Finland, the major national players in media archiving are the Finnish Broadcasting Company YLE and the National Audiovisual Institute KAVI. For example, YLE has notable historic media archives containing 1,000,000 still images, 200,000 hours of audio and 200,000 hours of video. Collaboration with both YLE and KAVI has been active in COIN's research projects.

Many emerging application areas in video and image processing require large-scale visual concept detection. Visual concept detection facilitates high-level querying of audiovisual data by organizing the database in terms of mid-level concepts such as objects, persons, locations or events. Current state-of-the-art automatic concept detectors rely on bag-of-visual-words histograms of local features and computationally heavy kernel-based classifiers such as non-linear SVMs. There are, however, many scenarios today that require extremely fast classification, preferably real-time or better for a typically large number (hundreds or thousands) of concepts.

Fast linear classifiers have been shown to achieve competitive performance on very high-dimensional problems such as in document classification. However, they cannot typically match the performance of non-linear SVMs in vision problems, in which the dimensionalities are not as high (typically in the thousands). One approach to improve accuracy of linear classifiers is to use linear approximations for additive non-linear kernels, such as intersection or χ^2, as the approximations provide results very similar to the original non-linear kernels, but require only a fraction of the detection time. We have studied speeding up visual concept detection while preserving the classification accuracy [3, 4] and applied the technique in large-scale evaluations [5, 6]. Our results show that it is possible to reach the performance of non-linear additive kernels on large video databases with computational requirements on the level of linear classifiers using either homogeneous kernel maps [7] or the power mean SVM [8].

Another modeling challenge with the information retrieval systems is that they often utilize various kinds of metadata, such as captions and tags, but it is not always possible to tag new documents or images in a new dataset quickly and efficiently enough to meet the requirements of an information retrieval system. To address these issues, in recent years there have been many attempts to incorporate the user feedback into a search session directly, in an online manner. However, systems of this kind often face usability issues: in order to incorporate the user feedback, the system needs to perform complex optimizations which typically take a long time to execute. Thus, the resulting system is either unrealistically slow, not allowing real-time feedback, or it is limited to be functional only with a small amount of data. We approach the problem by developing fast on-line learning techniques [1] that can scale up to realistic data sizes while maintaining fast response times. We are currently also studying how to utilize parallelization in this context, and are developing a fast implementation of a number of Gaussian Process algorithms on CUDA GPUs, and are also considering other parallelization techniques used earlier in other contexts [2].

References

[1] K. Konyushkova and D. Głowacka, Content-Based Image Retrieval with Hierarchical Gaussian Process Bandits with Self-Organizing Maps. Pp. 267–272 in Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2013.

[2] Medlar, A. and Głowacka, D. and Stanescu, H. and Bryson, K. and Kleta, R., SwiftLink: Parallel MCMC linkage analysis utilising multicore CPU and GPU. Bioinformatics 29 (2013), 420 – 427.

[3] Mats Sjöberg, Markus Koskela, Satoru Ishikawa, and Jorma Laaksonen. Real-time large-scale visual concept detection with linear classifiers. In *Proceedings of 21st International Conference on Pattern Recognition*, Tsukuba, Japan, November 2012.

[4] Mats Sjöberg, Markus Koskela, Satoru Ishikawa, and Jorma Laaksonen. Large-scale visual concept detection with explicit kernel maps and power mean SVM. In *Proceedings of ACM International Conference on Multimedia Retrieval (ICMR2013)*, pages 239–246, Dallas, Texas, USA, April 2013. ACM.

[5] Mats Sjöberg, Satoru Ishikawa, Markus Koskela, Jorma Laaksonen, and Erkki Oja. PicSOM experiments in TRECVID 2012. In *Proceedings of the TRECVID 2012 Workshop*, Gaithersburg, MD, USA, November 2012.

[6] Satoru Ishikawa, Markus Koskela, Mats Sjöberg, Jorma Laaksonen, Erkki Oja, Ehsan Amid, Kalle Palomäki, Annamaria Mesaros, and Mikko Kurimo. PicSOM experiments in TRECVID 2013. In *Proceedings of the TRECVID 2013 Workshop*, Gaithersburg, MD, USA, November 2013.

[7] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2010)*, 2010.

[8] Jianxin Wu. Power mean SVM for large scale visual classification. In *Proceedings of The IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, Providence, USA, June 2012.

Chapter 2

C2: Learning from multiple data sources

Samuel Kaski, Arto Klami, Markus Koskela, Jorma Laaksonen, Jaakko Peltonen, Ehsan Amid, Muhammad Ammad-ud-din, José Caldas, Ritabrata Dutta, Ali Faisal, Elisabeth Georgii, Jussi Gillberg, Mehmet Gönen, Melih Kandemir, Suleiman A. Khan, Eemeli Leppäaho, Pekka Marttinen, Anna Maria Mesaros, Kristian Nybo, Juuso Parkkinen, Sami Remes, Sohan Seth, Tommi Suvitaival, Seppo Virtanen,

2.1 Introduction

Big data is not only big amounts of homogeneous data, but also massive collections of heterogenenous data sources, or in practice data sets. As the size of the collection of data sets grows, finding and analysing the relationships between the sets becomes more and more important. We have developed methods for *multi-view learning*, where the observations are shared across the data sets even though the variables are different, and generalizations where some of the observations or variables are shared. A new line of work is *data set search*, to find sets useful when analysing a new set.

2.2 Unsupervised multi-view learning

A central application for unsupervised multi-view settings is data integration or fusion. Given co-occurring observations from multiple data sources, the task is to provide a latent representation that combines joint variation (statistical dependencies) between the views. In this section, we present new Bayesian latent variable models and inference methods for learning dependencies between the views, and apply developed models to solving problems in bioinformatics, neuroinformatics and intelligent information access.

Canonical correlation analysis (CCA) is a classical method for seeking dependencies between two data sources. Bayesian models and inference methods provide a solid framework for building hierarchical extensions of CCA [17] and for coping with uncertainty, especially with large dimensionalities and small sample sizes. We have introduced a new Bayesian solution for the CCA problem [17] that is more robust and efficient than earlier approaches. In particular, the new model advances the state of the art enabling applications that were previously not possible to solve.

Factor analysis is a traditional technique for capturing dependencies between multiple univariate variables. We extend this formulation to groups of variables (that is, views), presenting a novel problem formulation called group factor analysis (GFA) [28]. The task GFA solves is to find factors that capture dependencies between any subset of the views. This task generalizes Bayesian CCA for more than two views in a flexible way. We have introduced an efficient Bayesian inference method for GFA.

GFA assumes views with co-occurring observations. However, in many applications feature spaces also co-occur and each view can be represented as a tensor. We recently extend the problem formulation of GFA to multi-way tensors [13].

Applications in neuroinformatics. A very attractive application area for unsupervised multi-view learning is *neuroinformatics*. The developed models and inference methods, described above, are particularly relevant for coping with extremely noisy data. In [18], we succeeded in decoding natural speech that the subjects heard in an MEG scanner. Usually decoding is done with simplified stimuli. In [23], we inferred the roles of two subjects in a novel brain imaging experiment—one where brain measurements for two subjects, interacting with Skype, were recorded at the same time. We also proposed a method for automatically selecting the most relevant voxels in fMRI data [20]. The proposed method helps in improving the weak signal to noise ratio of fMRI experiments.

Applications for intelligent information access. In [10], we inferred user's auditory attention from multiple bio-sensors, including EEG, using Bayesian extensions of CCA. The model extends (mixtures of) Bayesian CCA by introducing dynamics for the latent signals. In [27], we capture dependencies between web images and co-occurring text documents, combining Bayesian CCA-type factor models with topic modeling that is needed to represent text data. For that, we presented a novel non-parametric Bayesian Hierarchical Dirichlet Process-based multi-view topic model.

Applications in bioinformatics. In [12], we addressed a novel multi-view cancer drug response problem with sparse extension of GFA finding previously unknown relationships. In [26], we addressed data translation/ANOVA problem with sparse GFA. Other data integration works can be found in F2.

2.3 Beyond multi-view learning

The methods described in the previous section provide justified solutions for multi-view setups with co-occurring observations in all views. While many data integration tasks match such setups, there are also several problems that come with richer data and hence require more flexible models. As one core direction of C2 we have worked towards extending the multi-view learning models for such setups.

The first extension considers multi-view setups with no known co-occurrence. Given two views with arbitrary features, the task is to learn which object in one view corresponds to which in the other, based on the observed data alone. We presented first a variational Bayesian solution to this problem in [15], receiving the best-paper award of the ACML'12 conference for our work, and then a Gibbs-sampling based solution in [16] that outperforms both the earlier variational solution and all the competing approaches for learning the correspondence. Besides providing the leading method for solving the cross-domain object matching problem, these works contribute to Bayesian inference over permutations and hence relate to inference for structured models studied in C3.

The second extension generalizes multi-view learning methods for setups with arbitrary collections of matrices with some shared entities, illustrated in Figure 2.1. Besides multiple views for the same objects, we can have, for example, additional representatios for the features in some of the views. The task of learning joint factorization for all of the data sources is called *collective matrix factorization* (CMF). Even though the setup is much more flexible than multi-view learning, it turns out that the group-wise sparse multi-view method of [28] can be extended also for learning the shared and private factors for such arbitrary collections. The resulting *group-sparse CMF* method [14] also supports non-Gaussian likelihoods and has efficient variational inference for large matrices with missing data.

2.4 Supervised multi-view learning and multi-task learning

Multiple kernel learning methods [6] have been successful in integrating several data sources in supervised learning tasks. In large p small n domains, of large dimensionality p compared to small sample size n, which are common in computational biology and

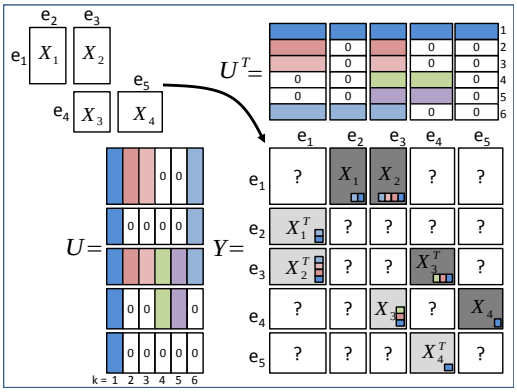


Figure 2.1: The top left corner illustrates a complex setup of related matrices that goes beyond multi-view learning; views X_1 and X_2 share the same row entities, whereas views X_2 and X_3 share the same column entites, etc. The rest of the plot shows how *group-sparse collective matrix factorization* learns joint low-rank factorizations for all of the views, identifying both factors shared by multiple views (factors 1 and 6) as well as factors private to each view (factors 2-5). The figure is taken from [14].

medicine (F2), it is important to flexibly control and assess uncertainty of the solutions. Bayesian solutions based on Gaussian processes would be natural choices but are computationally demanding, and we have developed efficient Bayesian multiple kernel learning methods [7]. The methods were generalized to matrix factorization [8] given multiple side information sources, both empirically outperfoming existing methods. The result is effectively a flexible, Bayesian non-linear recommender engine which can use the side information sources to make out-of-matrix predictions. The method is directly applicable to personalized medicine prediction problems of F2.

In addition to personalized medicine, the multiple kernel learning are widely applicable to data integration tasks in other fields. We have successfully applied them in estimating user's state and relevance for intelligent information access tasks of F1, for instance in [11].

A central reason why the new multiple kernel learning methods have been successful in personalized medicine is multi-task learning, that they have been used to predict for multiple drugs simultaneously, sharing some but not all of the parameters across the drugs. This strategy is only useful if the tasks are sufficiently related, otherwise "negative transfer" may even decrease the performance. For asymmetric multi-task learning, where the other tasks are only important to the extent they are useful for modelling one "task of interest", we have developed a price-winning (ECML2011 best paper award) Gaussian process-based solution [19].

Another scenario that can be seen as multi-task learning is planning under uncertainty for multiple agents, by optimizing action policies that take both the short-term and long-term consequences of actions into account. Such optimization is computationally hard, and efficient solutions are needed to scale up such planning to larger planning problems and a richer variety of planning situations. The planning is often based on decentralized partially observable Markov decision processes (Dec-POMDPs), but current methods must de-emphasize long-term effects of actions by a discount factor. In tasks like wireless networking, agents are evaluated by average performance over time, both short and long-

term effects of actions are crucial, and discounting based solutions can perform poorly. In [22] we show that under a common set of conditions expectation maximization (EM) for average reward Dec-POMDPs is stuck in a local optimum. We introduce a new average reward EM method; it outperforms a state of the art discounted-reward Dec-POMDP method in experiments. In [21] we consider the wireless networking domain, where medium access control (MAC) determines which devices (agents) can access the wireless channel at each time. The MAC performance depends on spatial locations and traffic patterns of agents. We propose a Dec-POMDP based MAC solution that adapts to the spatial and temporal opportunities, is able to handle network dynamics described by a Markov model, and takes both sensor noise and partial observations into account. The method yields MAC policies optimized for the network dynamics model, given a freely chosen goal such as maximal throughput or minimal latency. We make approximate optimization efficient by exploiting factorization of problem structure in the Dec-POMDP, yielding compact policy representations. Our approach yields higher throughput and lower latency than CSMA/CA based comparison methods.

2.5 Retrieval of experiments

We study the task of retrieving relevant experiments given a query experiment, that incorporates measurement information in the retrieval task, instead of relying only on annotations. This line of work is motivated by the opportunity and need to relate new measurement data sets of experimental sciences to earlier research, piloted in F2 on computational biology and medicine. The methods are more general, however, and by an experiment we here mean essentially a data set. Depending on what we can assume about the commonality of the sets, we need slightly different types retrieval methods.

In the first works [2, 3] the idea of modelling-driven retrieval was introduced to molecular biology. The experiments were reduced to a set of data samples, and the whole data collection was modelled with a big probabilistic model, which gave a principled distance measure for the retrieval. By changing the model the retrieval system can be made more targeted [5].

This mode of operation has two weaknesses when scaling up: (1) computation of one big joint model of all the experiments becomes prohibitive, and (2) managing the practical effort of including all prior knowledge of all experiments in the joint model becomes hard. We have started studying the alternative scenario (Fig. 2.2) where the experimenters deposit their models of their data to public repositories, hence storing both the prior knowledge and the new evidence in the data.

Given a collection of fixed models, the problem of relating a new (query) data set to them can be formulated as a mixture modelling problem. The models which get large mixture weights are relevant to the new data [4]. This problem can be solved efficiently with linear or even sublinear complexity for maximum likelihood estimates of the mixture weights.

Moving to Bayesian modelling capturing even more of the experimenter's prior knowledge, we build a "supermodel" that models the set of earlier models and operates on their posterior distributions. A general and practical way of representing the posterior distributions as sets of MCMC samples. The supermodel needs to be computationally light and learnable sequentially. We have used Particle-Learning-based sequential Dirichlet process mixtures (DPM) for this purpose [24], and derive the relevance measure for retrieval from

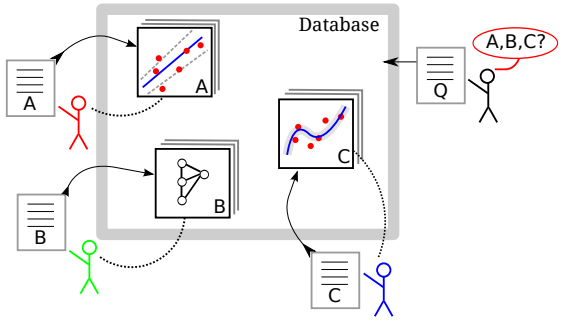


Figure 2.2: The setup of retrieval of models of experiments. Our general objective is to retrieve experiments A, B or C, given query experiment Q. We assume that the database contains models of experiments learned by the experimenters along with the experimental details. Each model is represented in terms of posterior samples.

the mixture representation. We have recently also extended the work to models coming from different model families, and learning to reduce the number of stored MCMC samples for computational efficiency [25]. We have demonstrated the efficacy of our approaches on simulated data with linear regression, Bayesian lasso for sparse linear regression, logistic regression and factor model as the models, and on real world molecular biological datasets.

2.6 Learning from multimodal media data

Digital video has become commonplace, both in professional and consumer use. The exponentially growing amounts of video content necessitates development of novel and multimodal technologies for analyzing, indexing, and retrieving relevant videos based on the audio and visual content of the video. In the visual content one can detect generic visual concepts, such as “vehicle” and “marching people”, recognize known persons, objects, buildings and locations, or perform optical character recognition. In the aural content, speech and speaker recognition can be done and music and environmental sounds can be classified. Combining all these parallel sources of information poses a challenging machine learning task.

COIN researchers have participated in NIST’s annual TREC Video Retrieval Evaluation TRECVID since 2005. In TRECVID 2013, we participated for the first time in the Multimedia Event Detection (MED) task. The research task is to build an automated system that can learn to determine whether an event is present in a video clip using the content of the video clip only. In TRECVID 2013, the considered multimedia events have included, e.g. “felling a tree”, “fixing musical instrument”, and “horse riding competition” and the total number of the video clips that were analyzed exceeded 90,000 and their duration exceeded 3,700 hours, The events are described additionally with a concise definition, a longer precise explication, and an evidential description, referring to some visual, acoustic, and temporal attributes that are often indicative of the event instance. Depending on the search task, either 0, 10 or 100 positive and negative example videos depicting the type of event are also provided and can be used for learning the properties of the event. In our experiments, we studied different combinations of visual and aural content indexing methods [9]. In particular, a novel unsupervised method for aural feature extraction was developed. The results show [1] that the proposed method is capable of extracting more applicable features for multimedia event detection than those commonly used in speech and audio recognition.

References

[1] Ehsan Amid, Annamaria Mesaros, Kalle J. Palomäki, Jorma Laaksonen, and Mikko Kurimo. Unsupervised feature extraction for multimedia event detection and ranking using audio content. In *Proceedings of EEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, Florence, Italy, May 2014.

[2] José Caldas, Nils Gehlenborg, Ali Faisal, Alvis Brazma, and Samuel Kaski. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics*, 25:i145–i153, 2009. (ISMB/ECCB 2009).

[3] José Caldas, Nils Gehlenborg, Eeva Kettunen, Ali Faisal, Mikko Rönty, Andrew G. Nicholson, Sakari Knuutila, Alvis Brazma, and Samuel Kaski. Data-driven information retrieval in heterogeneous collections of transcriptomics data links SIM2s to malignant pleural mesothelioma. *Bioinformatics*, 28(2):246–253, 2012. Supplementary data and source code are available from http://www.ebi.ac.uk/fg/research/rex.

[4] Ali Faisal, Jaakko Peltonen, Elisabeth Georgii, Johan Rung, and Samuel Kaski. Toward computational cumulative biology by combining models of biological datasets. *arxiv*, Submitted for publication.

[5] Elisabeth Georgii, Jarkko Salojärvi, Mikael Brosché, Jaakko Kangasjärvi, and Samuel Kaski. Targeted retrieval of gene expression measurements using regulatory models. *Bioinformatics*, 28(18):2349–2356, 2012.

[6] M. Gönen and E. Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

[7] Mehmet Gönen. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics*, 28(18):2304–2310, 2012.

[8] Mehmet Gönen and Samuel Kaski. Kernelized Bayesian matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear.

[9] Satoru Ishikawa, Markus Koskela, Mats Sjöberg, Jorma Laaksonen, Erkki Oja, Ehsan Amid, Kalle Palomäki, Annamaria Mesaros, and Mikko Kurimo. PicSOM experiments in TRECVID 2013. In *Proceedings of the TRECVID 2013 Workshop*, Gaithersburg, MD, USA, November 2013.

[10] Melih Kandemir, Arto Klami, Akos Vetek, and Samuel Kaski. Unsupervised inference of auditory attention from biosensors. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012)*, Lecture Notes in Computer Science, pages 403–418, Heidelberg, Germany, 2012. Springer.

[11] Melih Kandemir, Akos Vetek, Mehmet Gönen, Arto Klami, and Samuel Kaski. Multitask and multi-view learning of user state. *Neurocomputing*, to appear.

[12] S.A. Khan, S. Virtanen, O. Kallioniemi, K. Wennerberg, A. Poso, and S. Kaski. Identification of structural features in chemicals associated with cancer drug response: A systematic data-driven analysis. arXiv: 1312.7734v1, 2013.

[13] Suleiman A. Khan and Samuel Kaski. Bayesian multiview tensor factorization. submitted to a conference.

[14] A. Klami, G. Bouchard, and A. Tripathi. Group-sparse embeddings in collective matrix factorization. arXiv: 1312.5921, 2013.

[15] Arto Klami. Variational Bayesian matching. In Steven C.H. Hoi and Wray Buntine, editors, *Proceedings of Asian Conference on Machine Learning*, volume 25 of *JMLR C&WP*, pages 205–220. JMLR, 2012. Best paper award.

[16] Arto Klami. Bayesian object matching. *Machine Learning*, 92:225–250, 2013.

[17] Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14:965–1003, 2013. Implementation in R available at http://research.ics.aalto.fi/mi/software/CCAGFA/.

[18] Miika Koskinen, Jaakko Viinikanoja, Mikko Kurimo, Arto Klami, Samuel Kaski, and Riitta Hari. Identifying fragments of natural speech from the listener's MEG signals. *Human Brain Mapping*, 34(6):1477–1489, 2013.

[19] Gayle Leen, Jaakko Peltonen, and Samuel Kaski. Focused multi-task learning in a Gaussian process framework. *Machine Learning*, 89(1-2):157–182, 2012.

[20] K. Nybo, J. Shawe-Taylor, S. Kaski, and J. Mourao-Miranda. Characterizing unknown events in MEG data with group factor analysis. In *Proceedings of the 3rd Workshop on Machine Learning and Interpretation in Neuroimaging (MLINI)*, 2013.

[21] Joni Pajarinen, Ari Hottinen, and Jaakko Peltonen. Optimizing spatial and temporal reuse in wireless networks by decentralized partially observable Markov decision processes. *IEEE Transactions on Mobile Computing*, 13(4):866–879, 2014. Published online March 2013.

[22] Joni Pajarinen and Jaakko Peltonen. Expectation maximization for average reward decentralized POMDPs. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Zelezny, editors, *Proceedings of ECML PKDD 2013, The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 129–144, Berlin Heidelberg, 2013. Springer-Verlag.

[23] Sami Remes, Arto Klami, and Samuel Kaski. Characterizing unknown events in meg data with group factor analysis. In *Proceedings of the 3rd Workshop on Machine Learning and Interpretation in Neuroimaging (MLINI)*, 2013.

[24] Samuel Kaski Ritabrata Dutta, Sohan Seth. Retrieval of experiments with sequential dirichlet process mixtures in model space. *arXiv:1310.2125 [stat.ML]*, submitted to a journal.

[25] Sohan Seth, John Shawe-Taylor, and Samuel Kaski. Retrieval of experiments by efficient estimation of marginal likelihood. *arXiv:1402.4653 [stat.ML]*, submitted to a conference.

[26] Tommi Suvitaival, Juuso A. Parkkinen, Seppo Virtanen, and Samuel Kaski. Cross-organism prediction of drug hepatotoxicity by sparse group factor analysis. In *12th Annual International Conference on Critical Assessment of Massive Data Analysis (CAMDA)*, 2013. Extended abstract.

[27] Seppo Virtanen, Yangqing Jia, Arto Klami, and Trevor Darrell. Factorized multi-modal topic model. In Nando de Freitas and Kevin Murphy, editors, *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 843–851, Corvallis, Oregon, 2012. AUAI Press.

[28] Seppo Virtanen, Arto Klami, Suleiman A. Khan, and Samuel Kaski. Bayesian group factor analysis. In Neil Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *JMLR W&CP*, pages 1269–1277. JMLR, 2012. Implementation in R available at http://research.ics.aalto.fi/mi/software/CCAGFA/.

Chapter 3

C3: Statistical Inference in Structured Stochastic Models

Jukka Corander, Lu Cheng, Michael Gutmann, Väinö Jääskinen, Luca Martino, Pekka Marttinen, Elina Numminen, Ville Parkkinen, Jukka Sirén, Lu Wei, Jie Xiong

3.1 Introduction

The research activities in C3 are broadly two-fold; firstly we develop highly structured stochastic models for a wide variety of application areas ranging from statistical genetics to general multivariate system modeling, secondly we develop inference methods for the needs of such models. More detailed report of applications of these methods within the F2 flagship is given in its own section. It should be noted that most activities reported in C2 are also based on highly structured stochastic models and related inference tools, however, to maintain sufficient brevity we have here chosen not to present the overlap between C2 and C3 explicitly.

