BIENNIAL REPORT

2014 – 2015

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

Coordinated by the Department of Computer Science

Aalto University School of Science

P.O. Box 15400

FI–00076 Aalto, Finland

S. Kaski & M. Lindqvist, editors

Otaniemi, October 2016

Unigrafia Oy
Helsinki 2016

Contents

Preface **5**

Personnel **7**

Awards and activities **11**

Doctoral dissertations **27**

Theses **55**

Introduction **63**

1 C1: Learning models from massive data **65**
1.1 Introduction . . . 66
1.2 Speeding up Unsupervised Learning . . . 66
1.3 Feature Selection for Supervised Learning . . . 67
1.4 Dimension Reduction and Visualization . . . 67
1.5 Applications . . . 68

2 C2: Learning from multiple data sources **71**
2.1 Introduction . . . 72
2.2 Group factor analysis . . . 72
2.3 Kernelized Bayesian matrix factorization . . . 73
2.4 Retrieval of experiments . . . 73
2.5 Learning from multimodal media data . . . 75

3 C3: Statistical Inference in Structured Stochastic Models **77**
3.1 Introduction . . . 78
3.2 Probabilistic graphical models . . . 78
3.3 Adaptive Monte Carlo and adaptive MCMC . . . 79
3.4 Inference for intractable models . . . 79

4 C4: Extreme Inference **83**
4.1 Introduction . . . 84
4.2 Contributions to ASP Methodology . . . 84
4.3 Contributions to SAT Methodology . . . 85
4.4 Probabilistic Graphical Models . . . 86
4.5 Further Applications . . . 87
4.5.1 Abstract Argumentation . . . 87
4.5.2 Temporal Planning . . . 88

5 F1: Intelligent Information Access **95**
5.1 Introduction . . . 96
5.2 Contextual information interfaces . . . 96
5.3 Interactive intent modelling and SciNet . . . 96
5.4 Biosignal feedback and brain activity . . . 101
5.5 Visual recognition of human actions . . . 102
5.6 Speech recognition . . . 105
5.7 Video content analysis for intelligent access . . . 109
5.8 Deep neural network research . . . 110

6 F2: Computational Molecular Biology and Medicine **113**
6.1 Introduction . . . 114
6.2 Protein structure prediction by direct coupling analysis . . . 114
6.3 Global epistatic analysis . . . 114
6.4 Computational inference for microbiology and infectious disease epidemiology 115
6.5 Timing of gene expression . . . 116
6.6 Probabilistic models of multiple data sources . . . 117
6.7 Detection and prediction of multivariate associations . . . 120

7 Publications of COIN 2014-2015 **123**

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. This Biennial Report covers the activities of COIN during the years 2014 and 2015.

COIN operates within three departments of two universities: Department of Computer Science (CS), of the School of Science of Aalto University, and the departments of Computer Science (CS) and Mathematics and Statistics (MS) of the University of Helsinki. The majority of COIN's researchers have a double affiliation with the Helsinki Institute for Information Technology HIIT, which is a joint department of the two universities. Academy Professor Samuel Kaski is the director of COIN since 2015 and Professor Jukka Corander the vice director; until 2014 COIN was directed by Aalto Distinguished Professor Erkki Oja and Samuel Kaski was the vice director. COIN consists of six research groups, two led by Kaski and Corander, and the others by Professors Ilkka Niemelä, Erik Aurell (jointly between Aalto and the Royal Institute of Technology in Stockholm, Sweden), Petri Myllymäki, and senior researcher Jorma Laaksonen.

By end 2015, COIN has published 455 peer-reviewed papers. Highlights of 2014-2015 include: Organization of one of the main machine learning conferences AISTATS, chaired by Kaski and Corander, in Iceland, in 2014; 25 COIN doctoral students graduated; Kaski was selected as an Academy Professor for 2016-2020; Corander received the Cozzarelli Prize for excellent and original scientific publication in PNAS, 2014; and Myllymäki was selected as the director of HIIT, 2015-2019.

During 2014-2015 the Scientific Advisory Board of COIN, consisting of Professors Adnan Darwiche, Dan Geiger and Roderick Murray-Smith, met on May 8, 2014 and September 2, 2015.