3.2 Probabilistic graphical models

Probabilistic graphical models (GMs) are ubiquitous in statistics and machine learning; one of major themes in C3 is to develop a large family of different generalizations of graphical models that conceptualize and enable capture of local, context-specific independencies (CSIs), in a more comprehensive way than the earlier proposals in the literature do allow. We have introduced a family of labeled directed acyclic graphs (LDAGs), which generalizes the Bayesian multinets and CPT-trees by allowing a compact and intuitive representation of CSIs such that exact Bayesian learning about model structure is possible. A related family of undirected stratified graphical models (SGMs), also termed as labeled Markov networks, has also been developed. These models allow for CSIs similar to LDAGs, and our theoretical results show that they are divided into locally decomposable and non-decomposable subclasses, the former of which allows for exact Bayesian inference. We have shown that LDAGs/SGMs in addition to being conceptually appealing, provide a powerful way to encode sparse dependencies in predictive classification that leads to higher classification accuracy compared to Bayesian networks which have partially irrelevant edges present in the DAGs. Furthermore, we have developed a class of sparse Markov chains (SMCs), which generalizes variable-length and variable-order Markov chains (VLMCs/VOMs) that are widely used in bioinformatics and natural language modeling applications. The SMC model class is a special case of LDAGs for time-series data in a finite state-space. To generalize the concept of CSIs to continuous variables, we have developed a novel class of stratified Gaussian graphical models (SGGMs), where an edge is allowed to be absent in a convex subset of values for its neighbors. It is proven that SGGMs represent a curved exponential family.

Inference and structural learning algorithms for GMs and their generalizations have been developed with three approaches in parallel to explore different possibilities. Firstly, we have generalized the family of non-reversible parallel (population) MCMC algorithms introduced earlier by Corander et al. by combining the non-reversible stochastic process with greedy hill-climbing, which seems to offer a very promising hybrid solution and balance of exploration-exploitation schemes. Secondly, we have created translations of the statistical learning problem to answer set programming and performed model optimization using existing solvers. This approach to model learning appears to be highly promising and is described more in detail in the section relating to C4. Thirdly, we have recently introduced a marginal pseudolikelihood (MPL) method, which appears to be among the first truly Bayesian versions of pseudolikelihood inference. We have proven its consistency for discrete graphical models that are not forced to be chordal. In computational ex-

periments MPL was both more accurate and faster than the state-of-the-art approaches based on regularized logistic regression and conditional mutual information. We are currently exploring generalization of MPL to several other model classes, including SGMs, and investigating its performance for identifying 3D contact patterns in proteins.

3.3 Adaptive Monte Carlo and adaptive MCMC

Monte Carlo methods, such as importance sampling, and MCMC have in general struggled considerably with the pace of increase in model complexity. One of the most popular solutions to the issue of slow convergence to the target distribution is to adapt the importance or proposal densities used in the sampling algorithm. Such an adaptation can be done by changing the locations and/or variance-covariance structure of the samplers. We have both developed adaptive importance samplers (e.g. APIS) and MCMC algorithms as well as demonstrated the usefulness of adaptive inference in challenging applications in computational biology. One of the key ideas in our algorithm development is to utilize a population of samplers which jointly attempt to adapt to better enable escape from local modes and to better represent difficult distribution shapes. The nonreversible population MCMC mentioned above is one example of such an approach where the proposal operators are allowed to be data-driven and have complicated, only algorithmically calculable expressions for proposal probabilities, since their values need not be explicitly known in the acceptance step of the Metropolis-Hastings transition kernel. The proposals can be adapted in several different manners, e.g. by driving a population of importance samplers with a MCMC kernel to allow a more thorough exploration of the parameter space to detect novel modes.

3.4 Inference for intractable models

Intractable models are in general understood as statistical models for which evaluation of likelihood terms is in practice a computationally intractable problem, when the model is of realistic size from the application perspective. Inference for such models has recently received considerable interest via the Approximate Bayesian Computation (ABC) framework and also via revival of the pseudolikelihood approach. As mentioned above, our Bayesian MPL method offers a very attractive opening of new research direction. In the application to non-chordal Markov networks we are considering a central class of intractable models and one of the particular strengths of the method is its parallelizability which paves way to very large-dimensional applications met e.g. in computational biology.

Many Bayesian models are intractable due to latent variables whose correlation structure or distribution assumptions in general make likelihood expression underivable in closed form (or numerically). ABC inference for such models replaces the likelihood by filtering results of forward simulation with given parameter values and it is one of the most intensive research areas in Bayesian statistics at the moment. The challenge related to use of ABC has three major components: 1) choice of the summary statistics to mimic the likelihood, 2) choice of metric to represent closeness of forward simulation output with that in the observed data, 3) filtering algorithm of forward simulation output to yield a reliable approximation of the posterior. In the context of computational biology, we have introduced a novel model for estimating transmission dynamics of bacteria from

cross-sectional incidence data. This model has a very high-dimensional latent variable component and an intractable likelihood. To enable learning of such models from data, we developed a sequential adaptive Monte Carlo algorithm which approaches the problem of comparing simulation output with the data summaries in a novel way, based on adapting the sampler separately with respect to the distributions of each summary statistic rather than using a single Euclidean distances as in most ABC methods. This adaptation scheme makes the sequential sampler much more efficient. Our second highlight work on ABC and indirect inference introduces a Bayesian optimization scheme for choosing optimally points in parameter space where forward simulation is going to be maximally informative about the likelihood approximation. Our experiments suggest that this approach can be several orders of magnitude faster than the standard sampling methods widely used in ABC. The third highlight solves the problem of choosing summary statistics and distance measures for comparing forward simulation output with data summaries, by combining these two problems translating the result into a classification problem. To the best of our knowledge, such an approach is entirely novel and offers several advantages. We have proved the consistency of the classifier-ABC estimates and shown that it can automatically yield a more infromative data representation compared with expert curated choice of summary statistics.

References

[1] Corander, J., Xiong, J., Cui, Y. and Koski, T. Optimal Viterbi Bayesian Predictive Classification for Data from Finite Alphabets. *Journal of Statistical Planning and Inference*, doi:10.1016/j.jspi.2012.07.013 (2012)

[2] Corander, J., Janhunen, T., Rintanen, J., Nyman, H. and Pensar, J. Learning chordal Markov networks by constraint satisfaction. In: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (eds.) *Advances in Neural Information Processing Systems* 26, pp. 1349-1357 (2013)

[3] Jääskinen, J., Xiong, J., Koski, T. and Corander, J. Sparse Markov Chains for Sequence Data. *Scandinavian Journal of Statistics*, doi: 10.1111/sjos.12053. (2013)

[4] Martino, L. Elvira, V. Luengo, D. and Corander, J. An Adaptive Population Importance Sampler. In *Proceedings of the 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, in press. (2014)

[5] Numminen, E., Cheng, L., Gyllenberg, M. and Corander, J. Estimating the transmission dynamics of Streptococcus pneumoniae from strain prevalence data. *Biometrics*, doi: 10.1111/biom.12040. (2013)

[6] Nyman, H., Pensar, J., Koski, T. and Corander, J. Stratified graphical models - context-specific independence in graphical models. arXiv:1309.6415v1 [stat.ML] (2013)

[7] Nyman, H., Xiong, J., Pensar, J. and Corander, J. Marginal and simultaneous predictive classification using stratified graphical models. arXiv:1401.8078v1 [stat.ML] (2014)

[8] Nyman, H., Pensar, J. and Corander, J. Stratified Gaussian graphical models. Submitted to *Biometrika*. (2013)

[9] Pensar, J., Nyman, H., Koski, T., Corander, J. Labeled directed acyclic graphs: a generalization of context-specic independence in directed graphical models. arXiv:1310.1187v1 [stat.ML] (2013)

[10] Pensar, J. Nyman, H. and Corander, J. Marginal Pseudo-Likelihood Inference for Markov Networks. arXiv:1401.4988v1 [stat.ML] (2014)

[11] Sirén, J., Hanage, W.P. and Corander, J. Inference on Population Histories by Approximating Infinite Alleles Diffusion. *Molecular Biology and Evolution*, doi: 10.1093/molbev/mss227. (2012)

[12] Sunnåker, M., Busetto, A.G., Numminen, E., Corander, J., Foll, M. and Dessimoz, C. Approximate Bayesian Computation. *PLoS Comput Biol* 9(1): e1002803. doi:10.1371/journal.pcbi.1002803 (2013)

Chapter 4

C4: Extreme Inference

Ilkka Niemelä, Tomi Janhunen, Tommi Junttila, Jussi Rintanen, Jeremias Berg, Jori Bomanson, Jukka Corander, Martin Gebser, Matti Järvisalo, Roland Kindermann, Tero Laitinen, Guohua Liu, Brandon Malone

4.1 Introduction

The goal of challenge area C4 is to develop efficient and scalable learning and inference techniques, in particular, to tackle problems arising in C1, C3, and F1. This is to be achieved by combining existing expertise on machine learning and constraint-based search and optimization, which is presently available in COIN. The problems of interest include, e.g., learning probabilistic models from data, clustering based on different optimization criteria, and fast inference techniques demanded by applications. The idea is to exploit massively distributed computing environments, such as clusters and computational grids, when solving the most demanding problem instances and to develop new algorithmic techniques required. To this end, there is substantial hardware[1] at our disposal.

During the first two years of COIN, the contributions in the area of C4 are two-fold. First, there is substantial progress made in the core reasoning techniques and, in particular, the development of new methodology for *answer set programming* (ASP) as well as *Boolean satisfiability checking* (SAT) and its extensions such as maximum satisfiability (MaxSAT). Achievements in these paradigms are reported in Sections 4.2 and 4.3, respectively. As regards the main applications in the COIN agenda, the problem of learning graphical models from data has been tackled using constraint-based techniques both in the case of Markov networks and Bayesian networks as reported in Section 4.4. Further applications are addressed in Section 4.5. In addition to COIN-specific application domains, also others are emerging due to general applicability of logic-based methods.

4.2 Contributions to ASP Methodology

Answer set programming is a declarative programming paradigm where problems are first formalized as logic programs (sets of rules) and then solved by computing answer sets for programs. In addition to developing native ASP solvers, there is interest towards translations that enable the implementation of ASP using other back-end solvers.

In the translation-based approach, one of the main concerns is the treatment of extended rule types such as *choice*, *cardinality*, and *weight* rules. If such constructs are not supported by the target formalism, they have to be translated away. In [4], we develop new schemes for the *normalization* of cardinality rules. The designs are based on merging and sorting circuits and lead to $n \times (\log_2 k)^2$ blow-up for n-literal cardinality rules having k as the bound. There is ongoing work that extends these designs for weight rules as well as objective functions used in optimizing variants of ASP. Since auxiliary variables have to be introduced, the notion of *visible strong equivalence* developed in [14] is central when addressing the correctness of normalization. For reasonably small rules, the respective verification steps can be fully automated using appropriate translations and ASP solvers. Normalization is also exploited by native ASP solvers and from time to time it may boost the search for answer sets as auxiliary atoms contribute to propagation and branching.

We have previously developed a translation from ASP into the satisfiability modulo theories (SMT) framework and, in particular, to the fragment corresponding to *difference logic*. In [28], an analogous translation into *fixed-width bit-vector* theories is presented, thus enabling the use of further SMT solvers for the computation of answer sets. The case of *mixed integer programming* (MIP) is covered in [26]: the presented translation shows

[1] For instance, the Triton cluster of Aalto University has 6900 cores available for the moment.

how different rule types used in ASP can be transformed into linear inequalities of form $a_1x_1 + \ldots + a_nx_n \geq k$. The performance of MIP solvers is promising on ASP instances involving optimization. Moreover, we suggest to enrich ASP languages to support linear constraints as native primitives in rules. This enables the treatment of potentially infinite domains such as integers within answer set programs. The challenges arising from the incorporation of real numbers are addressed in a follow-up paper [27].

Many knowledge representation tasks involve trees or similar structures as abstract datatypes. In this respect, graphical models addressed in Section 4.4 form an excellent example. Compact and efficient ASP encodings of acyclicity properties are devised and experimentally evaluated in [8]. There is ongoing work on extending these results for other tree-like properties and other target formalisms than ASP.

4.3 Contributions to SAT Methodology

The goal of this research is to develop fundamental techniques for SAT solving and to integrate them in state-of-the-art SAT solver technology.

A class of constraints for a number of core applications of SAT solving, including cryptanalysis, circuit verification and model-checking, that is not always handled effectively by the resolution rule implicit in general-purpose SAT solvers, is *parity constraints*. For example, logical cryptanalysis instances usually have a non-linear part that can be presented in clausal form, and a linear part more conveniently presented with parity constraints, i.e., linear equations in modulo 2 arithmetic. For instance, the constraints for the shift register bit 1 in the Trivium cipher may look like this:

$$(\neg t_i \vee s_{286,i}) \wedge (\neg t_i \vee s_{287,i}) \wedge (t_i \vee \neg s_{286,i} \vee \neg s_{287,i}) \wedge (s_{1,i+1} \oplus s_{243,i} \oplus s_{288,i} \oplus t_i \oplus s_{69,i} \equiv \text{false})$$

An incremental Gauss-Jordan-elimination based deduction method for parity constraints that can capture all implied literals is given in [23]. In [22] we present a new conflict-analysis technique that can (i) easily solve some instances that are hard for resolution and (ii) produce new parity constraints that can be learned. We have also studied structural properties of parity constraint sets and defined easily detectable classes in which weaker, more efficiently implementable reasoning techniques already give full propagation power [21]. In addition, it is shown how to use structural properties to decompose parity constraint sets into components that communicate through interface variables [23, 25].

Furthermore, we provide translation techniques that aim at improving the propagation power of unit propagation, the most elementary (and efficiently implementable) deduction technique, by adding new parity constraints into the instance. One of the translation technique is tailored for instances on which equivalence reasoning gives full propagation power [21] and the another one is for the unrestricted case [24]. We also show that a polynomial amount of additional constraints is enough to make unit propagation a complete propagation engine for parity constraint sets with bounded tree-width [25].

A central line of research in core SAT solvers has dealt with the *inprocessing* paradigm, in which the core CDCL SAT solver routine is interleaved with complex combinations of relatively inexpensive inference techniques, thus re-introducing more reasoning to the core SAT solver search. The research has established the formal underpinnings of inprocessing SAT solving via an abstract inprocessing framework [16], which generally covers all inprocessing techniques implemented in current inprocessing SAT solvers, captures sound

solution reconstruction in a unified way, and allows for checking new inprocessing techniques for correctness. There has also been progress in the analysis and development of further inprocessing techniques. In [15], it is shown that CNF-level reasoning can be surprisingly effective both in theory and in practice: without explicit knowledge of the underlying problem structure, specific inexpensive CNF-level inference techniques achieve the same level of simplification as a combination of structure-based techniques and previously suggested polarity-based CNF encodings. Further work on inprocessing has included development and analysis of more powerful clause elimination techniques [11] as well as CNF-level techniques for equivalence reasoning [12]. Extending the work on core SAT techniques, a study of the applicability and the effectiveness of SAT preprocessing within SAT-based algorithms for the extracting minimally unsatisfiable subformulas (MUSes) of Boolean formulas was presented [1]. Most recently, there is on-going work on extensions of inprocessing techniques to the more general context of quantified Boolean formulas (QBFs) and MaxSAT, as well as formal lower-bound analysis of known CNF-level reasoning techniques such as failed literal elimination [19].

Further SAT-related work has included foundational studies on the interplay between proof complexity measures and practical hardness of SAT [18], providing new theoretical and empirical insights into the possibility of resolution space complexity as a fine-grained indicator of practical hardness of SAT instances, as well as work relating Boolean function circuit complexity under restricted circuit classes with the strong exponential-time hypothesis, i.e., the existence of non-trivial exponential-time algorithms for CNF-SAT [17].

4.4 Learning Graphical Models by Constraint Satisfaction

Integration of constraint programming with data mining and machine learning has recently been identified as an important research direction with high potential. The research in COIN contributes in this direction by developing new methodology for solving optimally different structure learning problems by employing state-of-the-art solver technology. For instance, combinatorial search methods have been developed in order to learn *probabilistic graphical models*, such as Bayesian networks, from data. The learning problem can be solved by formalizing the structure requirements in terms of constraints and using a constraint solver to find an optimal solution. This idea has been successfully applied in the case of Bayesian networks. In [5], the substantially harder problem of learning undirected graphical models, known as *Markov networks* or *Markov random fields*, is tackled. Figure 4.1 illustrates an optimal network for the *Econ* dataset with 8 variables. Starting from initial log likelihood scores computed from the data, an encoding of Markov network structure is presented in terms of constraints expressible in MaxSAT or ASP frameworks. The encoding relies on a novel characterization of *junction trees* used to define the *separators* between cliques (marked red in Figure 4.1) and to score the entire

network. Ensuring the *chordality* of the underlying network is one of the central constraints. Using existing ASP and MaxSAT solver technology, it is then possible to prove optimal certain network structures previously discovered by stochastic search methods such as Markov chain Monte Carlo. Interestingly, the optimality proof can be carried out in certain cases much faster than what is takes by a stochastic algorithm to converge. In ongoing work, we have improved the encoding for Markov networks using *perfect elimination orderings* as basis. Together with with various strategies for solution pruning, they offer a dramatic improvement both in terms of time and memory complexity. The method is also extended for a more general class of models, called *labeled* Markov networks, which have an astronomically larger model space.

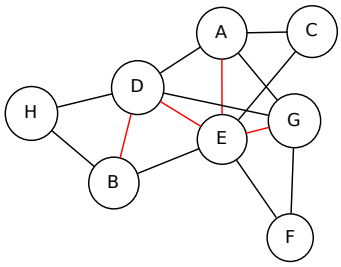


Figure 4.1: Optimal Markov network amongst 30.888.596 candidates

In [3], a novel score-based approach to the NP-hard problem of learning optimal bounded tree-width Bayesian networks, was presented, based on casting the problem as weighted partial Maximum satisfiability problem. Bayesian network structure learning is the well-known problem of finding a directed acyclic graph structure that optimally describes given data. An underlying motivation for this work is that, while exact inference in Bayesian networks is in general NP-hard, it is tractable in networks with low tree-width. Empirically, the approach scales notably better than the current state-of-the-art exact dynamic programming algorithm for the problem. Moreover, in [13], a very general approach to learning the structure of causal models is presented. The approach is based on d-separation constraints, obtained from any given set of overlapping passive observational or experimental data sets. The procedure allows for both directed cycles (feedback loops) and the presence of latent variables. Our approach is based on a logical representation of causal pathways, which permits the integration of quite general background knowledge, and inference is performed using a SAT solver iteratively, associating the causal structure learning task with the backbone of Boolean formulas. Many existing constraint-based causal discovery algorithms can be seen as special cases of the procedure, tailored to circumstances in which one or more restricting assumptions apply.

4.5 Further Applications

We have also continued devising methods for the model checking analysis of safety critical timed systems employing real-valued clocks. Such systems are usually modeled with *timed automata*. Techniques for model checking timed automata have been studied for two decades. However, most of the techniques studied do not support quantitative specifications on the timing of events. In [20], we consider the specification language $\mathrm{MITL}_{0,\infty}$ allowing one to write specifications like

$$(\mathbf{G}^{\mathrm{s}}_{<5}\neg \mathit{ack}) \Rightarrow \mathbf{F}^{\mathrm{s}}_{\leq 6}(\mathit{reset}\ \mathbf{R}^{\mathrm{s}}_{\leq 10}\ \mathit{alarm})$$

meaning that either an acknowledgment must be received in less than five seconds or otherwise the alarm is activated within six seconds and played for the next ten seconds or until a reset button is pushed. The previously presented semantics of $\mathrm{MITL}_{0,\infty}$ are extended to support super-dense time (allowing a finite number of instantaneous actions to take place between time-consuming actions) that is applied in the timed automata class used in several standard tools such as Uppaal. As a consequence of this, the dualities

that have been used in the previous works on model checking $\mathrm{MITL}_{0,\infty}$ do not hold anymore. In order to obtain symbolic model checking techniques for $\mathrm{MITL}_{0,\infty}$, a symbolic encoding that evaluates the truth values of $\mathrm{MITL}_{0,\infty}$ formulas on the symbolically presented executions of the system under verification is developed. The encoding is proven sound and complete, and it is described how it can be instantiated in the symbolic model checking technique called SMT-based bounded model checking. To our best knowledge, we also provide the first implementation of a tool for model checking $\mathrm{MITL}_{0,\infty}$ specifications and provide preliminary experimental evidence supporting that model checking such quantitative specifications can be feasible in practice, too.

Finding execution paths in transition systems is a core problem in control and management of discrete systems. In this research, we have developed first practical methods for deriving approximate upper bounds for transition sequence lengths based on decomposition of state-variable dependency graphs [30]. These bounding methods strengthen the power of highly parallel use of SAT solvers in finding transition sequences in control applications (with the possibility of carrying over the ideas to applications in verification) by bounding the lengths of transition sequences. No practical general methods have been earlier available for this purpose, which has led either to the use of unnecessarily loose upper bounds and substantially increased search effort, or increased risk of incompleteness due to uninformed and too tightly chosen bounds.

The development of effective methods for decision-making under imperfect information has been a long standing problem. The success of combinatorial search methods such as SAT in sequential decision-making with perfect information is the driving motivation for the use of generalizations and extensions of these methods also for the imperfect information case. The research has identified new classes of contingent scheduling problems amenable to solution with quantified extensions of standard satisfiability and constraint-satisfaction problems such as QBF and QCSP [29]. The results place a number of imperfect-information scheduling problems in the complexity classes Σ_2^p, Π_2^p, and PSPACE, showing the suitability of $\exists\forall$- and $\forall\exists$-quantified, and unlimited QBF and QCSP.

There has also been work on novel SAT and ASP approaches to arising topics in AI, especially, computational argumentation [6, 7] as highlighted next and computational creativity [31], extending further the application domains covered by COIN. Abstract *argumentation frameworks* (AFs) provide the basis for various reasoning problems in knowledge representation and artificial intelligence. Efficient evaluation of AFs has thus been identified as an important research challenge. In [6, 7], we presented a generic approach for reasoning over AFs, based on the novel concept of complexity-sensitivity. We established theoretical foundations of this approach, providing further understanding on the sources of intractability of AF reasoning problems via complexity-theoretical analysis, and presented instantiations of the generic complexity-sensitivity framework via harnessing SAT solver technology for solving argumentation problems in an iterative way.

Further logic-based interdisciplinary work has considered the relation of various restricted models of distributed computing and modal logic [10, 9]. In another line of work [2], we have developed an extensible framework for correlation clustering by harnessing the MaxSAT paradigm. The approach allows for finding cost-optimal clusterings, and extends to constrained correlation clustering, by allowing for easy integration of user-defined domain knowledge in terms of hard constraints over the clusterings of interest, as well as overlapping correlation clustering.

References

[1] Anton Belov, Matti Järvisalo, and Joao Marques-Silva. Formula preprocessing in MUS extraction. In Nir Piterman and Scott Smolka, editors, *Proceedings of the 19th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS 2013)*, volume 7795 of *Lecture Notes in Computer Science*, pages 110–125. Springer, 2013.

[2] Jeremias Berg and Matti Järvisalo. Optimal correlation clustering via MaxSAT. In Wei Ding, Takashi Washio, Hui Xiong, George Karypis, Bhavani M. Thuraisingham, Diane J. Cook, and Xindong Wu, editors, *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW 2013)*, pages 750–757. IEEE Press, 2013.

[3] Jeremias Berg, Matti Järvisalo, and Brandon Malone. Learning optimal bounded treewidth Bayesian networks via maximum satisfiability. In Jukka Corander and Samuel Kaski, editors, *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS 2014)*, 2014 (to appear).

[4] Jori Bomanson and Tomi Janhunen. Normalizing cardinality rules using merging and sorting constructions. In *Logic Programming and Nonmonotonic Reasoning*, volume 8148 of *LNCS*, pages 187–199. Springer, 2013.

[5] Jukka Corander, Tomi Janhunen, Jussi Rintanen, Henrik Nyman, and Johan Pensar. Learning chordal Markov networks by constraint satisfaction. In *Advances in Neural Information Processing Systems*, volume 26 of *Advances in Neural Information Processing Systems*, pages 1349–1357, 2013.

[6] Wolfgang Dvořák, Matti Järvisalo, Johannes Peter Wallner, and Stefan Woltran. Complexity-sensitive decision procedures for abstract argumentation. In Thomas Eiter and Sheila McIlraith, editors, *Proceedings of the 13th International Conference on Principles of Knowledge Representation and Reasoning (KR 2012)*, pages 54–64. AAAI Press, 2012.

[7] Wolfgang Dvořák, Matti Järvisalo, Johannes Peter Wallner, and Stefan Woltran. Complexity-sensitive decision procedures for abstract argumentation. *Artificial Intelligence*, 206:53–78, 2014.

[8] Martin Gebser, Tomi Janhunen, and Jussi Rintanen. ASP encodings of acyclicity properties. In *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning, KR 2014*, In Press.

[9] Lauri Hella, Matti Järvisalo, Antti Kuusisto, Juhana Laurinharju, Tuomo Lempiäinen, Kerkko Luosto, Jukka Suomela, and Jonni Virtema. Weak models of distributed computing, with connections to modal logic. *Distributed Computing*.

[10] Lauri Hella, Matti Järvisalo, Antti Kuusisto, Juhana Laurinharju, Tuomo Lempiäinen, Kerkko Luosto, Jukka Suomela, and Jonni Virtema. Weak models of distributed computing, with connections to modal logic. In Darek Kowalski and Alessandro Panconesi, editors, *Proceedings of the 31st Annual ACM Symposium on Principles of Distributed Computing (PODC 2012)*, pages 185–194. ACM, 2012.

[11] Marijn Heule, Matti Järvisalo, and Armin Biere. Covered clause elimination. In Andrei Voronkov, Geoff Sutcliffe, Matthias Baaz, and Christian Fermüller, editors, *Short*

Paper Proceedings of the 17th International Conference on Logic for Programming, Artificial Intelligence and Reasoning (LPAR-17 / 2010), volume 13 of *EasyChair Proceedings in Computing*, pages 41–46, 2013.

[12] Marijn Heule, Matti Järvisalo, and Armin Biere. Revisiting hyper binary resolution. In Carla Gomes and Meinolf Sellmann, editors, *Proceedings of the 10th International Conference on Integration of Artificial Intelligence and Operations Research Techniques in Constraint Programming (CPAIOR 2013)*, volume 7874 of *Lecture Notes in Computer Science*, pages 77–93, 2013.

[13] Antti Hyttinen, Patrik Hoyer, Frederick Eberhardt, and Matti Järvisalo. Discovering cyclic causal models with latent variables: A general SAT-based procedure. In Ann Nicholson and Padhraic Smyth, editors, *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pages 301–310. AUAI Press, 2013.

[14] Tomi Janhunen and Ilkka Niemelä. Applying visible strong equivalence in answer-set program transformations. In *Correct Reasoning – Essays on Logic-Based AI in Honour of Vladimir Lifschitz*, volume 7265 of *LNCS*, pages 363–379. Springer, 2012.

[15] Matti Järvisalo, Armin Biere, and Marijn Heule. Simulating circuit-level simplifications on CNF. *Journal of Automated Reasoning*, 49(4):583–619, 2012.

[16] Matti Järvisalo, Marijn Heule, and Armin Biere. Inprocessing rules. In Bernhard Gramlich, Dale Miller, and Uli Sattler, editors, *Proceedings of the 6th International Joint Conference on Automated Reasoning (IJCAR 2012)*, volume 7364 of *Lecture Notes in Computer Science*, pages 355–370. Springer, 2012.

[17] Matti Järvisalo, Petteri Kaski, Mikko Koivisto, and Janne H. Korhonen. Finding efficient circuits for ensemble computation. In Alessandro Cimatti and Roberto Sebastiani, editors, *Proceedings of the 15th International Conference on Theory and Applications of Satisfiability Testing (SAT 2012)*, volume 7317 of *Lecture Notes in Computer Science*, pages 369–382. Springer, 2012.

[18] Matti Järvisalo, Arie Matliah, Jakob Nordström, and Stanislav Živný. Relating proof complexity measures and practical hardness of SAT. In Michela Milano, editor, *Proceedings of the 18th International Conference on Principles and Practice of Constraint Programming (CP 2012)*, volume 7514 of *Lecture Notes in Computer Science*, pages 316–331, 2012.

[19] Matti Järvisalo and Janne. H. Korhonen. Conditional Lower Bounds for Failed Literals and Related Techniques. In *Proceedings of the 17th International Conference on Theory and Applications of Satisfiability Testing (SAT 2014), Lecture Notes in Computer Science*, Springer, 2014.

[20] Roland Kindermann, Tommi Junttila, and Ilkka Niemelä. Bounded model checking of an MITL fragment for timed automata. In *Proceedings of the 13th International Conference on Application of Concurrency to System Design, ACSD 2013*, pages 216–225. IEEE, 2013.

[21] Tero Laitinen, Tommi Junttila, and Ilkka Niemelä. Classifying and propagating parity constraints. In *Principles and Practice of Constraint Programming, CP 2012*, volume 7514 of *LNCS*, pages 357–372. Springer, 2012.

[22] Tero Laitinen, Tommi Junttila, and Ilkka Niemelä. Conflict-driven XOR-clause learning. In *Theory and Applications of Satisfiability Testing, SAT 2012*, volume 7317 of *LNCS*, pages 383–396. Springer, 2012.

[23] Tero Laitinen, Tommi Junttila, and Ilkka Niemelä. Extending clause learning SAT solvers with complete parity reasoning. In *Proceedings of the 24th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2012*. IEEE Computer Society Press, 2012.

[24] Tero Laitinen, Tommi Junttila, and Ilkka Niemelä. Simulating parity reasoning. In *Proceedings of the 19th International Conference on Logic Programming and Automated Reasoning, LPAR 2013*, volume 8312 of *LNCS*, pages 568–583. Springer, 2013.

[25] Tero Laitinen, Tommi Junttila, and Ilkka Niemelä. Simulating parity reasoning (extended version). *CoRR*, abs/1311.4289, 2013.

[26] Guohua Liu, Tomi Janhunen, and Ilkka Niemelä. Answer set programming via mixed integer programming. In *Proceedings of the 13th International Conference on Principles of Knowledge Representation and Reasoning, KR 2012*, pages 32–42. AAAI Press, 2012.

[27] Guohua Liu, Tomi Janhunen, and Ilkka Niemelä. Introducing real variables and integer objective functions to answer set programming. In *Declarative Programming and Knowledge Management*, LNCS, 2014. Revised Selected Papers of INAP'13.

[28] Mai Nguyen, Tomi Janhunen, and Ilkka Niemelä. Translating answer-set programs into bit-vector logic. In *Applications of Declarative Programming and Knowledge Management*, volume 7773 of *LNCS*, pages 95–113. Springer, 2013. Revised Selected Papers of INAP'11.

[29] Jussi Rintanen. Scheduling with contingent resources and tasks. In *Proceedings of the International Conference on Automated Planning and Scheduling, ICAPS 2013*, pages 189–196. AAAI Press, 2013.

[30] Jussi Rintanen and Charles Orgill Gretton. Computing upper bounds on lengths of transition sequences. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI 2013*, pages 2365–2372. AAAI Press, 2013.

[31] Jukka M. Toivanen, Matti Järvisalo, and Hannu Toivonen. Harnessing constraint programming for poetry composition. In Mary Lou Maher, Tony Veale, Rob Saunders, and Oliver Bown, editors, *Proceedings of the 4th International Conference on Computational Creativity (ICCC 2013)*, pages 160–167. The University of Sydney, 2013.

Chapter 5

F1: Intelligent Information Access

Jorma Laaksonen, Erkki Oja, Samuel Kaski, Petri Myllymäki, Markus Koskela, Ville Viitaniemi, Mats Sjöberg, Antti Ajanki, Jaakko Peltonen, Cristina Gonzalez-Caro, Zhirong Yang, Teemu Roos, Dorota Głowacka, Matti Karppa, Marcos Luzardo, He Zhang, Mikko Kurimo, Kalle Palomäki, Andre Mansikkaniemi, Annamaria Mesaros, Janne Pylkkönen, Reima Karhila, Sami Keronen, Heikki Kallasjoki, Ulpu Remes, Seppo Enarvi, Dhananjaya Gowda, Ville Turunen, Peter Smit, Matti Järvisalo, Ziyuan Lin, Antti Oulasvirta, Manuel Eugster, Arto Klami

5.1 Introduction

The goal of COIN flagship application F1 Intelligent Information Access is to break the conventional keyboard-mouse-display based human–computer interaction scheme and allow the user to access contextual information in the real world. This goal can be reached by applying solid computational inference methods that can make use of the massive interrelated information sources when selecting what information to present and how to visualize it to the user. The inference methods need to be fast and do the inference online, learning relevance from the user's responses, both explicit and implicit. For the user input we develop techniques for analyzing diverse search cues and semantic indications, such as visual gestures, gaze patterns, audible background, recognized speech, physiological measurements, and sensory data, which together can reveal the target of the user's current information need.

We report our work on contextual information interfaces, including of a prototype of a mobile personal history browser, in Section 5.2. Basic research on interactive visualization methods is discussed in Section 5.3. SciNet, an information access system for scientific literature involving interactive visualization and interactive user intent modeling is described in Section 5.4. We have studied a multitude of input modalities for facilitating intelligent information access. In Section 5.5 we report our work on biosignals, such as EEG, galvanic skin response and eye tracking for implicit feedback. Section 5.6 describes studies on computer-vision-based human activity recognition, usable for contextual information access as both explicit and implicit search cues. Similarly, Section 5.7 addresses research on automatic speech and aural environment recognition, the methods of which can be used for providing user input for intelligent information interfaces as well as for indexing the multimodal data repositories. In Section 5.8, we detail our research on visual analysis techniques for efficient content recognition and indexing of multimodal information, necessary for facilitating intelligent, proactive and fast retrieval and presentation.

5.2 Contextual information interfaces

The increasing pervasiveness of mobile computers and digital recording devices in daily life has made it easy to capture and store a lot of data in digital form. Consequently, the sizes of digital photograph collections, textual document archives, lifelog recordings and other forms of personal digital histories are growing rapidly, and new kinds of retrieval tools are required to effectively find data items from such collections. We have introduced a personal history browser (PHB) mobile application for recording and browsing events, images, notes and other personal data [1]. The interface intelligently emphasizes potentially relevant items on a timeline, so that an item can easily be recognized and selected for further inspection. If items are shown in a meaningful order, as is the case with personal history items along a timeline, the emphasized items have another important role as search cues. The purpose is to support fast retrieval of heterogeneous personal data. For example, PHB can help retrieve notes the user wrote in a particular spatial location, or find an article she remembers reading after a discussion with a particular person. We have previously explored a similar approach in a desktop setting [2].

In our PHB interface (Figure 5.1), implemented on a smartphone, data items are primarily represented as thumbnail images. The purpose is to allow quick spotting of familiar images. We base the relevance predictions on a multi-domain representation of the history items,

Figure 5.1: Prototype dynamic timeline interface in the PHB system. The textual search box is at the top, the list of the most relevant items with relevance feedback buttons in the middle, and the scrollable view of time-ordered images scaled by their relevance at the bottom. Relevance feedback for the items can be given with the plus and minus controls.

such as associated images, videos, textual notes, or other descriptions recorded during the same calendar event or otherwise close by in time. Our system learns a separate ranking function on each domain and combines them to produce the final prediction. Because a PHB is inherently a mobile application, it can be used to record events during daily life. Mobility is also important for searching, because one is likely to find information when needed, if the browser is handily available in a mobile form. We conducted a user study to compare the dynamic timeline of our PHB prototype to a more conventional textual search in a news story retrieval task [2]. The results of the study showed that the dynamic timeline interface was significantly more effective than and preferred over the reference methods.

In recent work, we have studied the integration of PHB with wearable devices, such as wearable cameras and mobile eye tracking glasses. The implicit or automatic collection of the personal history can be implemented with an application that is autonomously working as a background process. Wearable cameras provide a natural way to gather a photographic history for PHB-type applications. Furthermore, mobile eye tracking glasses enable us to infer the user's interests with respect to the environment from the eye movement patterns and use this data for information filtering.

References

[1] Antti Ajanki, Markus Koskela, Jorma Laaksonen, and Samuel Kaski. Adaptive timeline interface to personal history data. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 229–236, New York, NY, USA, 2013. ACM.

[2] Antti Ajanki and Samuel Kaski. Probabilistic proactive timeline browser. In *Proc. 21st International Conference on Artificial Neural Networks (ICANN), Part II*, pages 357–364, 2011.

5.3 Interactive visualization

The new visualization approaches discussed in Section 1.4 consider static visualization situations, where the interest of the analyst is sufficiently well known and encoded as a metric of the original data or annotation available for the data, and the remaining task is to show the interesting information in the data on the display as well as possible. The approaches for scaling up the visualizations to larger data sets [3, 4, 9], to more accurate visualizations [1], to principled incorporation of supervision [2], and to meta-visualization [5], can potentially be applied in interactive situations as well, by combining them with the approaches discussed below.

In [6], *interactive visualization* is formalized as a task of information retrieval under uncertainty about the user's tacit knowledge and interests in the data. An interactive visualization method is created which iteratively gathers user feedback, as pairs of points the user wishes to be close-by or far apart, then infers the user's interest from the feedback as a metric of relevant differences between data features, and visualizes the data using the relevant metric, by optimizing information retrieval of the relevant neighbors from the display based on our recent formalization [8]. The interaction goes on as more user feedback is received and the metric and the resulting visualization improve to better match the user's interests.

Moreover, the accurate visualization approach [1] can easily incorporate user feedback as user-assigned constraints to yield optimal visualizations satisfying such constraints. Lastly, the interactive visual search approach [7] discussed in Section 5.4 involves an advanced visual interface optimized for the estimated user's intent and for alternative future intents; thus the method can be seen as interactive visualization coupled to an advanced user modeling approach and information retrieval approach.

References

[1] Kerstin Bunte, Matti Järvisalo, Jeremias Berg, Petri Myllymäki, Jaakko Peltonen, and Samuel Kaski. Optimal neighborhood preserving visualization by maximum satisfiability. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*, AAAI Press, 2014.

[2] Mehmet Gönen. Bayesian supervised dimensionality reduction. *IEEE Transactions on Cybernetics*, 43(6):2179–2189, 2013.

[3] Jaakko Peltonen and Konstantinos Georgatzis. Efficient optimization for data visualization as an information retrieval task. In Ignacio Santamaría, Jerónimo Arenas-García, Gustavo Camps-Valls, Deniz Erdogmus, Fernando Pérez-Cruz, and Jan Larsen, editors, *Proceedings of MLSP 2012, the 2012 IEEE International Workshop on Machine Learning for Signal Processing*, Piscataway, NJ, 2012. IEEE.

[4] Jaakko Peltonen and Ziyuan Lin. Multiplicative update for fast optimization of information retrieval based neighbor embedding. In Saeid Sanei, Paris Smaragdis, Asoke Nandi, Anthony TS Ho, and Jan Larsen, editors, *Proceedings of MLSP 2013, the 2013 IEEE International Workshop on Machine Learning for Signal Processing*, Piscataway, NJ, 2012. IEEE.

[5] Jaakko Peltonen and Ziyuan Lin. Information retrieval perspective to meta-visualization. In *Proceedings of ACML 2013, Fifth Asian Conference on Machine Learning*, JMLR W&CP, volume 29, pages 165–180, 2013. JMLR.

[6] Jaakko Peltonen, Max Sandholm, and Samuel Kaski. Information retrieval perspective to interactive data visualization. In M. Hlawitschka and T. Weinkauf, editors, *Proceedings of Eurovis 2013, The Eurographics Conference on Visualization*. The Eurographics Association, 2013.

[7] Tuukka Ruotsalo, Jaakko Peltonen, Manuel J. A. Eugster, Dorota Głowacka, Ksenia Konyushkova, Kumaripaba Athukorala, Ilkka Kosunen, Aki Reijonen, Petri Myllymäki, Giulio Jacucci, and Samuel Kaski. Directing exploratory search with interactive intent modeling. In *Proceedings of CIKM 2013, the ACM International Conference of Information and Knowledge Management*, pages 1759-1764, New York, NY, 2013. ACM.

[8] 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.

[9] Zhirong Yang, Jaakko Peltonen, and Samuel Kaski. Scalable optimization of neighbor embedding for visualization. In *Proceedings of ICML 2013, the 30th International Conference on Machine Learning*, volume 28 of *JMLR W&CP*, pages 127–135. JMLR, 2013.

5.4 Interactive intent modelling and SciNet

Inferring a user's intention in human-computer interaction is a key research issue for developing personalized systems. In this line of research, we focused on the information retrieval setup. Traditional search engines support user needs in scenarios where the user is aware of what they are looking for. However, systems that would support *exploratory search activities*, requiring learning and investigating the information space, have turned out to be more difficult to design. One reason is that in an exploratory search setting the searcher is not a priori familiar with the information and hence requires iteration of interpretation, synthesis, and evaluation of the found information to accomplish their task. We propose that better support for exploration can be provided through learning from feedback on higher level representations of the data sets, such as topics or keywords, that are extracted from document features.

In [3, 1] we proposed new methodology for interactive information search, namely *Interactive Intent Modeling*, where the user's search intent and its alternatives are modeled and displayed for feedback on an interactive display. This feedback enables applying machine learning techniques such as reinforcement learning to improve relevance, novelty and diversity of results.[1]

Based on the interactive intent modeling approach, we built SciNet [1, 2, 3] – an information access system that couples advanced machine learning techniques for interactive intent modeling with advanced information visualization and interaction to boost exploratory

[1] In detail, the searchers intents are estimated by a reinforcement-learning based intent model by simultaneously maximizing the relevance of estimated search intents for the searcher and minimizing the uncertainty of the intent estimates of the system.

search. The primary goal of the system is to assist scientists in finding and exploring the relevant literature on a given research topic quickly and effectively, although the approach can additionally be easily adapted to other domains.

The interactive intent modeling approach yields greatly improved information seeking task performance in user studies. In detail, we could show that users using our approach achieved significantly better task performances, retrieved relevant information items more effectively, interacted more without a decrease of the quality of information, and were also more pleased with their search experience.

The interactive interface In the interactive interface, instead of only typing queries at each iteration, the user can navigate by manipulating keywords on a visual display shown in Figure 5.2. The manipulations of keywords are used as feedback which the system uses to improve its estimation of the user's search intent. This results in new keywords appearing on the screen as well as a new set of documents being presented to the user. The search starts with the user typing in a query, which results in a set of keywords being displayed in the exploratory view on the left hand-side of the screen and a set of articles being displayed on the right hand-side of the screen. The user can manipulate the keywords in the exploratory view to indicate their relevance: the closer to the center a given keyword is, the more relevant it is. The user can manipulate as many keywords as she likes. After each iteration, new keywords and new articles are displayed. The search continues until the user is satisfied with the results.

Information seeking with interactive intent modeling The interactive intent modeling approach brings many benefits to the way users perform their information seeking tasks:

- It allows users to direct their search using the offered keyword cues at any point of time without getting trapped in a context, or having to provide tedious document–level relevance feedback, or relying on implicit feedback mechanisms that may take long to converge.
- The users can actively engage in an exploratory search loop where they manipulate article features such as keywords, and the underlying machine learning system offers them navigation options (keywords, articles) using an exploration–exploitation paradigm. The search becomes significantly faster by allowing exploration and easier query manipulation.
- We have found a suitable abstract level on which it is convenient for the users to direct their search (in our case, the document keywords are the navigation options users can use to direct their search), and use observed interaction together with feedback to feed reinforcement learning–based optimization of further navigation options. This can support users in better directing the exploratory search nearer or further from the current context and following a direction.

Main components of the intent modeling process The visualization allow the user to give feedback (assign relevance scores) to the displayed keywords by moving them within the exploratory view provided by the system (keyword manipulation), which is

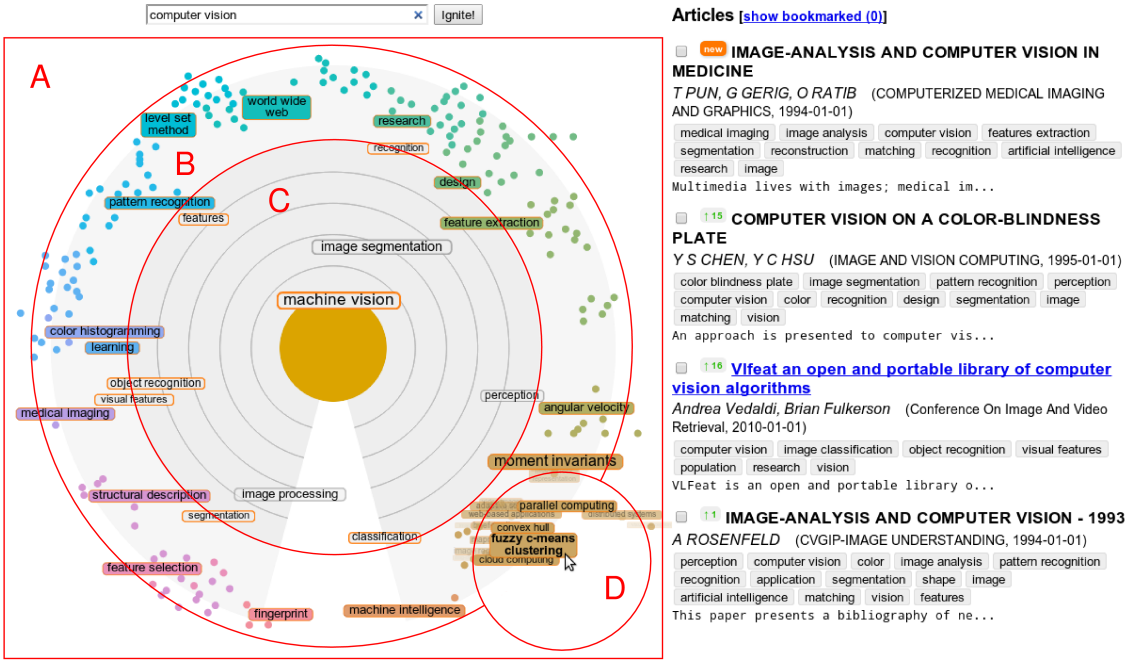


Figure 5.2: A novel information access system based on interactive intent modeling, where users are shown a representation of their estimated present and future intents on an interactive visual interface, and they can give natural feedback on the interface to tune the search towards relevant information. Search intents are visualized through keywords on a radial layout (A). The orange center area represents the user: the closer a keyword is to the center the more relevant it is to the estimated intent. The intent model used for retrieval is visualized as keywords in the inner circle (C); projected future intents are visualized as keywords in the outer circle (B). Keywords can be inspected with a fisheye lens (D).

then provided as feedback to reinforcement learning (RL) methods. Through keyword manipulation, the user can direct the search according to her interest, while the inbuilt RL mechanism helps the system to form a model of user's interests and suggest appropriate keywords in the next search iteration. The learning of the user's intent and the corresponding retrieval of new relevant documents and keywords is composed of three main modules: (1) Information Retrieval and Ranking, (2) Keywords Exploration, and (3) Document Diversification. The process of modeling the user's intent is restarted once the user types in a new query and we build a new user model for each session to avoid the issue of "over-personalization".

Three main blocks of the system are responsible for the initial retrieval and ranking of documents, and exploration in the keyword and the document spaces using RL. The initial set of documents and their rankings are obtained through the Information Retrieval and Ranking module. Having received feedback on keywords, the system enters the exploratory loop. The explicit user feedback is sent to the Keywords Exploration and the Document Diversification modules. The Keywords Exploration module implements user model estimation using RL techniques. The user model is a representation of the system's belief about the user's informational need at the current iteration of retrieval. The component receives feedback from the user and produces a list of keywords with weights which are passed on to the Information Retrieval and Ranking module, which predicts a new set

of documents for the new search iteration based on the predicted user model. Thus, the dataset in the system is not static and it changes at every iteration based on the present, best estimation of the user model.

The Document Diversification module is responsible for determining the set and order of documents that are passed on to the Interface. The module uses exploration–exploitation techniques to sample a set of documents to display to the user, while keeping the ranking obtained from the Information Retrieval and Ranking module. The new set of documents is used in Keywords Exploration module to capture dependencies between keywords. The user model is visualized in the exploratory view, which allows the user to give feedback to the system through keyword manipulation. A list of articles is also presented to the user. The system gets new feedback from the user and continues in the iterative feedback loop.

References

[1] D. Głowacka, T. Ruotsalo, K. Konyushkova, K. Athukorala, S. Kaski, and G. Jacucci. Directing exploratory search: Reinforcement learning from user interactions with keywords. In *18th International Conference on Intelligent User Interfaces (IUI)*, pages 117–128, 2013.

[2] T. Ruotsalo, K. Athukorala, D. Głowacka, K. Konyushkova, A. Oulasvirta, S. Kaipiainen, S Kaski, and G. Jacucci. Supporting exploratory search tasks with interactive user modeling. In *76th Annual Meeting of the Association for Information Science and Technology (ASIST)*, 2013.

[3] T. Ruotsalo, J. Peltonen, M. Eugster, D. Głowacka, K. Konyushkova, K. Athukorala, I. Kosunen, A. Reijonen, P. Myllymäki, G. Jacucci, and S. Kaski. Directing exploratory search with interactive intent modeling. In *22nd ACM International Conference on Information & Knowledge Management (CIKM)*, 2013.

5.5 Feedback from biosignals

In interactions between humans a lot of information is exchanged implicitly—this motivates human-computer interaction related research to go beyond explicit user feedback and exploit, for example, biosignals like electroencephalogram (EEG), galvanic skin response (GSR) and eye tracking, for implicit feedback. In this line of research we specifically focused on proposing methodologies for learning the user's attention and relevance of information items to the user from biosignals. These methodologies are mainly based on (probabilistic) machine learning methods for supervised and unsupervised multi-view and multi-task learning developed in C2.

We [2] inferred information about relevance of the objects the user inspected, based on gaze tracking information, in mobile settings. In [3], we studied ways of automatically inferring the level of attention a user is paying to auditory content, with applications for example in automatic podcast highlighting and auto-pause. In this work, we introduced a novel time-dependent Bayesian CCA model by encoding time-dependent interactions in the generative description. We learned the model from the coupled physiological signals (EEG, body movement, and pupil dilation) and features computed for the audio content, and then measured the amount of correlation to represent the level of attention. We

showed that the correlation reveals the level of attention with accuracy comparable to a user-independent supervised models. This means, that we were able to directly detect auditory attention as correlation between the two signals, without needing any training data.

Currently, we are working on methodologies for learning relevance of information items from brain signals; and on methodologies for improved information search by leveraging search history and biosignals. In [1] we use EEG data separated into different views (e.g., frequency bands and ERPs) to automatically detect relevance of text information directly in an information retrieval setup using Bayesian Multi Kernel Learning.

References

[1] Manuel J. A. Eugster, Tuukka Ruotsalo, Michiel Sovijärvi-Spapé, Ilkka Kosunen, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. Predicting term-relevance from brain signals. *Submitted to a conference.*

[2] Melih Kandemir and Samuel Kaski. Learning relevance from natural eye movements in pervasive interfaces. In Louis-Philippe Morency and Dan Bohus, editors, *Proceedings of the International Conference on Multimodal Interaction, ICMI '12*, pages 85-82, New York, NY, 2012. ACM.

[3] Melih Kandemir, Arto Klami, Akos Vetek, and Samuel Kaski. Unsupervised inference of auditory attention from biosensors. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012), Lecture Notes in Computer Science*, pages 403-418, Heidelberg, Germany, 2012. Springer.

5.6 Visual recognition of human actions

Analysing human actions in videos has long been an important area of computer vision, constantly receiving the attention of researchers. Humans and their actions are often central in deciding the meaning and interpretation of the contents of a given piece of video material. The human action analysis is used, e.g., in surveillance and patient monitoring systems, and in various kinds of human-computer interfaces. Applications in information retrieval are also becoming more common.

Action and gesture recognition from motion capture and RGB-D (RGB and depth) camera sequences has recently raised considerable research interest. Starting from either video, motion capture, depth data, or some combination of these modalities, many action and gesture recognition methods have been applied in various fields. Motion capture (mocap) systems capture human motion with high frequency and accuracy by calculating the joints' coordinates and the angular information of the human skeleton using a system setup consisting of multiple calibrated cameras in a dedicated space. On the other hand, the recently introduced commodity RGB-D sensors, such as the Microsoft Kinect, provide depth information along with the RGB video with portability and low cost. Algorithms have been developed to extract the skeleton model from the depth frames in real-time [1].

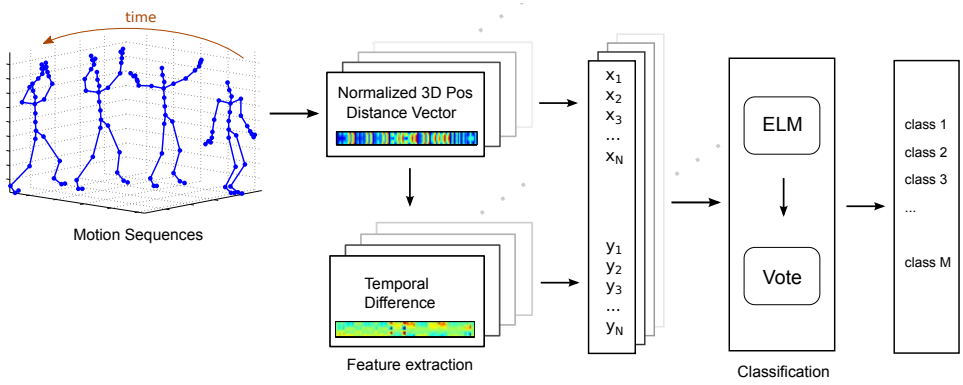


Figure 5.3: An overview of the skeleton-based activity recognition framework.

They provide analogous, albeit noisier, data to mocap, and this enables the action classification methodologies developed for mocap skeletons to be applied for RGB-D data as well.

Our approach to skeleton-based activity recognition [2], is to first extract features from the raw skeletal data of each frame, classify the actions on the frame level, and then build a model of the whole action sequence by aggregating the frame-wise results to get the final classification result. Frame-level features are classified with Extreme Learning Machines, which can provide high accuracy and, at the same time, both classification and the training of the models are very fast compared to many other non-linear classification methods. A graphical overview of the classification framework can be seen in Fig. 5.3.

The skeletal features capture the movement of the whole body or a body region, but are not able to capture hand gestures, which often present meaningful linguistic symbols. For this reason, we have extended the skeleton-based recognition framework by locating the hand regions from the RGB-D frames, and extracting histogram of oriented gradients (HOG) features from these regions [3, 4]. Experiments with the ChaLearn 2013 dataset[2] show that the HOG-based hand features provide a valuable addition that can reduce the overall error rate even with low spatial resolution and the presence of strong motion blur.

Analysis of sign language videos is a very special case of human action analysis as in sign language the movements and postures carry the very information the signers want to communicate. This directly defines one type of ground truth against which the results of video analysis can be evaluated. From the point of computer vision research, sign language analysis is scientifically interesting as it entails challenging problems involving complex body movements and skin-coloured articulators that occlude each other. Our research is targeted at developing automatic video analysis tools that would help the linguistic research of sign languages. Current sign language research often utilises corpus-based approaches where large collections of videos would need to be annotated at least for signs on the basis of information concerning, for example, the locations, shapes, and movements of the hands producing them [5]. Also non-manual aspects of the videos would often be of importance. We have made the algorithms and methods developed publicly availabe in our SLMotion video analysis toolkit [6].

In linguistic analysis the stream of signing is routinely segmented into signs and inter-

[2]http://gesture.chalearn.org/2013-multi-modal-challenge

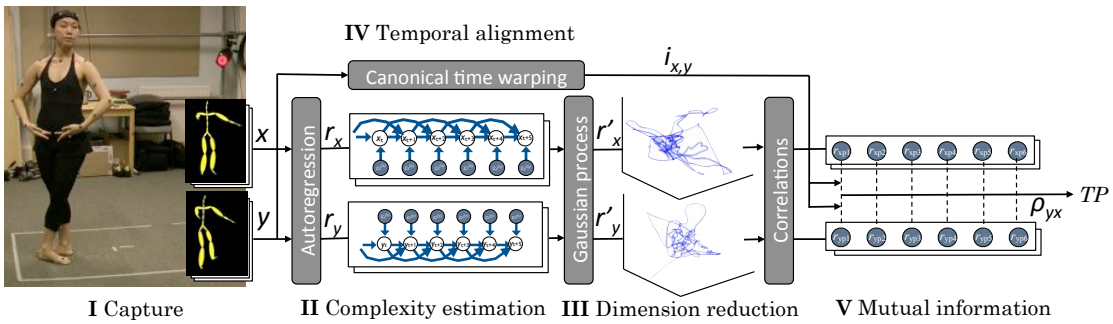


Figure 5.4: Overview of computation steps in calculating information capacity (TP) in full-body movement. Image source: [14].

sign transitions. One of the themes in our work has been developing an automatic system capable of this. In order to make this goal more concrete and measurable, we have compiled and published an annotated benchmark data collection for spotting specific signs [7], based on the video material of the Suvi online dictionary of Finnish sign language[3]. The manual segmentation process has often been enhanced with quantitative measurements concerning the hand movement (e.g. [8]). For this task, the most accurate method has always been motion capturing with specialised equipment, but it cannot be used for pre-recorded material and it is always tied to laboratory settings. In [9] we have compared the motion capture measurements with a computer vision based method that enables tracking and measuring the motion of the hand and other articulators. Our study showed that the movement and position information obtained by video analysis is often very accurate.

Typical approaches for hand tracking, based on skin colour segmentation, have difficulties when the skin blobs are merged because of the hands touching or occluding each other or the head region. In [10] we proposed and studied a method for detecting hand-head occlusions that is based on local tracking of skin-coloured points in the neighbourhoods of the head. The local approach is combined with global tracking of the hand movements to reduce the number of false positive detections. It is inconceivable to try to understand sign language without recognising also the handshapes. In our recent study [11], we studied which visual feature extraction methods would be the most useful for handshape recognition.

In sign languages the movements and poses of the head in whole as well as those of the individual facial elements express important communicative, grammatical, prosodic and emotional information. In [12] we propose a method for automatic detection of the three head pose angles—yaw, pitch, and roll—from images. The method is based on two kinds of visual features: tonal segmentation masks of skin-like colours within the face bounding box. In addition to head pose estimation, in [13] we propose and evaluate methods for estimating the state of facial elements—eyes, eyebrows and mouth —in the context of sign language. The applicability of methods is naturally not limited to sign language analysis even though they have been devised specifically for this application. For example, we have demonstrated their use for speaker identification in news broadcast videos.

An important part of human-computer interaction research addresses the design of practical user interface technologies and user studies that evaluate their performance. In

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

addition, there is also need for theoretical models characterizing the fundamental limits of different interaction scenarios. In [14], we propose a measure that can be used to study the capacity of human movement from an information-theoretic point of view. The paper presents case studies ranging from classical ballet and in-air gesturing to mouse pointing. A follow-up paper [15] presents an application measuring the security of gesture-based authentication systems in comparison to conventional symbolic passwords. Further applications of the metric can be envisioned in, e.g., medical diagnostics, rehabilitation, and sports science.

References

[1] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *Proc. Computer Vision and Pattern Recognition*, June 2011.

[2] Xi Chen and Markus Koskela. Classification of RGB-D and motion capture sequences using extreme learning machine. In *Proceedings of the 18th Scandinavian Conference on Image Analysis*, volume 7944 of *LNCS*, Espoo, Finland, June 2013. Springer Verlag.

[3] Xi Chen and Markus Koskela. Online RGB-D gesture recognition with extreme learning machines. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction (ICMI 2013)*, pages 467–474, Sydney, Australia, December 2013. ACM.

[4] Xi Chen, Markus Koskela, and Jorma Laaksonen. Using appearance-based hand features for dynamic rgb-d gesture recognition. In *Proceedings of 22nd International Conference on Pattern Recognition (ICPR 2014)*, Stockholm, Sweden, August 2014.

[5] Trevor Johnston. Guidelines for annotation of the video data in the Auslan corpus. Online publication http://media.auslan.org.au/media/upload/attachments/ Annotation_Guidelines_Auslan_CorpusT5.pdf, 2009. Dept. of Linguistics, Macquarie University, Sydney, Australia.

[6] Matti Karppa, Ville Viitaniemi, Marcos Luzardo, Jorma Laaksonen, and Tommi Jantunen. SLMotion – an extensible sign language oriented video analysis tool. In *Proceedings of 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavík, Iceland, May 2014. European Language Resources Association.

[7] Ville Viitaniemi, Tommi Jantunen, Leena Savolainen, Matti Karppa, and Jorma Laaksonen. S-pot – a benchmark in spotting signs within continuous signing. In *Proceedings of 9th Language Resources and Evaluation Conference (LREC 2014)*, 2014.

[8] Tommi Jantunen. Signs and transitions: Do they differ phonetically and does it matter? *Sign Language Studies*, 13(2):211–237, 2013.

[9] Matti Karppa, Tommi Jantunen, Ville Viitaniemi, Jorma Laaksonen, Birgitta Burger, and Danny De Weerdt. Comparing computer vision analysis of signed language video with motion capture recordings. In *Proceedings of 8th Language Resources and Evaluation Conference (LREC 2012)*, pages 2421–2425, Istanbul, Turkey, May 2012.

Available online at `http://www.lrec-conf.org/proceedings/lrec2012/pdf/321\_Paper.pdf`.

[10] Ville Viitaniemi, Matti Karppa, Jorma Laaksonen, and Tommi Jantunen. Detecting hand-head occlusions in sign language video. In *Proceedings of the 18th Scandinavian Conference on Image Analysis*, volume 7944 of *LNCS*, Espoo, Finland, June 2013. Springer Verlag.

[11] Ville Viitaniemi, Matti Karppa, and Jorma Laaksonen. Experiments on recognising the handshape in blobs extracted from sign language videos. In *Proceedings of 22nd International Conference on Pattern Recognition (ICPR 2014)*, Stockholm, Sweden, August 2014.

[12] Marcos Luzardo, Matti Karppa, Jorma Laaksonen, and Tommi Jantunen. Head pose estimation for sign language video. In *Proceedings of the 18th Scandinavian Conference on Image Analysis*, volume 7944 of *LNCS*, Espoo, Finland, June 2013. Springer Verlag.

[13] Marcos Luzardo, Ville Viitaniemi, Matti Karppa, Jorma Laaksonen, and Tommi Jantunen. Estimating head pose and state of facial elements for sign language video. In *Proceedings of 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavík, Iceland, May 2014. European Language Resources Association.

[14] A. Oulasvirta, T. Roos, A. Modig, and L. Leppanen. nformation capacity of full-body movements. In *Proc. of CHI'13*, 2013.

[15] M. Sherman, G. Clark, Y. Yang, S. Sugrim, A. Modig, J. Lindqvist, A. Oulasvirta, and T. Roos. User-generated free-form gestures for authentication: security and memorability. In *Proc. 12th International Conference on Mobile Systems, Applications, and Services (Mobisys-2014)*, ACM, 2014. arXiv:1401.0561.

5.7 Speech recognition

Training and adaptation of acoustic models Acoustic modeling in speech technology 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. Commonly used feature vector consist of mel-frequency cepstral coefficients (MFCC) or linear predictor (LP) based features. MFCCs are the result of the discrete cosine transform (DCT) applied to the logarithmic mel-scaled filter bank energies. LP-based analysis methods model the vocal tract formants more directly. 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 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.

Note that similar models are used for speech recognition as for speech synthesis. Different training techniques can be used for adapting the model to the task at hand.

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, 4].

As accoustic 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 [5] to transform a generic model in to a specific model using a limited amount of data.

The HMM-based acoustic modeling framework of an ASR system can be inverted and used to generate speech with some modifications. Text-to-speech (TTS) systems take text prompts as input, predict prosodic elements related to duration and stress, and use the acoustic models to generate vocoder parameters for synthetic speech. The acoustic models for TTS are often trained separately for each speaker, and try to capture the expressiveness of the speech and the personal characteristics of the speaker. Model clustering is used to get more robust approximations for similar phones, as well as to allow synthesis of previously unseen phone sequences. In a similar fashion to ASR, an average acoustic model can be adapted to a new speaker with a small amount of speech data [6]. The speaker-adaptive system is also very robust against noise in the adaptation data [7]. With high-quality average voice models, it is possible to create high-quality adapted voices even when there is a presence of noise in the adaptation data. Beside speaker adaptation, it is possible to adapt a TTS voice to a different speaking style.

Noise robust speech recognition Reasonably accurate speech recognition has been possible for years in controlled conditions where the noise levels are low and words are clearly articulated. The continuously increasing computational power has enabled the study of complex speech recognition systems trained on thousands of hours of speech data. The recent advances in neural networks have set the current research trend towards hybrid multilayer-perceptron and HMM structures that are displacing the traditional HMM-GMM structures as the basis of modern ASR systems. Despite the progress, the performances of the most complex systems still degrade in the presence noise. The work presented in this section is focused on methods that model the uncertainty in the observed or reconstructed (or cleaned) speech features when the clean speech signal is corrupted with noise from an unknown source. In addition to the uncertainty-based methods,

we have studied in bandwidth extension that aimed to improve the quality of perceived telephone speech [8].

The missing data methods, which draw inspiration from the human auditory system, are based on the assumption that the noise corrupted speech signal can be divided into reliable speech-dominant and unreliable noise-dominant time-frequency regions as illustrated in Figure 5.5.

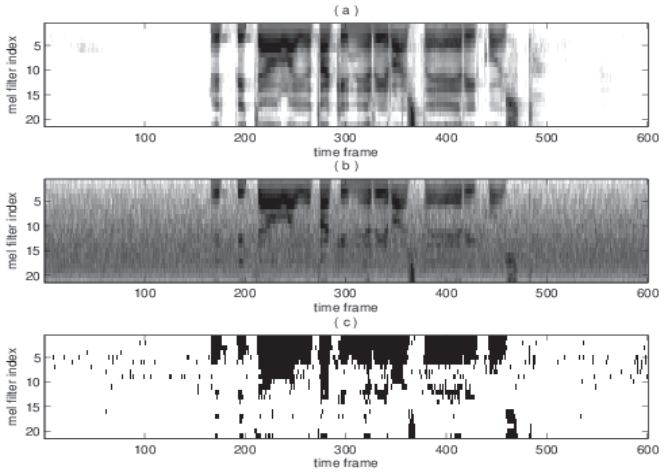


Figure 5.5: Logarithmic Mel spectrogram of (a) an utterance recorded in a quiet environment and (b) the same utterance corrupted with additive noise. The noise mask (c) constructed for the noisy speech signal indicates the speech dominated regions in black and the noise dominated regions in white.

The speech and noise segregation can be simplified to a binary classification problem e.g. by extracting acoustic features that are important for the auditory organization of speech. Such features are, for example, interaural time difference and interaural level difference which measure the differences in arrival time and intensity of a sound signal between two ears.

A noise robust missing data method based on multifeature mask estimation was introduced in [9] to counter the detrimental effects of low SNR environments. The article also proposed retraining acoustic models on imputed data and extensively evaluated several acoustic features for their ability to discriminate reliable and unreliable mask information. In addition, two feature reconstruction methods were compared.

Instead of founding the mask estimation on a pre-designed set of features, the set of features can be automatically learned via unsupervised training of acoustical patterns with neural networks. The automatic features may enhance the mask estimation accuracy by capturing information that the pre-designed features can not capture. A method to automatically learn the set of features for missing data mask estimation was proposed in [10] by implementing the Gaussian-Bernoulli restricted Boltzmann machine. The automatic features were derived from cross-correlation vectors computed from bandpass filtered stereophonic speech signals.

In our missing data approaches, the missing clean speech information is reconstructed either by cluster-based imputation or sparse imputation in windows that span several time frames. The cluster-based imputation is 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. The imputed missing clean speech features can also be associated with an approximate posterior distribution to model uncertainty in the reconstruction. Noise-robust speech recognition based on the approximate posterior proposed in [11] improved speech recognition performance compared to baseline cluster-based imputation.

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 a statistical language model (LM).

Statistical language models are usually trained on large quantities of newspaper texts. 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. In our efforts to improve the recognition of conversational Finnish speech, we have developed methods to retrieve texts from the Web and select texts which are more likely to be of a conversational and informal nature [12]. An LM trained on filtered Web data significantly improves the recognition of conversational speech.

We often want to adapt the LM to a certain topic. Usually we can't find enough data to train a reliable standalone topic-specific LM. The standard setting for language model adaptation is to combine a background model trained on newspaper texts (large text set) with an adapted model trained on topic-specific texts (small text set).

One focus of our research has been to develop unsupervised language model adaptation for Finnish speech recognition [13]. We have developed an adaptation framework where the topic is estimated from first-pass ASR output. Topic-related texts are retrieved from an online source and used to train a small topic LM which is combined with the background LM through linear interpolation. The adapted LM is then used in second-pass recognition.

Another focus point has been to adapt the vocabulary to improve the recognition of foreign words in Finnish speech recognition. This involves detecting foreign word candidates in topic-specific texts and generating pronunciation variants for them. Adding several new pronunciation variants to the vocabulary can increase acoustic confusability between words. A challenge has been to improve recognition of foreign words while maintaining recognition accuracy of native words intact. We have developed a method to remove harmful pronunciation variants based on lattice rescoring [14]. Pronunciation variants which give a net increase of word error rate are removed.

Content based audio retrieval While multimedia content available online in the internet grows exponentially every day, searches are still often based on textual labels. Considering the user contributed content in services like YouTube, for instance, searches can be conducted on clip titles or other key words, but there is yet no possibility to search in within these clips. More intelligent access would allow a direct and precise access to the multimedia content. Movies could be searched based on what the actors said, what are the images in the video, or what are the sounds in the audio track. Imagine a personal media

clip from wedding parties. With an intelligent search the user could find the moment when the wedding cake was cut, when the band in the weddings started playing wedding waltz or the moment of the honeymoon when a lion roared loudly in a safari. In this section, we focus on the development of new approaches to access information in generic audio namely content based audio retrieval. With this approach the beginning of the wedding waltz or roaring of lion could be spotted.

Our research in audio based retrieval addressed first challenges in diversely labeled audio data, in which users could name sounds freely as they have heard them. Sounds that are acoustically similar can be labeled with different semantic descriptions depending on the user and his or her life experiences and preferences. To investigate the issue, we conducted an analysis that combined the semantic and acoustic aspects. Our results showed that for a given audio event the acoustically closest neighbor is on the average either one level higher or lower in the semantic hierarchy [15]. In addition, less than half of the test data samples had a synonym among the ten samples that were acoustically closest. We have two main application targets for the joint acoustic semantic analysis of audio. First, the analysis could be used to develop automatic methods to refine the labeling of diverse real-world audio. For example acoustically and semantically near neighbors could be given the same labels. Alternatively, a semantic parent label could be given to sounds that are acoustically similar, but have unnecessarily detailed labels. The second application scenario we have already addressed is retrieval based on queries that have both textual description and an audio example [16]. Figure 5.6 illustrates a query for crickets. In this case, the text description is simply "crickets" and then an audio example is used to refine the retrieval with desired sound qualities. In a practical scenario, after an initial textual search user could pick some examples of hits with desired audio content to retrieve more audio samples that share some desired qualities with those preferred hits.

Figure 5.6: Illustration of the content based audio retrieval system. Results are shown for a query of crickets. The query is conducted using both semantic description and an audio example. The left pane illustrates the case when semantics is weighted more than the audio example in the query. The left pane illustrates the case when their weight is equal. Using The equal weight results in more examples with labels other than crickets.

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] Janne Pylkkönen and Mikko Kurimo. Improving discriminative training for robust acoustic models in large vocabulary continuous speech recognition. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*, pages 1–4, 2012.

[4] Janne Pylkkönen and Mikko Kurimo. Analysis of extended baum-welch and constrained optimization for discriminative training of hmms. *IEEE Transcations on Audio, Speech, and Language Processing*, 20(9):2409–2419, November 2012.

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

[6] J. Yamagishi, Thousands of Voices for HMM-Based Speech Synthesis-Analysis and Application of TTS Systems Built on Various ASR Corpora. In *IEEE Transactions on Audio, Speech and Language Processing* , vol. 18, pp. 984-1004, 2010.

[7] R. Karhila, HMM-based speech synthesis adaptation using noisy data: analysis and evaluation methods. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013.

[8] H. Pulakka, U. Remes, S. Yrttiaho, K. J. Palomäki, M. Kurimo, and P. Alku. Bandwidth Extension of Telephone Speech to Low Frequencies Using Sinusoidal Synthesis and a Gaussian Mixture Model, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2219–2231, 2012.

[9] S. Keronen, H. Kallasjoki, U. Remes, G. J. Brown, J. F. Gemmeke, and K. J. Palomäki. Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment, *Computer Speech and Language*, vol. 27, no. 3, pp. 798–819, 2013.

[10] S. Keronen, K. H. Cho, T. Raiko, A. Ilin, and K. J. Palomäki. Gaussian-Bernoulli Restricted Boltzmann Machines and Automatic Feature Extraction for Noise Robust Missing Data Mask Estimation, *Proc. ICASSP 2013*, Vancouver, Canada, pp. 6729–6733, May 2013.

[11] U. Remes. Bounded conditional mean imputation with an approximate posterior. *Proc. INTERSPEECH*, pp. 3007–3011, August 2013.

[12] S. Enarvi and M. Kurimo, Studies on Training Text Selection for Conversational Finnish Language Modeling. In *Proceedings of the 10th International Workshop on Spoken Language Translation*, 2013.

[13] A. Mansikkaniemi and M. Kurimo, Unsupervised Topic Adaptation for Morph-based Speech Recognition, *Proceedings of Interspeech*, 2013.

[14] S. Enarvi and M. Kurimo, A Novel Discriminative Method for Pruning Pronunciation Dictionary Entries. In *Proceedings of the 7th International Conference on Speech Technology and Human-Computer Dialogue*, 2013.

[15] A. Mesaros, T. Heittola and K. Palomäki . Query-by-example retrieval of sound events using an integrated similarity measure of content and label, *14th International Workshop on Image and Audio Analysis for Multimedia Interactive services (WIA2MIS 2013)*, Paris, France, 2013.

[16] A. Mesaros, T. Heittola and K. Palomäki . Analysis of acoustic-semantic relationship for diversely annotated real-world audio data, *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013.

5.8 Video content analysis for intelligent access

The inference problem in multimedia event detection is to build an automated system that can learn to determine, using the content of the video clip only, whether a prespecified event occurs in a video clip. We have studied multimedia event detection and related problems while participating in the TRECVID evaluations. In the TRECVID 2013 evaluation [1], the considered multimedia events have included, e.g. "felling a tree", "fixing musical instrument", and "horse riding competition". The media event detection system we have implemented in our PicSOM media analysis framework not only detects events in video clips but also recounts the evidences used to identify the event. These recountings help the user quickly and accurately locate their event of interest within the clips detected by the system. The system task is to provide a recounting of the important evidence that a video clip contains an instance of an event of interest.

In the analysis of TV broadcast material, a particularly important topic is to identify the persons appearing in the TV programmes to be analyzed. For this purpose, we have developed a multimodal approach consisting of *speaker segmentation*, *speaker clustering*, and *facial clustering*.

After the segments of speech in an audio stream have been detected and labelled, it is important to know if they are composed of only one or several speakers. The aim is to label each segment with an unique speaker id, so we have to break any multi speaker segments into unique speaker segments before clustering. The approach we are using is based on a growing window and the use of the Bayesian Information Criterion (BIC) as a distance measure. In speaker clustering, we label each speaker turn of the same speaker with the same label to identify who spoke when. Since the problem of measuring the similarity between segments is the same as in segmentation, BIC can also be used as a metric here. If we assume that the previous segmentation has good quality and segments contain only one speaker, we can perform a standard hierarchical clustering on the segments, by merging those that are similar as being produced by the same speaker.

When creating a multimodal summary of video content, the identities of the visible persons play a key role. Even if the names or other exact indicators of the persons cannot be assigned to the seen faces, it is still useful to identify the persons with some consistent tags [2]. When creating a summary, these tags can then be used to ensure that each person will appear in the summary exactly once. The result of the person identification process can be used as a temporal video segmentation parallel to the speaker segmentation result.

These segmentations can then be fused to generate a visual–aural summary of the video and to produce a multimodal person database comprising of facial image and speech voice sample pairs of the persons included.

The term *affect* denotes a broad category encompassing feelings, emotions and moods of humans. There are many application areas for which computational models of affect would have great value, for example movie indexing and recommendation systems, as well as image content classification [3].

In [4] we performed a set of experiments to predict affective content for 14 movie clips, taken from popular mainstream movies made between 1955 and 2009 encompassing several genres [5]. Ground truth data was collected in a user experiment in which 72 participants were shown a series of movie clips and asked to assess their stylistic, aesthetic and affective attributes. The human-provided ratings were then used to train the algorithms used in the computational prediction. We have made the collected data and low-level features extracted from the clips publicly available[4].

Two prediction methods are compared: multiple linear regression and the recent neural-network-based Extreme Learning Machine (ELM) [6] algorithm. Our study found that felt affect was the easiest to predict, while style was the second easiest category to predict, followed by perceived affect, and lastly, aesthetics. The finding is interesting in the sense that though both affect and aesthetics are abstract concepts, the former appears to be more closely linked to low-level features than the latter. Our feature-specific results corroborate earlier findings that aural features are suited for arousal modelling, and that temporal features generally perform well in affect modeling.

Another concrete application of affective content prediction is detecting violent scenes in movies. In 2013 we participated in an international team taking part in the MediaEval 2013 Affect Task [7] which challenged participants to develop algorithms for finding violent scenes in popular Hollywood movies. The violence detection system combines standard visual and auditory features together with hierarchical detection system that first detects a set violence-related concepts such as “blood”, “firearms”, “screams” and “explosions”, and then uses their outputs as features for the final violence predictor. Both the violence and mid-level concept classifiers are multi-layer perceptrons utilizing a random dropout scheme to improve generalization. The system had the best performance in 2012, and among the best also in 2013.

In the Helsinki Privacy Experiment project we conducted a large-scale longitudinal study on the effects of ubiquitous surveillance at home. The goal of the project was to understand the effects of continuous computerized surveillance on individuals, and we instrumented ten (voluntary) Finnish households with video cameras, microphones, and logging software for personal computers, wireless networks, smartphones, TVs, and DVDs for a period of 12 months [9]. Our preliminary results [8] expose a range of negative changes in the experience and behavior of the volunteers. All in all, the project produced over 50 terabytes of rich, multi-modal data, but the problem is not only the complexity and size of the data, but the private nature of it: according to the project agreement, only a limited subset of researchers can have access to the raw data. The data is also not necessarily of very high quality as, e.g., the video recorded with the web cameras has a very low resolution. We

[4] http://research.ics.aalto.fi/cbir/data/

are currently focusing our efforts on the analysis of the audio data, and hope to be able to give structure to the massive data set through automatic annotation of daily events.

References

[1] Satoru Ishikawa, Markus Koskela, Mats Sjöberg, Jorma Laaksonen, Erkki Oja, Ehsan Amid, Kalle Palomäki, Annamaria Mesaros, and Mikko Kurimo. PicSOM experiments in TRECVID 2013. In *Proceedings of the TRECVID 2013 Workshop*, Gaithersburg, MD, USA, November 2013.

[2] Subhradeep Kayal. Experiments on the LFW database using curvelet transforms and a random forest-kNN cascade. In *Digital Information Processing and Communications (ICDIPC), 2012 Second International Conference on*, pages 146–149, Klaipeda, Lithuania, July 2012.

[3] He Zhang, Zhirong Yang, Mehmet Gönen, Markus Koskela, Jorma Laaksonen, Timo Honkela, and Erkki Oja. Affective abstract image classification and retrieval using multiple kernel learning. In *Proceedings of 20th International Conference on Neural Information Processing (ICONIP 2013)*, Daegu, South Korea, 2013. Springer.

[4] Jussi Tarvainen, Mats Sjöberg, Stina Westman, Jorma Laaksonen, and Pirkko Oittinen. Content-based prediction of movie style, aesthetics and affect. *IEEE Transactions on Multimedia*, 2013. Submitted.

[5] Jussi Tarvainen, Stina Westman, and Pirkko Oittinen. Stylistic Features for Affect-based Movie Recommendations. *Proc. Fourth International Workshop on Human Behavior Understanding*, pages 52–63, 2013.

[6] G.B. Huang, Q.Y. Zhu, and C.K. Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[7] C.H. Demarty, C. Penet, M. Schedl, B. Ionescu, V.L. Quang, and Y.G. Jiang. The MediaEval 2013 Affect Task: Violent Scenes Detection. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[8] A. Oulasvirta, A. Pihlajamaa, J. Perkiö, T. Vähäkangas, D. Ray, N. Vainio, P. Myllymäki, T. Hasu, Long-term Effects of Ubiquitous Surveillance at Home. In Proceedings of the 14th International Conference on Ubiquitous Computing (Ubicomp), September, 2012.

[9] Debarshi Ray, Data gathering in digital homes. M.Sc. thesis, University of Helsinki, Department of Computer Science, 2012.

Chapter 6

F2: Computational Molecular Biology and Medicine

Erik Aurell, Jukka Corander, Samuel Kaski, Antti Honkela, Lu Cheng, Brandon Malone, Sohan Seth, Martin Skwark, Johan Pensar, Onur Dikmen, Michael Gutmann, Väinö Jääskinen, Luca Martino, Pekka Marttinen, Elina Numminen, Ville Parkkinen, Jukka Sirén, Lu Wei, Jie Xiong, Hande Topa, Kai Brügge, Muhammad Ammad-ud-din, José Caldas, Ritabrata Dutta, Ali Faisal, Elizabeth Georgii, Jussi Gillberg, Suleiman A. Khan, Juuso Parkkinen, Sohan Seth, Tommi Suvitaival, Seppo Virtanen, and Erik Aurell's group in Stockholm

6.1 Introduction

The research activities in F2 involve the development of stochastic models and related inference algorithms for computational biology and medicine, as well as their application to real data in collaboration with biological experts. The groups in COIN that have been involved in F2 have different backgrounds ranging over Biophysics (Aurell), Statistics (Corander) and Machine learning (Kaski). An effort was therefore made to bring the groups together by holding regular sub-project meetings. These efforts have born fruit in the form of joint papers, which are now in a stage of submission or already published, as described below. We foresee many opportunities for future collaborations in this exciting field. In the last year we have also seen involvement from members of most of the other research groups in COIN, either by participating in the COIN F2 meetings, or in discussions on future projects, or already in ongoing research.

Pekka Marttinen acted as F2 coordinating postdoc from project start until the summer of 2013, after which Marcin Skwark took over this role. The group of Aurell at Aalto is much smaller than the ones of Corander and Kaski, but this imbalance has to some extent been countered by the involvement of Aurell's second research group at KTH - Royal Institute of Technology in Stockholm, Sweden.

6.2 Metagenomics

It has become increasingly clear that the species composition and diversity of bacterial communities are important components of human health. Altogether a human body contains approximately 1-2 kg of bacterial cells by weight, and the numbers of such cells are typically 1-2 orders of magnitude larger than the number of human cells. The estimation of the composition of such bacterial communities, as well the composition of samples from soil and others parts of the environment, is hence an important task which has until very recently been out of reach. Classically, bacteria were discovered and classified by their morphology as seen through a microscope, and then characterized in culture. This is however, even when proper culture conditions are known, not an immediate process, often requiring about a week or more in a hospital setting, for instance to determine the disease agent in a bacterial infection.

Modern high-throughput sequencing techniques promise to give an unbiased view of all the components of a bacterial community, either based on whole genomes or some important features. Jukka Corander and his group have developed two very accurate Bayesian unsupervised methods (BeBAC,BACDNAS, see below) for analyzing high-throughput data on the 16S ribosomal RNA gene which exist in all bacteria and which has been the basis of many previous classification methods. While this method is state-of-the-art it is also somewhat slow taking on the order of days of computation time on realistic problems, and faster though less accurate methods are therefore also of value.

Spectacular progress has been made over the last ten years in sparse signal processing and especially in compressive sensing, where a source is determined from what is (superficially) too few observations. After extensive discussion at COIN F2 meetings we have developed a method SEK which solves the same problem as BeBAC with slightly lower accuracy, but in seconds. The main idea of SEK is to reduce the data to the abundances of k-mers (subsequences of finite length) in windows along the 16S gene and then to formulate first

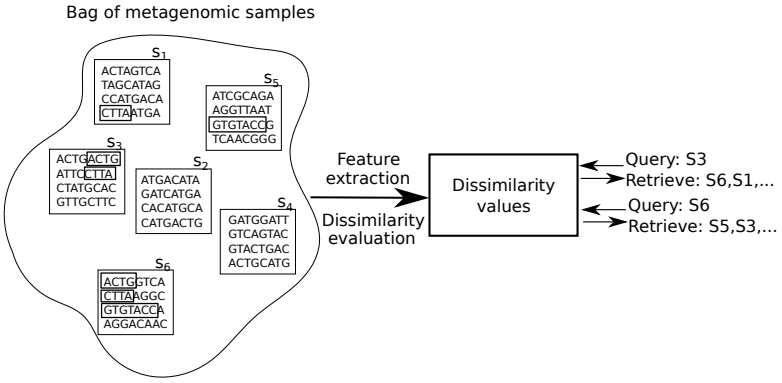


Figure 6.1: Retrieval of whole metagenome sequencing samples

order moment matching as a sparse reconstruction problem which can be solved efficiently. Extensions of the method involving higher moment matching and other feature vectors are under development.

In a parallel development Aurell has started a collaboration on the analysis of compressed sending algorithms in generalized settings with Yoshiyuki Kabashima of Tokyo Tech, Japan, which has resulted in one joint publication. Prof Kabashima is a co-PI of a new JSPS/MEXT "Initiative for High-Dimensional Data-driven Science through Deepening of Sparse Modeling" (http://www.sparse-modeling.jp/index_e.html) for which Aurell is an external advisor. We foresee extensive collaboration with Prof Kabashima and his group in the coming years.

We are also working on simultaneously extracting species, pathway and enzyme information from massive metagenomic datasets using non-negative matrix factorization (NMF). Unlike traditional NMF, the new algorithm factorizes the data into three matrices, which capture the species, pathway and enzyme information, respectively. Despite theoretical non-identifiability concerns, initial results show the method can extract useful information when given sufficient data.

In another line of work [4] (Fig. 6.1), we have developed a content-based retrieval method for whole metagenome sequencing samples. We apply a distributed string mining framework to efficiently extract all informative sequence k-mers from a pool of metagenomic samples, and use them to measure the dissimilarity between two samples. We evaluated the performance of the proposed approach on two human gut metagenome data sets and observe significant enrichment for diseased samples in results of queries with another diseased sample.

References

[1] Saikat Chatterjee, David Koslicki, Siyuan Dong, Nicolas Innocenti, Lu Cheng, Yueheng Lan, Mikko Vehkaperä, Mikael Skoglund, Lars K. Rasmussen, Erik Aurell and Jukka Corander SEK: Sparsity exploiting k-mer-based estimation of bacterial community composition *Submitted to Bioinformatics* [second revision]

[2] Cheng, L., Walker, A. and Corander, J. Bayesian estimation of bacterial community composition from 454 sequencing data. *Nucleic Acids Research*, doi:

10.1093/nar/gks227. (2012)

[3] Jääskinen, V., Parkkinen, V., Cheng, L., Corander, J. Bayesian clustering of DNA sequences using Markov chains and a stochastic partition model. *Statistical Applications in Genetics and Molecular Biology*, doi:/10.1515/sagmb-2013-0031 (2013)

[4] Sohan Seth, Niko Välimäki, Samuel Kaski, and Antti Honkela. Exploration and retrieval of whole-metagenome sequencing samples. *arXiv:1308.6074 [q-bio.GN]*, Submitted to a journal.

[5] Mikko Vehkaperä, Yoshiyuki Kabashima, Saikat Chatterjee, Erik Aurell, Mikael Skoglund, Lars Rasmussen Analysis of Sparse Representations Using Bi-Orthogonal Dictionaries *IEEE ITW*, 2012. [arXiv:1204.4065]

6.3 Protein structure prediction by direct coupling analysis

To predict the structure of a protein from its amino acid sequence has been a central goal of computational biology since the 1970ies. Today this is usually possible if the protein of interest is similar to another protein for which the structure has been determined experimentally, an approach known as homology modelling. Ab initio modelling, based only on the sequence, has however remained an unsolved problem even though the force fields between amino acids are largely known, and very large computational resources have been used. One reason for this difficulty is that protein folding happens on a time scale of microseconds or even milliseconds, and it is therefore inherently hard to "integrate out" the motion on the atomic level from the scale of picoseconds; another is presumably that even small inaccuracies in the descriptions of the force fields can matter for such large and complex processes.

The field has been revolutionized in the last five years when it was realized that the parameters of exponential models representing the amino acid sequences in a multiple sequence alignment can be used as very accurate predictors for the spatial proximity of pairs of amino acids in a structure. By this approach, known as *direct coupling analysis*, protein structure prediction is turned into a problem of computational inference. While we did not invent this approach, we have contributed plmDCA which is the most accurate stand-alone method available, and which has been taken up by other groups such as EVfold (http://evfold.org/). The key idea of plmDCA is to learn the model parameters by a pseudo-likelihood maximization instead of by variational inference, as would otherwise seem natural for an exponential family, see *e.g.* [1]. Marcin Skwark, in work done before but published after he joined COIN, is the lead author of an ensemble method PconC which combines plmDCA and other predictors to achieve the highest published method accuracy to date.

Over the last year we have developed a new version of plmDCA which is many times faster than the previous method by using a different output routine [4]. We have also, for the first time, introduced more than bi-linear (pair-wise) terms in the exponential model and shown that this leads to prediction accuracy comparable to PconC, but with a stand-alone method [5]. Marcin Skwark, Erik Aurell, Jukka Corander and Johan Pensar have an ongoing project to develop a Bayesian approach to the same problem, and to evaluate it against plmDCA and other methods, and potentially also to include it in a metapredictor such as PconC.

The world-wide protein structure prediction community organizes the bi-annual CASP competitions where a COIN team led by Marcin Skwark plans to particpate this year.

References

[1] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2), 1–305, December 2008.

[2] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, Erik Aurell Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models *Phys. Rev. E*, 87, 012707 (2013)

[3] Marcin J Skwark, Abbi Abdel-Rehim, Arne Elofsson PconsC: Combination of direct information methods and alignments improves contact prediction. *Bioinformatics* 15;29(14):1815-6 (2013)

[4] Magnus Ekeberg, Tuomo Hartonen, Erik Aurell Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences Submitted to *J Comp Physics* (2014) [arXiv:1401.4832]

[5] Christoph Feinauer, Marcin J. Skwark, Andrea Pagnani, Erik Aurell Improving contact prediction along three dimensions Submitted to *PLoS Comp Biology* (2014) [arXiv:1403.0379]

6.4 Computational inference for microbiology and infectious disease epidemiology

Bacteria and viruses are an inevitable part of all life on earth, but they also pose a considerable threat to human and animal health. Recently, resistance to antimicrobial agents has become a widespread problem in health care, in particular nosocomial infections have escalated in certain regions, causing significant losses of human life. One of the major reasons for rapid spread of antibiotic resistance is horizontal gene transfer through bacterial recombination, which allows acquisition of novel genome elements from other evolutionary lineages within a named species or alternatively from other species. Recombination plays also a central role in the adaptation of bacteria into novel niches. We have developed statistical methods for the study of recombinogenic bacteria by using either limited core gene variation or whole-genome sequences. Given the high rate of diversification of many bacteria, whole-genome data pose a tremendous challenge for inference algorithms when horizontal gene transfer needs to be acknowledged or explicitly modeled. Our Bayesian population genomic methods implemented in software packages BAPS and BratNextGen have gained considerable popularity for analyses of bacterial genome data. Given that a single multiple genome alignment may contain up to hundreds of thousands of variable positions and currently even thousands of bacteria, fitting Bayesian models to such data cannot be reliably done using any standard algorithms such as Gibbs sampler or basic Metropolis-Hastings. Our most recent update to the stochastic optimization algorithm in BAPS software has made model fitting an order of magnitude faster for large genome data sets, compared to the earlier version. Similarly, the use of large-scale parallel computation

has enabled the method implemented in BratNextGen to become the fastest available Bayesian method for estimating recombinations in bacterial genome data. The other currently available Bayesian methods are applicable only to data sets that are an order of magnitude smaller than those still handled by BratNextGen. Using these statistical tools in collaboration with biologists, we have made several important discoveries about the evolution of bacteria and transmission of resistance. In particular the two recent papers published in Nature Genetics present analyses of the largest bacterial sequence data sets ever produced, and highlight the importance of scalable inference methods to enable biological breakthroughs.

References

[1] Casali, N. et al. Evolution and transmission of drug resistant tuberculosis in a population: Insights from a 1000 genome study. *Nature Genetics*, doi:10.1038/ng.2878. (2014)

[2] Castillo-Ramírez S, et al. Linking founder events with regional variation in recombination rates within a global clone of Methicillin Resistant Staphylococcus aureus (MRSA). *Genome Biology*, 13:R126. doi:10.1186/gb-2012-13-12-r126. (2012)

[3] Cheng, L., Connor, T.R., Sirén, J., Aanensen, D.M. and Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular Biology and Evolution*, doi: 10.1093/molbev/mst028. (2013)

[4] Chewapreecha, C. et al. Dense genomic sampling identifies highways of pneumococcal recombination. *Nature Genetics*, doi:10.1038/ng.2895. (2014)

[5] Connor, T.R., Corander, J. and Hanage, W.P. Population subdivision and the detection of recombination in non-typable Haemophilus influenzae. *Microbiology*, 158, 2958-2964. (2012)

[6] de Been, M., van Schaik, W., Cheng, L., Corander, J. and Willems, R.J.L. Recent recombination events in the core genome are associated with adaptive evolution in Enterococcus faecium. *Genome Biology and Evolution*, doi: 10.1093/gbe/evt111. (2013)

[7] Lebreton, F. et al. Emergence of epidemic multi-drug resistant Enterococcus faecium from animal and commensal strains. *mBio*, doi: 10.1128/mBio.00534-13. (2013)

[8] Marttinen, P., Hanage, W.P., Croucher, N., Connor, T.R., Harris, S., Bentley, S. and Corander J. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Research*, 40: e6. doi: 10.1093/nar/gkr928. (2012)

[9] McNally, A., Cheng, L., Harris, S.R. and Corander, J. The evolutionary path to extra intestinal pathogenic, drug resistant Escherichia coli is marked by drastic reduction in detectable recombination within the core genome. *Genome Biology and Evolution*, 5: 699-710. (2013)

[10] Willems, R.J.L. et al. Restricted gene flow among hospital subpopulations of Enterococcus faecium. *mBio*, 3, e00151-12. (2012)

[11] Reuter, S. et al. Parallel independent evolution of pathogenicity within the genus Yersinia. *PNAS*, in press. (2014)

[12] Sheppard, S.K. et al. Cryptic ecology among host generalist Campylobacter jejuni in domestic animals. *Molecular Ecology*, in press. (2014)

6.5 Probabilistic models of gene expression dynamics and RNA-seq

Transcript isoform level RNA-seq data analysis We have developed BitSeq, a Bayesian probabilistic method for analysis of RNA-sequencing (RNA-seq) data on the level of alternatively spliced transcript isoforms. BitSeq consists of two stages: in stage 1 we estimate the expression level of different transcripts, while in stage 2 we perform differential expression testing on isoform level utilising information from biological replicates. The BitSeq model is based on probabilistically assigning the reads to their transcripts of origin, taking fully into account multiply mapping reads, sequencing errors, as well as biases in the sequencing process. As illustrated in Figs. 6.2 and 6.3, BitSeq has state-of-the-art performance in both of these problems. Our paper [1] was the first to present a solution to the latter problem.

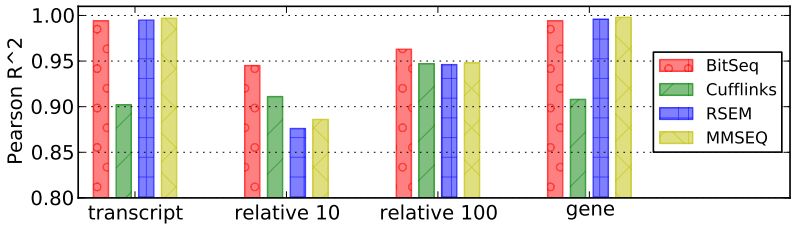


Figure 6.2: Comparison of BitSeq expression estimation accuracy (stage 1) using synthetic data.

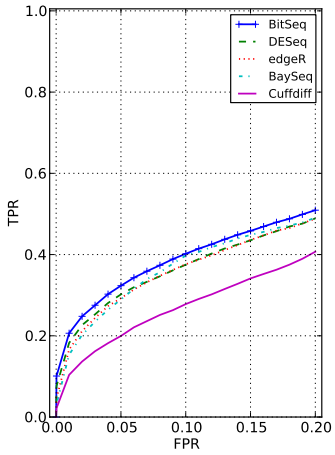


Figure 6.3: ROC curve from a comparison of BitSeq differential expression estimation accuracy (stage 2) using synthetic data.

More recently, we have applied a modern Riemannian collapsed variational Bayesian (VB) learning algorithm to the estimation of the model [2]. The VB algorithm can produce equally accurate estimates of mean expression as the original Gibbs sampler in significantly

shorter time, but the underestimation of variance makes differential expression analysis less reliable.

Modelling gene transcription and expression dynamics using Gaussian processes Gaussian processes (GPs) are an ideal tool for modelling genomic time series which are often short and unevenly sampled. In our earlier work we combined GPs with a linear ordinary differential equation model of gene transcription and used this model very successfully in ranking candidate targets of gene regulators called transcription factors (TFs) [3].

In [4], the same work is extended to non-linear regulation models with multiple regulators. Example model fits are illustrated in Fig. 6.4 and an evaluation of the accuracy of the regulator predictions in terms of percentage of predicted targets with evidence of TF binding near the gene is shown in Fig. 6.5.

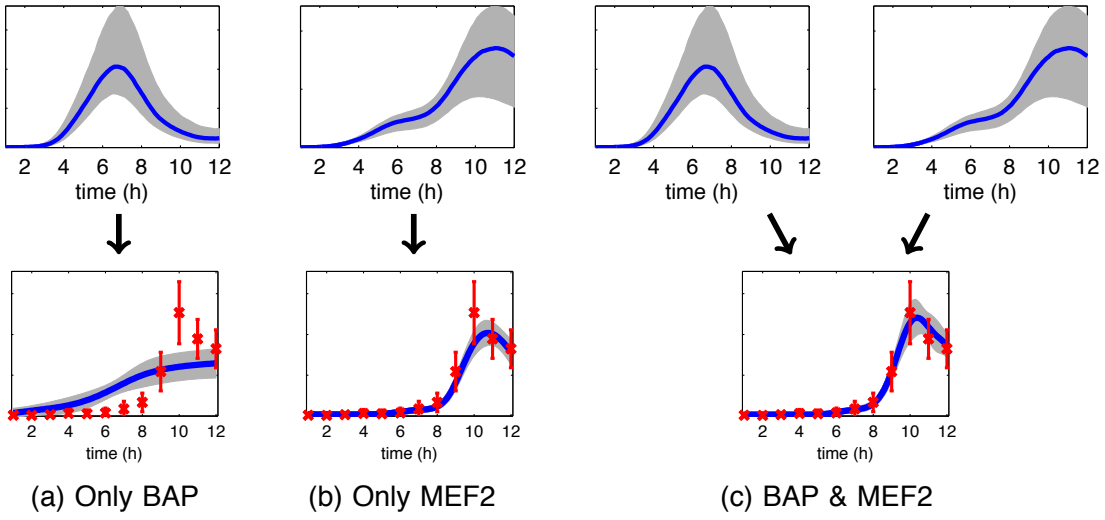
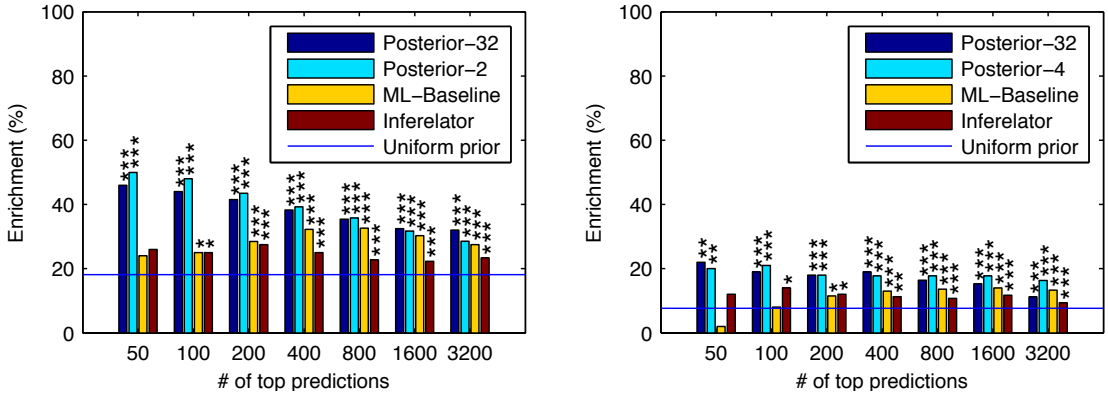


Figure 6.4: Illustration of modelling cooperatively gene regulation by two TFs. Red crosses show target gene expression data (12 time points) and blue lines show model predictions and associated credible regions. In the top row we show the activity profiles for each TF which are inferred during the training phase by fitting a regulation model on a network of known structure. In the bottom row we show the model fit during genome-wide scanning for this target gene. We show the target mRNA concentration profile inferred by fitted models of (a) regulation by BAP only, (b) regulation by MEF2 only and (c) regulation by BAP and MEF2. The candidate gene is confirmed as a joint target by independent ChIP-chip studies.

In [5], we have applied GPs to model the dynamics of RNA polymerase II (pol-II) in gene transcription. The model can be used to compute RNA transcription speed and infer the temporal pol-II activity at the gene promoter. We used the inferred promoter activity profile to determine genes that are responding in a coordinated manner to stimuli and are therefore potentially co-regulated. Employing this kind of modelling can significantly increase the accuracy of regulatory network inference in rapid signalling systems, where transcriptional and other delays in the system would otherwise make it difficult to reliably link causes and effects.



(a) Predicting regulation by one TF (b) Predicting regulation by two TFs

Figure 6.5: Enrichment of confident regulator predictions for ChIP binding. Plots show percentage of top ranked confident regulator predictions that had confirmed bindings by predicted regulators within 2000 base pairs of the putative target gene. Predictions were ranked by the posterior probability of (a) regulation by any single regulator; or (b) joint regulation by any two regulators. Both plots include rankings according to the marginal posterior probability of a set of regulators being active computed over all 2^5 models of 5 regulators (dark blue bars), posterior probability over a restricted set of models ignoring all other TFs leaving 2 models for single regulator and 4 models for two regulators (light blue bars) as well as maximum likelihood-based baseline model (yellow bars) and the Inferelator (red bars), compared to predicting regulators uniformly at random (blue line; link probability 0.5). p-values of results statistically significantly different from random are denoted by '***': $p < 0.001$, '**': $p < 0.01$, '*': $p < 0.05$.

References

[1] P. Glaus, A. Honkela, and M. Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728, Jul 2012.

[2] J. Hensman, P. Glaus, A. Honkela, and M. Rattray. Fast approximate inference of transcript expression levels from RNA-seq data. Aug. 2013. arXiv:1308.5953 [q-bio.GN].

[3] 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.

[4] M. K. Titsias, A. Honkela, N. D. Lawrence, and M. Rattray. Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison. *BMC Syst Biol*, 6:53, 2012.

[5] C. wa Maina, A. Honkela, F. Matarese, K. Grote, H. G. Stunnenberg, G. Reid, N. D. Lawrence, and M. Rattray. Inference of RNA polymerase II transcription dynamics from chromatin immunoprecipitation time course data. *PLOS Comput Biol*, 2014. Accepted for publication. Also available as arXiv:1303.4926 [q-bio.QM].

6.6 Probabilistic models of multiple data sources

Chemical systems biology Analysis of genome-wide effects of drugs is a central challenge for developing and tailoring modern treatments. Here the Connectivity Map (CMap) data set is particularly useful; it is a publicly available large collection of high-throughput molecular profiling measurements from drug-treatments on human cancer cell lines. We have addressed the problem of modelling the relationships between chemical structures of drugs causing specific gene expression responses. We started with Canonical correlation analysis to detect statistical dependencies between chemical descriptors and gene expression measurements [7], Figure 6.6. Later, we applied the novel multi-view data integration method (see C2), group factor analysis, to study multiple cancer types [6], creating testable predictions, and extended the work further to take into account the tensorial nature (drugs, genes, cancers) into account.

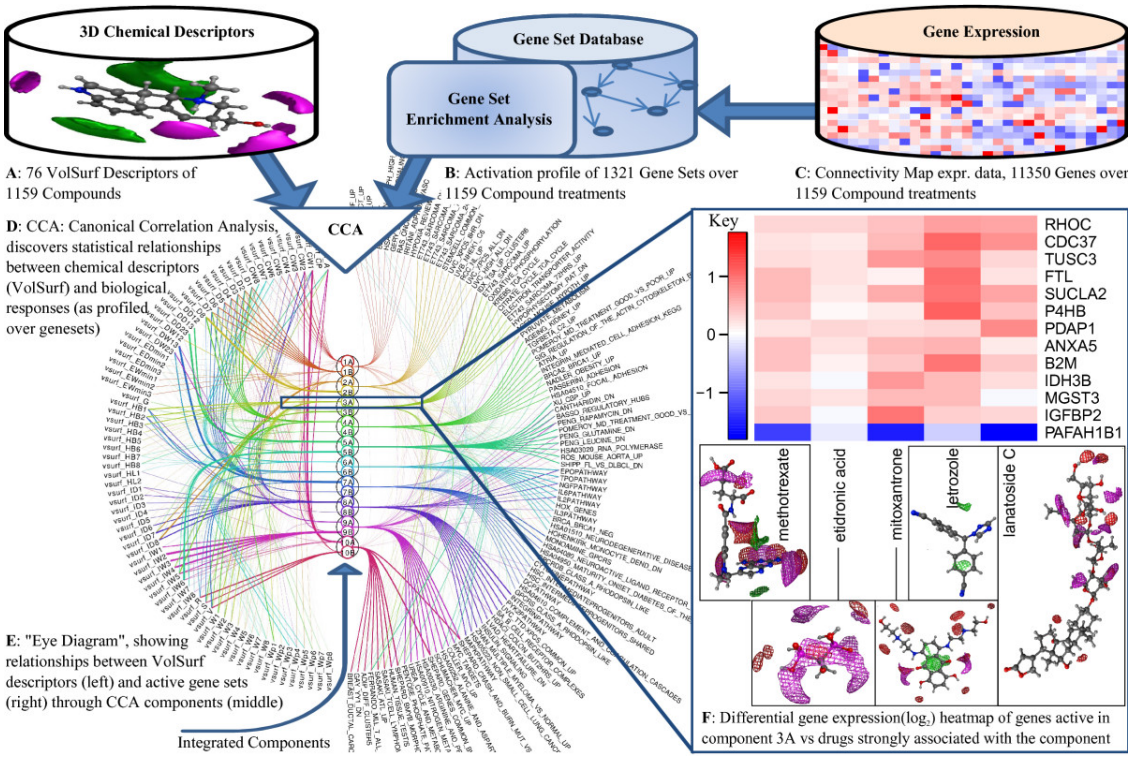


Figure 6.6: Data-driven search for statistical relationships between Chemical space (formed of VolSurf features) and Drug response space (gene expression).

In *toxigenomics* the goal is to identify associations between gene expression responses and toxicological outcomes, of drugs or more generally any potentially toxic chemicals. We participated [10] in the CAMDA 2013 toxicogenomic challenge where we detect cross-organism toxicogenomic associations based on group factor analysis (C2) and are finalizing work on a probabilistic toxicogenomics approach that effectively predicts human in vitro toxicity based on transriptomic response data from drug-treatments.

Metabolomics Metabolomics, analysis of metabolite levels in an organism, shares with genomics the problem of large p, small n, of having a large number p of potentially important variables compared to the sample size n. Probabilistic modelling and integration of evidence from multiple sources help solve the problem. We have developed more accurate

methods for quantifying metabolite levels from mass spectrometry, by combining multiple observations (mass spectrometry peaks) from the same molecules [11] and from correlating molecules (on-going work), in collaboration with University of Glasgow (Dr. Simon Rogers).

Genome-wide association studies A typical genome-wide association study (GWAS) searches for associations between single nucleotide polymorphisms (SNPs) and a univariate phenotype. However, there is a growing interest to investigate associations between genomics data and multivariate phenotypes, for example in gene expression or metabolomics studies.

Our work [8] is the first comprehensive comparison of existing approaches to GWASs with a rich phenotype using metabolite data. Motivated by this comparison we developed a new approach [9] that simultaneously tackles covariance in the high-dimensional phenotype induced by structured noise, and covariance induced by genes affecting multiple phenotype variables. The new approach detects new associations (with replications in other data) that are invisible to previous methods. Replication results for two new associations discovered with the new method are presented in Figure 6.7. Weak effects are the quintessential problem of GWASs: single covariates explain a small amount, less than $\approx 1\%$, of the variance of the target variables. In our unpublished work we address the problem of weak effects in a multiple-output regression setup and improve performance in genomic prediction by introducing a new principle of sharing information between the regression model and the explain-away model.

Personalized medicine With the recent biotechnological advances in large scale molecular profiling of cells (either extracted from patients or grown in cultures), it is now possible to build and test computational models for in-depth analysis of molecular biology of the disease and to predict most effective medicine. Hence, at the core of personalized medicine there is a computational problem: Given a set of molecular profiles of cells, for which some measurements of treatment outcomes exist, predict treatment outcomes for a new cell [2]. We have, in collaboration with the Institute for Molecular Medicine Finland FIMM, developed novel probabilistic multi-source machine learning methods (see C2) and demonstrated the usefulness of these methods in the competition NCI-DREAM 2012 Drug Sensitivity Prediction Challenge. Our methods showed the best predictive performance by outperforming other state-of-art methods proposed by 41 international teams [3]. The specific goal of the crowd-sourced competition was to predict effectiveness of the drugs on new cells based on omics measurements. Our method showed that combining multiple sources of information for the cells with the appropraite use of prior biological knowlegde, is essentially the key to make personalized predictions in cancer.

We further extended this line of research by proposing novel methods (see C2) that not only uitilize the omics measurements but can additionally incorporate chemical properties of drugs. This is necessary for making predictions for new drugs on existing cells; a step towards in-silico drug discovery for cancer. We showed that supplementing the models with chemical properties of the drugs improved the prediction performance [5]. More recently, we addressed a novel task of predicting responses of completely new drugs on new cells, in collaboration with Institute for Molecular Medicine Finland FIMM (manuscript submitted). This task challenging and required several different types of side information sources to enhance the prediction performance.

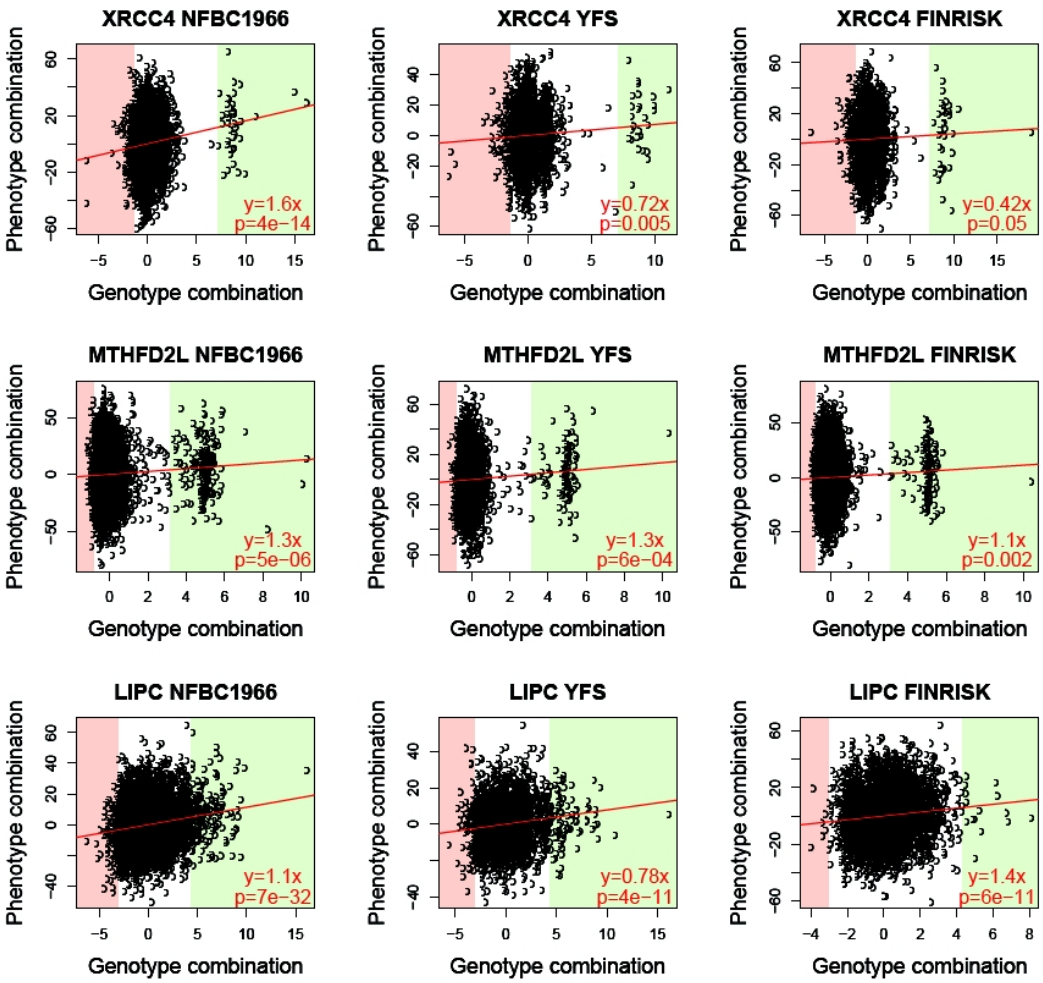


Figure 6.7: Genes XRCC4 and MTHFD2L were found to be associated with lipid metabolism. The well-known LIPC lipid locus is also shown for reference. Each panel shows the identified phenotype combination plotted against the genotype combination. The left column shows results in the NFBC1966 data set, in which the associations were detected. The center and right columns show replication results with the YFS and FINRISK data sets, where coefficient matrices learned with the NFBC1966 data were used to form the variable combinations.

Retrieval of relevant experiments We introduced the concept of retrieval of experiments (details in C2) motivated by the question of how to maximally benefit from the public repositories of molecular biological data. The motivation is that when a biologist makes a new experiment, such as a gene expression measurement, it would be useful to relate the results to earlier research, at best on the level of actual measurement data, and the data are available in the current large databases. This is content-based search, complementing the standard annotation-based searches, but content-based search where it is crucial to capture existing biological knowledge in the relevance metric. That we do based on probabilistic latent variable models that captures both relevant activity in the data and prior information, and retrieval is then performed in the model space. In addition to providing relevant search results, the model-based method helps in interpretating the results. For example, a previously unknown connection between SIM2s gene

and malignant mesothelioma suggested by the model was experimentally validated [1]. Another model was introduced for targeting the search specifically to relevant regulatory relationships between genes [4]. We have additionally applied the model-based retrieval framework to drug connectivity mapping, where the task is to find similar drugs from the CMap database based on their expression profiles. By using group factor analysis (see C2) to separate specific and shared effects between the multiple available cell lines we achieved superior retrieval performance compared to earlier connectivity mapping methods (manuscript submitted).

References

[1] José Caldas, Nils Gehlenborg, Eeva Kettunen, Ali Faisal, Mikko Rönty, Andrew G. Nicholson, Sakari Knuutila, Alvis Brazma, and Samuel Kaski. Data-driven information retrieval in heterogeneous collections of transcriptomics data links SIM2s to malignant pleural mesothelioma. *Bioinformatics*, 28(2):246–253, 2012. Supplementary data and source code are available from http://www.ebi.ac.uk/fg/research/rex.

[2] Jukka Corander, Tero Aittokallio, Samuli Ripatti, and Samuel Kaski. The rocky road to personalized medicine: computational and statistical challenges. *Personalized Medicine*, 9(2):109–114, 2012.

[3] J. C. Costello et al. *Nature Biotechnology*, Accepted for publication.

[4] Elisabeth Georgii, Jarkko Salojärvi, Mikael Brosché, Jaakko Kangasjärvi, and Samuel Kaski. Targeted retrieval of gene expression measurements using regulatory models. *Bioinformatics*, 28(18):2349–2356, 2012.

[5] Mehmet Gönen. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics*, 28(18):2304–2310, 2012.

[6] S.A. Khan, S. Virtanen, O. Kallioniemi, K. Wennerberg, A. Poso, and S. Kaski. Identification of structural features in chemicals associated with cancer drug response: A systematic data-driven analysis. arXiv: 1312.7734v1, 2013.

[7] Suleiman A. Khan, Ali Faisal, John Patrick Mpindi, Juuso A. Parkkinen, Tuomo Kalliokoski, Antti Poso, Olli P. Kallioniemi, Krister Wennerberg, and Samuel Kaski. Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs. *BMC Bioinformatics*, 13(112), 2012.

[8] P. Marttinen, J. Gillberg, A. Havulinna, J. Corander, and S. Kaski. Genome-wide association studies with high-dimensional phenotypes. *Statistical Applications in Genetics and Molecular Biology*, 12(4):413–431, 2013.

[9] Pekka Marttinen, Matti Pirinen, Antti-Pekka Sarin, Jussi Gillberg, Johannes Kettunen, Ida Surakka, Antti J. Kangas, Pasi Soininen, Paul O'Reilly, Marika Kaakinen, Mika Kähönen, Terho Lehtimäki, Mika Ala-Korpela, Olli T. Raitakari, Veikko Salomaa, Marjo-Riitta Järvelin, Samuli Ripatti, and Samuel Kaski. Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression. *Bioinformatics*, to appear.

[10] Tommi Suvitaival, Juuso A. Parkkinen, Seppo Virtanen, and Samuel Kaski. Cross-organism prediction of drug hepatotoxicity by sparse group factor analysis. In *12th Annual International Conference on Critical Assessment of Massive Data Analysis (CAMDA)*, 2013.

[11] Tommi Suvitaival, Simon Rogers, and Samuel Kaski. Stronger findings from mass spectral data through multi-peak modeling. arXiv:1403.4732 [q-bio.QM].

Publications of COIN 2012-2013

[1] M. Almeida, J. Dias, R. V. Vigário, and E. Oja. A comparison of algorithms for separatino of synchronous subspaces. *Bulletin of the Polish Academy of Sciences*, 60(3): 455–460, January, 2012.

[2] A. Ajanki, M. Koskela, J. Laaksonen, and S. Kaski. Adaptive timeline interface to personal history data. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 229–236, New York, NY, USA. ACM, 2013.

[3] E. Aurell and M. Ekeberg. Inverse Ising inference using all the data. *Phys. Rev. Lett.*, 108: 090201, 2012.

[4] E. Aurell and H. Mahmoudi. Dynamic mean-field and dynamic cavity for diluted Ising systems. *Phys. Rev. E*, 85: 031119, 2012.

[5] E. Aurell, C. Mejía-Monasterio, and P. Muratore-Ginanneschi. Boundary layers in stochastic thermodynamics. *Phys. Rev. E*, 85: 020103(R), 2012.

[6] E. Aurell, K. Gawędzki, C. Mejía-Monasterio, R. Mohayaee, and P. Muratore-Ginanneschi. Refined Second Law of Thermodynamics for fast random processes. *J. Stat. Phys.*, 147(3), 2012.

[7] E. Aurell. The physics of distributed information systems. *J. Phys.: Conf. Ser.*, 473: 012017, 2013.

[8] A. Balint, A. Belov, D. Diepold, S. Gerber, M. Järvisalo, and C. Sinz (eds.). *Proceedings of SAT Challenge 2012*, Volume B-2012-2 of Department of Computer Science Series of Publications B. University of Helsinki, 2012.

[9] M. de Been, W. van Schaik, L. Cheng, J. Corander, and R.J.L. Willems. Recent recombination events in the core genome are associated with adaptive evolution in Enterococcus faecium. *Genome Biology and Evolution*, 5(8): 1524–1536, 2013.

[10] A. Belov, M. Järvisalo, and J. Marques-Silva. Formula preprocessing in MUS extraction. In Nir Piterman and Scott Smolka, editors, *Proceedings of the 19th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS 2013)*, volume 7795 of *Lecture Notes in Computer Science*, pages 110–125. Springer, 2013.

[11] J. Berg and M. Järvisalo. Optimal correlation clustering via MaxSAT. In Wei Ding, Takashi Washio, Hui Xiong, George Karypis, Bhavani M. Thuraisingham, Diane J. Cook, and Xindong Wu, editors, *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW 2013)*, pages 750–757. IEEE Press, 2013.

[12] M. Berglund, T. Raiko, and K. Cho. Measuring the Usefulness of Hidden Units in Boltzmann Machines with Mutual Information In *Neural Information Processing*, volume 8226 of *Lecture Notes in Computer Science*, pages 482–489. Springer, 2013.

[13] S. Bhattacharya, S. Phithakkitnukoon, P. Nurmi, A. Klami, M. Veloso, and C. Bento. Gaussian process-based predictive modeling for bus ridership. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication (UbiComp'13 Adjunct)*, pages 1189–1198, 2013.

[14] S. Bhattacharya, P. Floréen, A. Forsblom, S. Hemminki, P. Myllymäki, P. Nurmi, T. Pulkkinen, and A. Salovaara. Ma$$ive- an intelligent mobile grocery assistant. In *Proceedings of Eighth International Conference on Intelligent Environments (IE 2012)*, pages 165–172, 2012.

[15] S. Bo, E. Aurell, R. Eichhorn, and A. Celani. Optimal stochastic transport in inhomogeneous thermal environments. *EPL*, 103: 10010, 2013.

[16] J. Bomanson and T. Janhunen. Normalizing cardinality rules using merging and sorting constructions. In *Logic Programming and Nonmonotonic Reasoning*, volume 8148 of *Lecture Notes in Computer Science*, pages 187–199. Springer, 2013.

[17] G. Bozkurt, M. Gönen, and F. Gürgen. Probabilistic and discriminative group-wise feature selection methods for credit risk analysis. *Expert Systems with Applications*, 39(14): 11709–11717, 2012.

[18] G.J. Brown, A. Beeston, K.J. Palomäki. Perceptual compensation for the effects of reverberation on consonant identification: A comparison of human and machine performance, In *Proc. Interspeech*, pages 1–4, 2012.

[19] R. Calandra, T. Raiko, F. Montesino Pouzols, and M.P. Deisenroth. Learning Deep Belief Networks from Non-Stationary Streams. In *Proceedings of Artificial Neural Networks and Machine Learning - ICANN 2012*, volume 7553 of *Lecture Notes in Computer Science*, pages 379–386. Springer, 2012.

[20] J. Caldas, N. Gehlenborg, E. Kettunen, A. Faisal, M. Rönty, 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): 246–253, 2012. Supplementary data and source code are available from http://www.ebi.ac.uk/fg/research/rex.

[21] S. Castillo-Ramírez S, et al. Linking founder events with regional variation in recombination rates within a global clone of Methicillin Resistant Staphylococcus aureus (MRSA). *Genome Biology*, 13:R126. doi:10.1186/gb-2012-13-12-r126, 2012.

[22] A. Celani, S. Bo, R. Eichhorn, and E. Aurell. Anomalous thermodynamics at the micro-scale *Phys. Rev. Lett.*, 109: 260603, 2012.

[23] X. Chen and M. Koskela. Classification of RGB-D and motion capture sequences using extreme learning machine. In *Proceedings of the 18th Scandinavian Conference on Image Analysis*, volume 7944 of *Lecture Notes in Computer Science*, Espoo, Finland, Springer Verlag, June 2013.

[24] X. Chen and M. Koskela. Online RGB-D gesture recognition with extreme learning machines. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction (ICMI 2013)*, pages 467–474, Sydney, Australia, ACM, December 2013.

[25] X. Chen and M. Koskela. Sequence Alignment for RGB-D and Motion Capture Skeletons. In *Proceedings of the International Conference on Image Analysis and Recognition (ICIAR 2013)*, volume 7950 of *Lecture Notes in Computer Science*. Springer, 2013.

[26] L. Cheng, A. Walker, and J. Corander. Bayesian estimation of bacterial community composition from 454 sequencing data. *Nucleic Acids Research*, 40(12): 5240–5249, 2012.

[27] L. Cheng, T.R. Connor, J. Sirén, D.M. Aanensen, and J. Corander. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular Biology and Evolution*, 30(5): 1224–1228, 2013.

[28] K. Cho. Boltzmann Machines for Image Denoising. In *Proceedings of the International Conference on Artificial Neural Networks and Machine Learning (ICANN 2013)*, 2013.

[29] K. Cho. Simple Sparsification Improves Sparse Denoising Autoencoders in Denoising Highly Corrupted Images. In *Proceedings of the International Conference on Machine Learning (ICML 2013)*, 2013.

[30] K. Cho. Understanding Dropout: Training Multi-Layer Perceptrons with Auxiliary Independent Stochastic Neurons. In *Proceedings of the International Conference on Neural Information Processing (ICONIP 2013)*, 2013.

[31] K. Cho, T. Raiko, and A. Ilin. Enhanced Gradient for Training Restricted Boltzmann Machines. *Neural Computation*, 25(3), 2013.

[32] K. Cho, T. Raiko, and A. Ilin. Gaussian-Bernoulli Deep Boltzmann Machines. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2013)*, 2013.

[33] K. Cho, T. Raiko, A. Ilin, Alexander, and J. Karhunen. A Two-stage Pretraining Algorithm for Deep Boltzmann Machines. In *Proceedings of the International Conference on Artificial Neural Networks and Machine Learning (ICANN 2013)*, 2013.

[34] T.R. Connor, J. Corander, and W.P. Hanage. Population subdivision and the detection of recombination in non-typable Haemophilus influenzae. *Microbiology*, 158: 2958-2964, 2012.

[35] J. Corander, T. Aittokallio, S. Ripatti, and S. Kaski. The rocky road to personalized medicine: computational and statistical challenges. *Personalized Medicine*, 9(2):109–114, 2012.

[36] J. Corander, J. Xiong, Y. Cui, and T. Koski. Optimal Viterbi Bayesian Predictive Classification for Data from Finite Alphabets. *Journal of Statistical Planning and Inference*, doi:10.1016/j.jspi.2012.07.013, 2012

[37] J. Corander, T.R. Connor, C.A. O'Dwyer, J.S. Kroll, and W.P. Hanage. Population structure in the Neisseria, and the biological significance of fuzzy species. *Journal of the Royal Society Interface*, 9(71): 1208–1215, 2012.

[38] J. Corander, T. Koski, T. Pavlenko, and A. Tillander. Bayesian block-diagonal predictive classifier for Gaussian data. In R. Kruse, M. R. Berthold, C. Moewes, M. Á. Gil, P. Grzegorzewski, and O. Hryniewicz, editos, *Synergis of Soft Computing and Statistic for Intelligent Data Analysis*, volume 190 of *Advances in Intelligent Systems and Computing*, pages 543–551, Springer Berlin Heidelberg, 2013.

[39] J. Corander, K.K. Majander, L. Cheng, and J. Merilä. High degree of cryptic population differentiation in the Baltic Sea herring Clupea harengus. *Molecular Ecology*, 22(11): 2931–2940, 2013.

[40] J. Corander, T. Janhunen, J. Rintanen, H. Nyman, and J. Pensar. Learning chordal Markov networks by constraint satisfaction. In: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (eds.) *Advances in Neural Information Processing Systems* volume 26, pages 1349–1357, 2013.

[41] W. Delezuch, P. Marttinen, H. Kokki, M. Heikkinen, K. Vanamo, K. Pulkki, and I. Matinlauri, Serum and CSF soluble CD26 and CD30 concentrations in healthy pediatric surgical outpatients. *Tissue Antigens*, 80(4): 368–375, 2012.

[42] O. Dikmen and A. Mesaros. Sound Event Detection Using Non-negative Dictionaries Learned From Annotated Overlapping Events. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2013)*, pp. 4, 2013.

[43] J. Dubrovin, T. Junttila, and K. Heljanko. Exploiting step semantics for efficient bounded model checking of asynchronous systems. *Science of Computer Programming*, 77(10- 11): 1095–1121, 2012.

[44] W. Dvořák, M. Järvisalo, J. Wallner, and S. Woltran. CEGARTIX: A SAT-based argumentation system. In *Proceedings of the 3rd Workshop on Pragmatics of SAT*, 2012.

[45] W. Dvořák, M. Järvisalo, J.P. Wallner, and S. Woltran. Complexity-sensitive decision procedures for abstract argumentation. In Thomas Eiter and Sheila McIlraith, editors, *Proceedings of the 13th International Conference on Principles of Knowledge Representation and Reasoning (KR 2012)*, pages 54–64, AAAI Press, 2012.

[46] R. Eggeling, T. Roos, P. Myllymäki, and I. Grosse. Comparison of NML and Bayesian scoring criteria for learning parsimonious Markov models. In *Proceedings of the 5th Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-2012)*, 2012.

[47] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models *Phys. Rev. E* 87: 012707, 2013

[48] S. Enarvi and M. Kurimo. A Novel Discriminative Method for Pruning Pronunciation Dictionary Entries. In *Proceedings of the 7th International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013)*, pages 113–116, 2013.

[49] S. Enarvi and M. Kurimo. Studies on Training Text Selection for Conversational Finnish Language Modeling. In *Proceedings of the 10th International Workshop on Spoken Language Translation (IWSLT 2013)*, pages 256–263, 2013.

[50] A. Faisal, J. Gillberg, J. Peltonen, G. Leen, and S. Kaski. Sparse Nonparametric Topic Model for Transfer Learning. In *Proceedings of 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 269–274, 2012.

[51] R. Gauriot, L. Gunaratnam, R. Moroni, T. Reinikainen, and J. Corander. Statistical challenges in the quantification of gunshot residue evidence. *Journal of Forensic Sciences*, 58(5): 1149–1155.

[52] J.F. Gemmeke, and U. Remes. Missing-Data Techniques: Feature Reconstruction. In *Techniques for Noise Robustness in Automatic Speech Recognition*, pages 399–432, John Wiley & Sons, Ltd, 2012.

[53] E. Georgii and K. Tsuda. Density-Based Set Enumeration in Structured Data. In *Statistical and Machine Learning Approaches for Network Analysis*, pages 261–301. John Wiley & Sons, Inc., 2012.

[54] E. Georgii, J. Salojärvi, M. Brosché, J. Kangasjärvi, and S. Kaski. Targeted retrieval of gene expression measurements using regulatory models. *Bioinformatics*, 28(18): 2349–2356, 2012.

[55] C.D. Giurcaneanu, P. Luosto, and P. Kontkanen. On The Performance Of Histogram-Based Entropy Estimators. In *Proceedings of the 2012 IEEE International Workshop on Machine Learning for Signal Processing (MLSP'2012)*, Santander, Spain, September 2012.

[56] P. Glaus, A. Honkela, and M. Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13): 1721–1728, Jul 2012.

[57] D. Głowacka, T. Ruotsalo, K. Konyushkova, K. Athukorala, S. Kaski, and G. Jacucci. Directing exploratory search: Reinforcement learning from user interactions with keywords. In *18th International Conference on Intelligent User Interfaces (IUI)*, pages 117–128, New York, NY, ACM, 2013. Best paper award.

[58] D. Głowacka, T. Ruotsalo, K. Konyushkova, K. Athukorala, S. Kaski, and G. Jacucci. SciNet: A system for browsing scientific literature through keyword manipulation. In *IUI'13 Companion, International Conference on Intelligent User Interfaces*, pages 61–62, New York, NY, ACM, March 2013.

[59] D. Gowda and M. Kurimo. Analysis of breathy, modal and pressed phonation based on low frequency spectral density. In *Proc. INTERSPEECH*, pages 3206–3210, 2013.

[60] D. Gowda, J. Pohjalainen, M. Kurimo, and P. Alku. Robust formant detection using group delay function and stabilized weighted linear prediction. In *Proc. INTERSPEECH*, pages 49–53, 2013.

[61] D. Gowda, J. Pohjalainen, P. Alku, and M. Kurimo. Robust spectral representation using group delay function and stabilized weighted linear prediction for additive noise degradations. In *Proceedings of the 7th International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013)*, pages 135–141, 2013.

[62] C. Gulcehre, K. Cho, P. Razvan, and Y. Bengio. Learned-Norm Pooling for Deep Neural Networks. In *NIPS 2013 Workshop on Deep Learning*, Lake Tahoe, USA, 2013.

[63] A. Gusmão and T. Raiko. Reinforcement Learning in Real-Time Strategy Games (Extended Abstract). In *Proceedings of the Federated Computer Science Event (YTP 2012)*, pages 1–4, 2012.

[64] A. Gusmão and T. Raiko. Towards Generalizing the Success of Monte-Carlo Tree Search beyond the Game of Go. In *Proceedings of the European Conference on Artificial Intelligence (ECAI 2012), Frontiers in Artificial Intelligence and Applications*, volume 242, pages 384–389, 2012.

[65] M. Gönen. Bayesian Efficient Multiple Kernel Learning. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1–8, 2012.

[66] M. Gönen. A Bayesian Multiple Kernel Learning Framework for Single and Multiple Output Regression. In *Proceedings of the 20th European Conference on Artifical Intelligence*, pages 354–359, 2012.

[67] M. Gönen. Bayesian Supervised Multilabel Learning with Coupled Embedding and Classification. In *Proceedings of the 12th SIAM International Conference on Data Mining*, pages 367–378, 2012.

[68] M. Gönen. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics*, 28(18): 2304–2310, 2012.

[69] M. Gönen. Probabilistic and Discriminative Group-Wise Feature Selection Methods for Credit Risk Analysis. *Expert Systems with Applications*, 39(14): 11709–11717, 2012.

[70] M. Gönen. Bayesian supervised dimensionality reduction. *IEEE Transactions on Cybernetics*, 43(6): 2179–2189, 2013.

[71] M. Gönen. Supervised Multiple Kernel Embedding for Learning Predictive Subspaces. *IEEE Transactions on Knowledge and Data Engineering*, 25(10): 2381–2389, 2013.

[72] M. Gönen, S.A. Khan, and S. Kaski. Kernelized Bayesian Matrix Factorization. In *Proceedings of ICML 2013, the 30th International Conference on Machine Learning*, pages 864–872, 2013.

[73] J. Hakkarainen, A. Ilin, A. Solonen, M. Laine, H. Haario, J. Tamminen, E. Oja, and H. Järvinen. On Closure Parameter Estimation in Chaotic Systems. *Nonlinear Processes in Geophysics*, 19: 127–143, 2012.

[74] R. Hakulinen, S. Puranen, J. V. Lehtonen, M. S. Johnson, and J. Corander. Probabilistic Prediction of Contacts in Protein-Ligand Complexes. *PLoS ONE* 7(11): e49216, 2012.

[75] T. Hao, T. Raiko, A. Ilin, and J. Karhunen. Gated Boltzmann Machine in Texture Modeling. In *Proceedings of 22nd International Conference on Artificial Neural Networks (ICANN 2012)*, in volume 7553 of *Lecture Notes in Computer Science*, pages 124–131. Springer, 2012.

[76] H. Heikinheimo, J.T. Eronen, A. Sennikov, C.D. Preston, E. Oikarinen, P. Uotila, H. Mannila, and M. Fortelius. Convergence in the distribution patterns of Europe's plants and mammals is due to environmental forcing. *Journal of Biogeography*, 39: 1633–1644 2012.

[77] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj. Supervised model training for overlapping sound events based on unsupervised source separation. In *Proceedings, 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 8677–8681, 2013.

[78] L. Hella, M. Järvisalo, A. Kuusisto, J. Laurinharju, T. Lempiäinen, K. Luosto, J. Suomela, and J. Virtema. Weak models of distributed computing, with connections to modal logic. In Darek Kowalski and Alessandro Panconesi, editors, *Proceedings of the 31st Annual ACM Symposium on Principles of Distributed Computing (PODC 2012)*, pages 185–194. ACM, 2012.

[79] M. Heule, M. Järvisalo, and A. Biere. Covered clause elimination. In Andrei Voronkov, Geoff Sutcliffe, Matthias Baaz, and Christian Fermüller, editors, *Short Paper Proceedings of the 17th International Conference on Logic for Programming, Artificial Intelligence and Reasoning (LPAR-17 / 2010)*, volume 13 of *EasyChair Proceedings in Computing*, pages 41–46, 2013.

[80] M. Heule, M. Järvisalo, and A. Biere. Revisiting hyper binary resolution. In Carla Gomes and Meinolf Sellmann, editors, *Proceedings of the 10th International Conference on Integration of Artificial Intelligence and Operations Research Techniques in Constraint Programming (CPAIOR 2013)*, volume 7874 of *Lecture Notes in Computer Science*, pages 77–93, 2013.

[81] A. Honkela, M. Rattray, and N.D. Lawrence. Mining regulatory network connections by ranking transcription factor target genes using time series expression data In *Methods in Molecular Biology*, pages 59–67, 2013.

[82] A. Hyttinen, P. Hoyer, F. Eberhardt, and M. Järvisalo. Discovering cyclic causal models with latent variables: A general SAT-based procedure. In Ann Nicholson and Padhraic Smyth, editors, *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pages 301–310, AUAI Press, 2013.

[83] A. Hyvärinen and N. Manthey. Designing scalable parallel SAT solvers. In A. Cimatti and R. Sebastiani, editors, *SAT 15th International Conference on Theory and Applications of Satisfiability Testing (SAT 2012) SAT*, volume 7317 of *Lecture Notes in Computer Science*, pages 214–227. Springer, 2012.

[84] N. Innocenti and E. Aurell. Lognormality and oscillations in the coverage of high-throughput transcriptomic data towards gene ends. *Journal of Statistical Mechanics: Theory and Experiment*, P10013, 2013.

[85] S. Ishikawa, M. Koskela, M. Sjöberg, J. Laaksonen, E. Oja, E. Amid, K. Palomäki, A. Mesaros, and M. Kurimo. PicSOM experiments in TRECVID 2013. In *Proceedings of the TRECVID 2013 Workshop*, Gaithersburg, MD, USA, November 2013.

[86] T. Janhunen and I. Niemelä. Applying visible strong equivalence in answer-set program transformations. In *Correct Reasoning – Essays on Logic-Based AI in Honour of Vladimir Lifschitz*, volume 7265 of *Lecture Notes in Computer Science*, pages 363–379, Springer, 2012.

[87] T. Janhunen and V. Luukkala. Meta programming with asnwer sets for smart spaces. In Markus Krötzsch and Umberto Straccia, editors, *Proceedings of the 6th International Conference on Web Reasoning and Rule Systems*, pages 106–121, Vienna, September, Springer, 2012.

[88] T. Jantunen, V Viitaniemi, M. Karppa, and J. Laaksonen. The head as a place of articulation: From automated detection to linguistic analysis. In *Proceedings of 11th Theoretical Issues in Sign Language Research conference (TISLR 2013)*, 2013.

[89] M. Järvisalo, A. Biere and M. Heule. Simulating circuit-level simplifications on CNF. *Journal of Automated Reasoning*, 49(4): 583–619, 2012.

[90] M. Järvisalo, M. Heule, and A. Biere. Inprocessing rules. In Bernhard Gramlich, Dale Miller, and Uli Sattler, editors, *Proceedings of the 6th International Joint Conference on Automated Reasoning (IJCAR 2012)*, volume 7364 of *Lecture Notes in Computer Science*, pages 355–370, Springer, 2012.

[91] M. Järvisalo, D. Le Berre, O. Roussel, and L. Simon. The international SAT solver competitions. *AI Magazine*, 33(1): 89–92, 2012.

[92] M. Järvisalo, P. Kaski, M. Koivisto, and J.H. Korhonen. Finding efficient circuits for ensemble computation. In Alessandro Cimatti and Roberto Sebastiani, editors, *Proceedings of the 15th International Conference on Theory and Applications of Satisfiability Testing (SAT 2012)*, volume 7317 of *Lecture Notes in Computer Science*, pages 369–382, Springer, 2012.

[93] M. Järvisalo, A. Matliah, J. Nordström, and S. Živný. Relating proof complexity measures and practical hardness of SAT. In Michela Milano, editor, *Proceedings of the 18th International Conference on Principles and Practice of Constraint Programming (CP 2012)*, volume 7514 of *Lecture Notes in Computer Science*, pages 316–331, 2012.

[94] V. Jääskinen, V. Parkkinen, L. Cheng, and J. Corander. Bayesian clustering of DNA sequences using Markov chains and a stochastic partition model. *Statistical Applications in Genetics and Molecular Biology*, 13(1): 105–21, 2013.

[95] V. Jääskinen, J. Xiong, T. Koski, and J. Corander. Sparse Markov Chains for Sequence Data. *Scandinavian Journal of Statistics*, doi: 10.1111/sjos.12053, 2013

[96] M. Kandemir and S. Kaski. Learning relevance from natural eye movements in pervasive interfaces. In Louis-Philippe Morency and Dan Bohus, editors, *Proceedings of the International Conference on Multimodal Interaction, ICMI '12*, pages 85–82, New York, NY, ACM, 2012.

[97] M. Kandemir, A. Klami, A. Vetek, and S. Kaski. Unsupervised inference of auditory attention from biosensors. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012), Lecture Notes in Computer Science*, pages 403–418, Heidelberg, Germany, 2012. Springer.

[98] R. Karhila, R.S. Doddipatla, M. Kurimo, and P. Smit. Creating synthetic voices for children by adapting adult average voice using stacked transformations and VTLN. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[99] R. Karhila, U. Remes, and M. Kurimo. HMM-based speech synthesis adaptation using noisy data: analysis and evaluation methods. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6930–6934, 2013.

[100] M. Karppa, T. Jantunen, V. Viitaniemi, J. Laaksonen, B. Burger and D. De Weerdt. Comparing computer vision analysis of signed language video with motion capture recordings. In *Proceedings of 8th Language Resources and Evaluation Conference (LREC 2012)*, pages 2421–2425, Istanbul, Turkey, May 2012. Available online at `http://www.lrec-conf.org/proceedings/lrec2012/pdf/321\_Paper.pdf`.

[101] S. Kayal. Experiments on the LFW database using curvelet transforms and a random forest-kNN cascade. In *Digital Information Processing and Communications (ICDIPC), 2012 Second International Conference on*, pages 146–149, Klaipeda, Lithuania, July 2012.

[102] S. Kayal. Face verification experiments on the LFW database with simple features, metrics and classifiers. In *Proceedings of the 8th International Workshop on Multidimensional Systems (nDS'13)*, pages 205–210, 2013.

[103] S. Kayal. Face Clustering in Videos: GMM-based Hierarchical Clustering using Spatio-temporal Data. In *Proceedings of the 13th UK Workshop on Computational Intelligence (UKCI 2013)*, pages 272–278, 2013.

[104] S. Kayal. Face Clustering Experiments on News Video Images. *Journal of Automation and Control Engineering*, 1(3): 213–216, 2013.

[105] S. Keronen, H. Kallasjoki, U. Remes, G.J. Brown, J.F. Gemmeke, and K.J. Palomäki. Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment. *Computer Speech and Language*, 27(3): 798–819, 2013.

[106] S. Keronen, K. Cho, T. Raiko, A. Ilinr, and K.J. Palomäki. Gaussian-Bernoulli restricted Boltzmann machines and automatic feature extraction for noise robust missing data mask estimation. In *Proceedings of the 38th International Conference on Acoustics, Speech and Signal Processing*, pages 6729–6733, 2013.

[107] S. Keronen, U. Remes, H. Kallasjoki, and K.J. Palomäki. Noise robust missing data mask estimation based on automatically learned features. in *the 2nd CHiME Workshop on Machine Listening in Multisource Environments*, pages 77–78, 2013.

[108] S.A. Khan, A. Faisal, J.P. Mpindi, J.A. Parkkinen, T. Kalliokoski, A. Poso, O.P. Kallioniemi, K. Wennerberg, and S. Kaski. Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs. *BMC Bioinformatics*, 13(112), 2012.

[109] R. Kindermann, T. Junttila, and I. Niemelä. Beyond Lassos: Complete SMT-Based Bounded Model Checking for Timed Automata. In *Formal Techniques for Distributed Systems*, volume 7273 of *Lecture Notes in Computer Science*, pages 84–100. Springer, 2012.

[110] R. Kindermann, T. Junttila, and I. Niemelä. SMT-Based Induction Methods for Timed Systems. In *Formal Techniques for Distributed Systems*, volume 7273 of *Lecture Notes in Computer Science*, pages 171–187. Springer, 2012.

[111] R. Kindermann, T. Junttila, and I. Niemelä. Bounded model checking of an MITL fragment for timed automata. In *Proceedings of the 13th International Conference on Application of Concurrency to System Design, ACSD 2013*, pages 216–225. IEEE, 2013.

[112] A. Klami. Variational Bayesian matching. In Steven C.H. Hoi and Wray Buntine, editors, *Proceedings of Asian Conference on Machine Learning*, volume 25 of *JMLR C&WP*, pages 205–220. JMLR, 2012. Best paper award.

[113] A. Klami. Bayesian object matching. *Machine Learning*, 92: 225–250, 2013.

[114] A. Klami, S. Virtanen, and S. Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14: 965–1003, 2013. Implementation in R available at http://research.ics.aalto.fi/mi/software/CCAGFA/.

[115] J. Klapuri, I.T. Nieminen, T. Raiko, and K. Lagus. Variational Bayesian PCA versus k-NN on a Very Sparse Reddit Voting Dataset. In *Advances in Intelligent Data Analysis XII*, volume 8207 of *Lecture Notes in Computer Science*, pages 249–260. Springer, 2013.

[116] K. Konyushkova and D. Głowacka, Content-Based Image Retrieval with Hierarchical Gaussian Process Bandits with Self-Organizing Maps. In *Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 267–272, 2013.

[117] 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*, 34(6): 1477–1489, 2013.

[118] J.-K. Kämäräinen, and M. Koskela (eds.). *Proceedings of 18th Scandinavian Conference on Image Analysis (SCIA 2013)*, volume 7944 of *Lecture Notes in Computer Science*. Springer, 2013.

[119] T.D. Laajala, J. Corander, N.M. Saarinen, K. Mäkelä, S. Savolainen, M.I. Suominen, E. Alhoniemi, S. Mäkelä, M. Poutanen, and T. Aittokallio. Improved statistical modeling of tumor growth and treatment effect in preclinical animal studies with highly heterogeneous responses in vivo. *Clinical Cancer Research*, 18(16): 4385–4396, 2012.

[120] T. Laitinen, T. Junttila, and I. Niemelä. Classifying and propagating parity constraints. In *Principles and Practice of Constraint Programming, CP 2012*, volume 7514 of *Lecture Notes in Computer Science*, pages 357–372, Springer, 2012.

[121] T. Laitinen, T. Junttila, and I. Niemelä. Conflict-driven XOR-clause learning. In *Theory and Applications of Satisfiability Testing, SAT 2012*, volume 7317 of *Lecture Notes in Computer Science*, pages 383–396, Springer, 2012.

[122] T. Laitinen, T. Junttila, and I. Niemelä. Extending clause learning SAT solvers with complete parity reasoning. In *Proceedings of the 24th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2012*, IEEE Computer Society Press, 2012.

[123] T. Laitinen, T. Junttila, and I. Niemelä. Simulating parity reasoning. In *Proceedings of the 19th International Conference on Logic Programming and Automated Reasoning, LPAR 2013*, volume 8312 of *Lecture Notes in Computer Science*, pages 568–583 Springer, 2013.

[124] F. Lebreton, et al. Emergence of epidemic multi-drug resistant Enterococcus faecium from animal and commensal strains. *mBio*, 4(4): e00534-13, 2013.

[125] G. Leen, J. Peltonen, and S. Kaski. Focused multi-task learning in a Gaussian process framework. *Machine Learning*, 89(1-2): 157–182, 2012.

[126] G. Liu, T. Janhunen, and I. Niemelä. Answer set programming via mixed integer programming. In *Proceedings of the 13th International Conference on Principles of Knowledge Representation and Reasoning, KR 2012*, pages 32–42, AAAI Press, 2012.

[127] G. Liu, T. Janhunen, and I. Niemelä. Introducing real variables and integer objective functions to answer set programming. In *Declarative Programming and Knowledge Management*,volume 1306 of *Lecture Notes in Computer Science*, pages 93–107. Springer, 2013. Revised Selected Papers of INAP'13.

[128] Z. Liu, B. Malone, and C. Yuan. Empirical evaluation of scoring functions for Bayesian network model selection. *BMC Bioinformatics*, 13(Suppl 15): S14, 2012.

[129] Z. Lu, Z. Yang, and E. Oja. Selecting β-divergence for nonnegative matrix factorization by score matching. In *Proceedings of 22nd International Conference on Artificial Neural Networks (ICANN 2012)*, pages 419–426, Lausanne, Switzerland, 2012.

[130] P. Luosto, C.D. Giurcaneanu, and P. Kontkanen. Construction of irregular histograms by penalized maximum likelihood: a comparative study. In *IEEE Information Theory Workshop 2012 (ITW)*, Lausanne, Switzerland, 3-7 September 2012.

[131] J. Luttinen and A. Ilin. Efficient Gaussian Process Inference for Short-Scale Spatio-Temporal Modeling. In *JMLR Workshop and Conference Proceedings (AISTATS 2012)*, volume 22, pages 741–750, 2012.

[132] J. Luttinen, A. Ilin, and J. Karhunen. Bayesian Robust PCA of Incomplete Data. *Neural Processing Letters*, 36(2), 189–202, 2012.

[133] J. Luttinen. Fast Variational Bayesian Linear State-Space Model. In *Machine Learning and Knowledge Discovery in Databases*, volume 8188 of *Lecture Notes in Computer Science*, pages 305–320. Springer, 2013.

[134] M. Luzardo, M. Karppa, J. Laaksonen, and T. Jantunen. Head pose estimation for sign language video. In *Proceedings of the 18th Scandinavian Conference on Image Analysis*, volume 7944 of *Lecture Notes in Computer Science*, Espoo, Finland, Springer Verlag, June 2013.

[135] B. Malone and C. Yuan. A bounded error, anytime parallel algorithm for exact Bayesian network structure learning. In *Proceedings of the 6th European Workshop on Probabilistic Graphical Models*, 2012.

[136] B. Malone and C. Yuan. A depth-first branch and bound algorithm for learning optimal Bayesian networks. In *3rd International Workshop on Graph Structures for Knowledge Representation and Reasoning*, 2013.

[137] B. Malone and C. Yuan. Evaluating anytime algorithms for learning optimal Bayesian networks. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, 2013.

[138] B. Malone and C. Yuan. Learning Optimal BAyesian Networks: A Shortest Path Perspective. *Journal of Artificial Intelligence Research*, 48: 23–65, 2013.

[139] A. Mansikkaniemi and M. Kurimo. Adaptation of Morpheme-based Speech Recognition for Foreign Entity Names. *Proceedings of the Fifth International Conference Human Language Technologies (HLT 2012) - The Baltic Perspective*, pages 129–137. IOS Press, 2012.

[140] A. Mansikkaniemi and M. Kurimo. Unsupervised Vocabulary Adaptation for Morph-Based Language Models In *Proceedings of the NAACL 2012 Workshop on the Future of Language Modeling for HLT*, pages 37–40. ACL, 2012.

[141] A. Mansikkaniemi and M. Kurimo. Unsupervised Topic Adaptation for Morph-based Speech Recognition. In *Proceedings of Interspeech 2013*, pages 2693–2697, IOS Press, 2013.

[142] P. Marttinen, W.P. Hanage, N. Croucher, T.R. Connor, S. Harris, S. Bentley, and J. Corander. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Research*, 40: e6, 2012

[143] P. Marttinen, J. Gillberg, A. Havulinna, J. Corander, and S. Kaski. Genome-wide association studies with high-dimensional phenotypes. *Statistical Applications in Genetics and Molecular Biology*, 12(4): 413–431, 2013.

[144] A. McNally, L. Cheng, S.R. Harris, and J. Corander. The evolutionary path to extra intestinal pathogenic, drug resistant Escherichia coli is marked by drastic reduction in detectable recombination within the core genome. *Genome Biology and Evolution*, 5: 699-710, 2013.

[145] A. Medlar, D. Głowacka, H. Stanescu, K. Bryson, and R. Kleta. SwiftLink: Parallel MCMC linkage analysis utilising multicore CPU and GPU. *Bioinformatics*, 29: 420–427, 2013

[146] A. Mesaros. Singing Voice Identification and Lyrics Transcription for Music Information Retrieval. In *Proceedings, 7th Conference on Speech Technology and Human-Computer Dialogue (SpeD2013)*, 10, 2013.

[147] A. Mesaros, T. Heittola, and K. Palomäki. Analysis of acoustic-semantic relationship for diversely annotated real-world audio data. In *Proceedings, 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 813–817, 2013.

[148] A. Mesaros, T. Heittola, and K. Palomäki. Query-by-example retrieval of sound events using an integrated similarity measure of content and label. In *Proceedings, 14th International Workshop on Image and Audio Analysis for Multimedia Interactive services (WIA2MIS 2013)*, pages 1–4, 2013.

[149] B. Molgaard, T. Hussein, J. Corander, K. Hämeri. Forecasting size-fractionated partical number concentrations in the urban atmosphere. *Atmospheric Environment*, 46: 155–163, 2012.

[150] B. Molgaard, W. Birmili, S. Clifford, A. Massling, K. Eleftheriadis, M. Norman, S. Vratolis, B. Wehner, J. Corander, K. Hämeri, and T. Hussein", Evaluation of a statistical forecast model for size-fractionated urban particle number concentrations using data from five European cities. *Journal of Aerosol Science*, 66: 96–110, 2013.

[151] R. Moroni, L. Aalberg, T. Reinikainen, and J. Corander. Bayesian adaptive approach to estimating sample sizes for seizures of illicit drugs. *Journal of Forensic Sciences*, 57(1): 80–85, 2012.

[152] M. Nguyen, T. Janhunen, and I. Niemelä. Translating answer-set programs into bit-vector logic. In *Applications of Declarative Programming and Knowledge Management*, volume 7773 of *Lecture Notes in Computer Science*, pages 95–113, Springer, 2013. Revised Selected Papers of INAP'11.

[153] E. Numminen, L. Cheng, M. Gyllenberg, and J. Corander. Estimating the transmission dynamics of Streptococcus pneumoniae from strain prevalence data. *Biometrics*, 69(3): 748–757, 2013.

[154] K. Nybo, J. Shawe-Taylor, S. Kaski, and J. Mourao-Miranda. Characterizing unknown events in MEG data with group factor analysis. In *Proceedings of the 3rd Workshop on Machine Learning and Interpretation in Neuroimaging (MLINI)*, 2013.

[155] H. Nyman, T. Talonen, A. Roine, M. Hupa, and J. Corander. Statistical approach to quality control of large thermodynamic databases. *Metallurgical and Materials Transactions B - Process Metallurgy and Materials Processing Science*, 43(5): 1113–1118, 2012.

[156] A. Oulasvirta, A. Pihlajamaa, J.P. Perkiö, D. Ray, T. Vähäkangas, T. Hasu, N. Vainio, and P. Myllymäki. Long-term effect of ubiquitous surveillance in the home. In *Proceedings of the 14th International Conference on Ubiquitous Computing (Ubicomp 2012)*, pages 41–50, ACM, 2012.

[157] A. Oulasvirta, T. Roos, A. Modig, and L. Leppänen. Information capacity of full-body movements. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1289–1298, 2013.

[158] J. Pajarinen and J. Peltonen. Expectation maximization for average reward decentralized POMDPs. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Zelezny, editors, *Proceedings of ECML PKDD 2013, The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 129–144, Berlin Heidelberg, Springer-Verlag, 2013.

[159] J. Pajarinen, A. Hottinen, and J. Peltonen. Optimizing spatial and temporal reuse in wireless networks by decentralized partially observable Markov decision processes. *IEEE Transactions on Mobile Computing*, 13(4): 866–879, 2014. Published online March 2013.

[160] T. Peltola, P. Marttinen, and A. Vehtari. Finite Adaptation and Multistep Moves in the Metropolis-Hastings Algorithm for Variable Selection in Genome-Wide Association Analysis. *PLoS ONE*, 7(11), e49445, 2012.

[161] J. Peltonen and K. Georgatzis. Efficient optimization for data visualization as an information retrieval task. In Ignacio Santamaría, Jerónimo Arenas-García, Gustavo Camps-Valls, Deniz Erdogmus, Fernando Pérez-Cruz, and Jan Larsen, editors,

Proceedings of MLSP 2012, the 2012 IEEE International Workshop on Machine Learning for Signal Processing, page electronic proceedings, Piscataway, NJ. IEEE, 2012.

[162] J. Peltonen and Z. Lin. Multiplicative update for fast optimization of information retrieval based neighbor embedding. In Saeid Sanei, Paris Smaragdis, Asoke Nandi, Anthony TS Ho, and Jan Larsen, editors, *Proceedings of MLSP 2013, the 2013 IEEE International Workshop on Machine Learning for Signal Processing*, page electronic proceedings, Piscataway, NJ. IEEE, 2012.

[163] J. Peltonen, T. Raiko, and S. Kaski. *Neuroinformatics, Special Issue on Machine Learning for Signal Processing 2010*, volume 80. Elsevier, 2012.

[164] J. Peltonen and Z. Lin. Information retrieval perspective to meta-visualization. In *Proceedings of ACML 2013, Fifth Asian Conference on Machine Learning*, JMLR W&CP, volume 29, pages 165–180, 2013. JMLR.

[165] J. Peltonen, M. Sandholm, and S. Kaski Information retrieval perspective to interactive data visualization. In M. Hlawitschka and T. Weinkauf, editors, *Proceedings of Eurovis 2013, The Eurographics Conference on Visualization*, the Eurographics Association, 2013.

[166] C. Prakash, D. Gowda, and S. Gangashetty. Analysis of acoustic events in speech signals using Bessel series expansion. *Circuits, Systems, and Signal Processing*, 32(6): 2915–2938, 2013.

[167] H. Pulakka, U. Remes, S. Yrttiaho, K. Palomäki, M. Kurimo, and P. Alku. Bandwidth extension of telephone speech to low frequencies using sinusoidal synthesis and a Gaussian mixture model. *IEEE Trans. Audio, Speech, and Language Processing*, 20(8): 2219–2231, 2012.

[168] T. Pulkkinen and P. Nurmi. AWESOM: Automatic discrete partitioning of indoor spaces for WiFi fingerprinting. In *Proceedings of the 10th International Conference on Pervasive Computing (Pervasive 2012)*, Lecture Notes in Computer Science, pages 271–288, Springer Berlin Heidelberg, 2012.

[169] J. Pylkkönen and M. Kurimo. Analysis of Extended Baum-Welch and Constrained Optimization for Discriminative Training of HMMs. *IEEE Transcations on Audio, Speech, and Language Processing*, 20(9): 2409–2419, 2012.

[170] J. Pylkkönen and M. Kurimo. Improving Discriminative Training for Robust Acoustic Models in Large Vocabulary Continuous Speech Recognition. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*, 2012.

[171] J. Pylkkönen and M. Kurimo. Optimization-Based Control for the Extended Baum-Welch Algorithm. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*, 2012.

[172] K. Qian, A. Di Lieto, J. Corander, P. Auvinen, and D. Greco. Re-analysis of Bipolar Disorder and Schizophrenia Gene Expression Complements *the Kraepelinian Dichotomy*, volume 736, pages 563–577. Springer, 2012.

[173] T. Raiko, H. Valpola, and Y. LeCun. Deep learning made easier by linear transformations in perceptrons. In *Proceedings of the 15th Int. Conf. on Artificial Intelligence and Statistics (AISTATS 2012)*, JMLR W&CP, volume 22, pages 924–932, 2012.

[174] J. Raitio, T. Raiko, and T. Honkela. Hybrid Bilinear and Trilinear Models for Exploratory Analysis of Three-Way Poisson Counts. In *Proceedings of Artificial Neural Networks and Machine Learning – ICANN 2012*, volume 7553 of *Lecture Notes in Computer Science*, pages 475–482. Springer, 2012.

[175] S. Remes, A. Klami, and S. Kaski. Characterizing unknown events in meg data with group factor analysis. In *Proceedings of the 3rd Workshop on Machine Learning and Interpretation in Neuroimaging (MLINI)*, 2013.

[176] U. Remes. Bounded conditional mean imputation with an approximate posterior. In *Proc. INTERSPEECH*, pages 3007–3011, 2013.

[177] U. Remes, R. Karhila, and M. Kurimo. Objective evaluation measures for speaker-adaptive HMM-TTS systems In *Proc. SSW8*, pages 177–181, 2013.

[178] N. Reyhani, J. Ylipaavalniemi, R. V. Vigário, and E. Oja. Consistency and asymptotic normality of FastICA and bootstrap FastICA. *Signal Processing*, 92(8): 1767–1778, 2012.

[179] J. Rintanen. Scheduling with contingent resources and tasks. In *Proceedings of the International Conference on Automated Planning and Scheduling, ICAPS 2013*, pages 189–196, AAAI Press, 2013.

[180] J. Rintanen and C.O. Gretton. Computing upper bounds on lengths of transition sequences. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI 2013*, pages 2365–2372, AAAI Press, 2013.

[181] J. Rissanen. Optimal Estimation of Parameters. Cambridge University Press, 2012.

[182] S. de Rooij, W. Kotlowski, J. Rissanen, P. Myllymäki, T. Roos, K. Yamanishi, editors. *Proceedings of the Fifth Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-2012)*, 2012.

[183] T. Roos, P. Myllymäki, and T. Jaakkola. Editorial: Special issue on the Fifth European Workshop on Probabilistic Graphical Models (PGM-2010). *International Journal of Approximate Reasoning*, 59, 2012.

[184] T. Roos and Y. Zou. Keep it simple stupid - On the effect of lower-order terms in BIC-like criteria *Proc. 2013 Information Theory and Applications Workshop, (ITA-2013)*, an invited paper, 2013.

[185] T. Ruokolainen. Applying Piecewise Approximation in Perceptron Training of Conditional Random Fields In *Advances in Intelligent Data Analysis XI - 11th International Symposium, IDA 2012*, volume 7619 of *Springer Lecture Notes in Computer Science*, pages 324–333. Springer, 2012.

[186] T. Ruotsalo, K. Athukorala, D. Głowacka, K. Konyushkova, A. Oulasvirta, S. Kaipainen, S Kaski, and G. Jacucci. Supporting exploratory search tasks with interactive user modeling. In *76th Annual Meeting of the Association for Information Science and Technology (ASIST)*, Silver Spring, MD, Association for Information Science and Technology, 2013.

[187] T. Ruotsalo, J. Peltonen, M. Eugster, D. Głowacka, K. Konyushkova, K. Athukorala, I. Kosunen, A. Reijonen, P. Myllymaki, G. Jacucci, and S. Kaski. Directing exploratory search with interactive intent modeling. In *22nd ACM International Conference on Information & Knowledge Management (CIKM)*, pages 1759–1764, New York, NY, ACM, 2013.

[188] H. Schulz, K. Cho, T. Raiko, and S. Behnke. Two-Layer Contractive Encodings with Linear Transformation of Perceptrons for Semi-Supervised Learning. In *Proceedings of the International Conference on Neural Information Processing (ICONIP 2013)*, 8226, pages 450–457, 2013.

[189] M.A. Shepheard, V.M. Fleming, T.R. Connor, J. Corander, E.J. Feil, C. Frase, and W.P. Hanage. Historical Zoonoses and Other Changes in Host Tropism of Staphylococcus aureus, Identified by Phylogenetic Analysis of a Population Dataset. *PLoS One*, 8(5), 2013

[190] M. Shubin. Analysing the course of the A(H1N1) influenza epidemic of 2009-2010 in Finland. In *Proceedings of Bioinformatics Research and Education Workshop*, 2012.

[191] M. Shubin and S.L. Varvio. Analyzing the course of the A(H1N1) influenza epidemic of 2009 in Finland. In *Proceedings of the Theory and Applications of System Analysis (Teorija i Praktika Sistemnogo Analiza)*, pages 62–68, 2012.

[192] M. Shubin. Decomposing the bacterial phenotypic time-series into biologically-meaningful components. In *Bioinformatics Research and Education Workshop*, 2013.

[193] L.M. Sihvonen, K. Jalkanen, E. Huovinen, S. Toivonen, J. Corander, M. Kuusi, M. Kurnik, A. Siitonen, and K. Haukka. Clinical isolated of Yersinia enterocolitica Biotype 1A represent two phylogenetic lineages with differing pathogenicity-related properties. *BMC Microbiology*, 12(208), 2012.

[194] J. Sirén, W.P. Hanage, and J. Corander. Inference on Population Histories by Approximating Infinite Alleles Diffusion. *Molecular Biology and Evolution*, doi: 10.1093/molbev/mss227, 2012.

[195] M. Sjöberg, M. Koskela, S. Ishikawa, and J. Laaksonen. Real-time large-scale visual concept detection with linear classifiers. In *Proceedings of 21st International Conference on Pattern Recognition*, Tsukuba, Japan, November 2012.

[196] M. Sjöberg, S. Ishikawa, M. Koskela, J. Laaksonen, and E. Oja. PicSOM experiments in TRECVID 2012. In *Proceedings of the TRECVID 2012 Workshop*, Gaithersburg, MD, USA, November 2012.

[197] M. Sjöberg, M. Koskela, S. Ishikawa, and J. Laaksonen. Large-scale visual concept detection with explicit kernel maps, and power mean SVM. In *Proceedings of ACM International Conference on Multimedia Retrieval (ICMR2013)*, pages 239–246, Dallas, Texas, USA, ACM, April 2013.

[198] M. Sjöberg, J. Schlüter, B. Ionescu, and M. Schedl. FAR at MediaEval 2013 Violent Scenes Detection: Concept-based Violent Scenes Detection in Movies. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, 2013.

[199] M.J. Skwark, A. Abdel-Rehim, A. Elofsson. PconsC: Combination of direct information methods and alignments improves contact prediction. *Bioinformatics* 15;29(14): 1815-6, 2013

[200] S. Srivastava, B. Malone, N. Sukhija, I. Banicescu, and F. M. Ciorba. Predicting the flexibility of dynamic loop scheduling using an artificial neural network. In *Proceedings of the 12th International Symposium on Parallel and Distributed Computing*, 2013.

[201] A. Suni, R. Karhila, T. Raitio, M. Kurimo, M. Vainio, and P. Alku. Lombard Modified Text-to-Speech Synthesis for Improved Intelligibility: Submission for the Hurricane Challenge 2013. *Proc. INTERSPEECH*, 2013.

[202] M. Sunnåker, A.G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. Approximate Bayesian Computation. *PLoS Computational Biology* 9(1): e1002803. doi:10.1371/journal.pcbi.1002803, 2013.

[203] T. Suvitaival, J.A. Parkkinen, S. Virtanen, and S. Kaski. Cross-organism prediction of drug hepatotoxicity by sparse group factor analysis. In *12th Annual International Conference on Critical Assessment of Massive Data Analysis (CAMDA)*, 2013. Extended abstract.

[204] S. Tarkoma, J.K. Kämäräinen, and T. Pahikkala. Evaluation Methods for Unsupervised Natural Language Learning. In *Proceedings of Federated Computer Science Event 2012*, pages 66-67, 2012.

[205] M.K. Titsias, A. Honkela, N. D. Lawrence, and M. Rattray. Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison. *BMC Syst Biol*, 6:53, 2012.

[206] J.M. Toivanen, M. Järvisalo, and H. Toivonen. Harnessing constraint programming for poetry composition. In Mary Lou Maher, Tony Veale, Rob Saunders, and Oliver Bown, editors, *Proceedings of the 4th International Conference on Computational Creativity (ICCC 2013)*, pages 160–167. The University of Sydney, 2013.

[207] V. Turunen, M. Kurimo, and S. Keronen. Results for variable speaker and recording conditions on spoken IR in Finnish. In *Proceedings of the 15th Internation Conference on Speech and Computer*, volume 8113 of *Lecture Notes in Computer Science*, pages 271–277. Springer, 2013.

[208] T. Vatanen, T. Raiko, H. Valpola, and Y. LeCun. Pushing Stochastic Gradient towards Second-Order Methods - Backpropagation Learning with Transformations in Nonlinearities. In *Neural Information Processing (ICONIP 2013)*, volume 8226 of *Lecture Notes in Computer Science*, pages 442–449. Springer, 2013.

[209] M. Vehkaperä, Y. Kabashima, S. Chatterjee, E. Aurell, M. Skoglund, L. Rasmussen. Analysis of Sparse Representations Using Bi-Orthogonal Dictionaries. In *Proceedings of the 2012 IEEE Information Theory Workshop*, pages 647-651, Lausanne, Switzerland, September 2012.

[210] V. Viitaniemi, M. Karppa, J. Laaksonen, and T. Jantunen. Detecting hand-head occlusions in sign language video. In *Proceedings of the 18th Scandinavian Conference on Image Analysis*, volume 7944 of *Lecture Notes in Computer Science*, Espoo, Finland, Springer Verlag, June 2013.

[211] S. Virpioja, P. Smit, S.-A. Grönroos, and M. Kurimo. Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline. *Aalto University publication series SCIENCE + TECHNOLOGY*, 25, pp. 38, 2013.

[212] S. Virtanen, Y. Jia, A. Klami, and T. Darrell. Factorized multi-modal topic model. In Nando de Freitas and Kevin Murphy, editors, *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 843–851, Corvallis, Oregon, AUAI Press, 2012.

[213] S. Virtanen, A. Klami, S.A. Khan, and S. Kaski. Bayesian group factor analysis. In Neil Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *JMLR W&CP*, pages 1269–1277. JMLR, 2012. Implementation in R available at http://research.ics.aalto.fi/mi/software/CCAGFA/.

[214] K. Watanabe, T. Roos, and P. Myllymäki Achievability of asymptotic minimax regret in online and batch prediction In *Proc. 5th Asian Conference on Machine Learning (ACML-2013)*, 2013.

[215] K. Watanabe, T. Roos, and P. Myllymäki. Non-achievability of asymptotic minimax regret without knowledge of the sample size. In *Proceedings of Information-Based Induction Sciences and Machine Learning (IBISML)*, 2013.

[216] R.J.L. Willems et al. Restricted gene flow among hospital subpopulations of Enterococcus faecium. *mBio*, 3, e00151-12, 2012.

[217] L.L. Wu, H.-J. Zhou, M. Alava, E. Aurell, and P. Orponen. Witness of unsatisfiability for a random 3-satisfiability formula. *Phys. Rev. E*, 87: 052807, 2013.

[218] Z. Yang and E. Oja. Clustering by low-rank doubly stochastic matrix decomposition. In *Proceedings of 29th International Conference on Machine Learning (ICML2012)*, pages 831–838, Edinburgh, United Kingdom, 2012.

[219] Z. Yang and E. Oja. Quadratic nonnegative matrix factorization. *Pattern Recognition*, 45(4): 1500–1510, 2012.

[220] Z. Yang, He Zhang, and E. Oja. Online projective nonnegative matrix factorization for large datasets. In *Proceedings of 19th International Conference on Neural Information Processing (ICONIP 2012)*, pages 285–290, Doha, Qatar, 2012.

[221] Z. Yang, T. Hao, O. Dikmen, X. Chen, and E. Oja. Clustering by nonnegative matrix factorization using graph random walk. In *Advances in Neural Information Processing Systems 25 (NIPS2012)*, pages 1088–1096, Lake Tahoe, USA, 2012.

[222] Z. Yang, J. Peltonen, and S. Kaski. Scalable optimization of neighbor embedding for visualization. In *Proceedings of ICML 2013, the 30th International Conference on Machine Learning*, volume 28 of *JMLR W&CP*, pages 127–135. JMLR, 2013.

[223] B. Yegnanarayana and D. Gowda. Spectro-temporal analysis of speech signals using zero-time windowing and group delay function. *Speech Communication*, 55(6): 782–795, 2013.

[224] C. Yuan and B. Malone. An improved admissible heuristic for finding optimal Bayesian networks. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, pages 2186–2191, 2012.

[225] C. Yuan and B. Malone. Learning optimal Bayesian networks: A shortest path perspective. *Journal of Artificial Intelligence Research*, 48: 23–65, 2013.

[226] H.-L. Zeng, M. Alava, E. Aurell, J. Hertz, and Y. Roudi. Maximum likelihood reconstruction for Ising Models with Asynchronous Updates. *Phys. Rev. Lett.*, 110: 210601, 2013.

[227] H. Zhang, Z. Yang, and E. Oja. Adaptive multiplicative updates for projective nonnegative matrix factorization. In *Proceedings of 19th International Conference on Neural Information Processing (ICONIP 2012)*, pages 277–284, Doha, Qatar, 2012.

[228] H. Zhang, T. Hao, Z. Yang, and E. Oja. Pairwise clustering with t-PLSI. In *Proceedings of 22nd International Conference on Artificial Neural Networks (ICANN 2012)*, pages 411–418, Lausanne, Switzerland, 2012.

[229] H. Zhang, M. Gönen, Z. Yang, and E. Oja. Predicting Emotional States of Images Using Bayesian Multiple Kernel Learning. In *Proceedings of 20th International Conference on Neural Information Processing (ICONIP 2013)*, Daegu, South Korea, Springer, 2013.

[230] H. Zhang, Z. Yang, M. Gönen, M. Koskela, J. Laaksonen, T. Honkela, and E. Oja. Affective abstract image classification and retrieval using multiple kernel learning. In *Proceedings of 20th International Conference on Neural Information Processing (ICONIP 2013)*, Daegu, South Korea, Springer, 2013.

[231] J.X. Zhou, M.D.S. Aliyu, E. Aurell, and S. Huang. Quasi-potential landscape in complex multi-stable systems. *J. R. Soc. Interface*, rsif20120434, 2012.

[232] Z. Zhu, Z. Yang, and E. Oja. Multiplicative updates for learning with stochastic matrices. In *Proceedings of the 18th conference Scandinavian Conferences on Image Analysis (SCIA 2013)*, pages 143–152, Espoo, Finland, 2013.