Samuel Kaski	*Jukka Corander*
Academy Professor	Professor
Director, COIN	Vice-Director, COIN

Personnel

Group of professor Samuel Kaski, Director of COIN, HIIT/Aalto-CS & UH-CS

Kaski, Samuel; D.Sc. (Tech.) Director of HIIT; professor
Mamitsuka, Hiroshi; D.Sc. (Tech.) FiDiPro professor
Peltonen, Jaakko; D.Sc. (Tech.) Academy research fellow, Professor
Honkela, Antti; D.Sc. (Tech.) Academy research fellow
Marttinen, Pekka; Ph.D. Academy research fellow
Bhadra, Sahely; Ph.D. Postdoctoral researcher (joint position with Prof. Rousu)
Blomstedt, Paul; Ph.D. Postdoctoral researcher
Chen, Yi; Ph.D. Postdoctoral researcher
Cheng, Lu; Ph.D. Postdoctoral researcher
Dutta, Ritabrata; Ph.D. Postdoctoral researcher
Eugster, Manuel; Ph.D. Postdoctoral researcher
Gisbrecht, Andrej; Ph.D. Postdoctoral researcher
Micallef, Luana; D.Sc. (Tech.) Postdoctoral researcher (joint position with Prof. Jacucci)
Mononen, Tommi; Ph.D. Postdoctoral researcher (joint position with Prof. Salmelin)
Parviainen, Pekka; Ph.D. Postdoctoral researcher
Peltola, Tomi; D.Sc. (Tech.) Postdoctoral researcher
Afrabandpey, Homayun; M.Sc. Doctoral student
Ammad Ud Din, Muhammad; M.Sc. Doctoral student
Daee, Pedram; M.Sc. Doctoral student
Faisal, Ali; M.Sc. Doctoral student
Gillberg, Jussi; M.Sc. (Tech.) Doctoral student
Kangasrääsiö, Antti; M.Sc. (Tech.) Doctoral student
Khan, Suleiman; M.Sc. Doctoral student
Leppäaho, Eemeli; M.Sc. (Tech.) Doctoral student
Lin, Ziyuan; M.Sc. Doctoral student
Parkkinen, Juuso; M.Sc. (Tech.) Doctoral student
Remes, Sami; M.Sc. Doctoral student
Sundin, Iiris; M.Sc. (Tech.) Doctoral student
Suvitaival, Tommi; M.Sc. (Tech.) Doctoral student
Topa, Hande; M.Sc. Doctoral student
Virtanen, Seppo; M.Sc. (Tech.) Doctoral student
Zhao, Xuran; M.Sc. (Tech.) Doctoral student

Group of professor Erik Aurell, Aalto-CS

Aurell, Erik; D.Sc. (Tech.) FiDiPro professor
Lemoy, Remí; Ph.D. Posdoctoral researcher
Skwark, Marcin; Ph.D. Posdoctoral researcher
Del Ferraro, Gino; M.Sc. Doctoral student
Innocenti, Nicolas; M.Sc. Doctoral student
Marino, Raffaele; M.Sc. Doctoral student
Zheng, HongLi; M.Sc. Doctoral student

Group of professor Jukka Corander, UH-MS

Corander, Jukka; Ph.D. Professor
Gutman, Michael; Ph.D. Postdoctoral researcher
Martino, Luca; Ph.D. Postdoctoral researcher
Wei, Lu; Ph.D. Postdoctoral researcher
Yang, Zhirong; D.Sc. (Tech.) 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
Leppä-aho, Janne; M.Sc. Doctoral student
Vehkala, Minna; M.Sc. Doctoral student
Numminen, Elina; M.Sc. Doctoral student
Pessia, Alberto; M.Sc. Doctoral student
Pusa, Taneli; M.Sc. Doctoral student
Shubin, Mikhail; M.Sc. Doctoral student
Xiong, Jie; M.Sc. Doctoral student

Group of Dr. Jorma Laaksonen, Aalto-CS

Laaksonen, Jorma; D.Sc. (Tech.) Teaching researcher
Oja, Erkki; D.Sc. (Tech.) Aalto Distinguished Professor
Kurimo, Mikko; D.Sc. (Tech.) Associate Professor
Raiko, Tapani; D.Sc. (Tech.) Assistant Professor
Koskela, Markus; D.Sc. (Tech.) Senior researcher
Palomäki, Kalle; D.Sc. (Tech.) Academy research fellow
Gowda, Dhananjaya; Ph.D. Postdoctoral researcher
Dikmen, Onur; Ph.D. Postdoctoral researcher
Gonzalez-Caro, Cristina; Ph.D. Postdoctoral researcher
Mesaros, Annamaria; Ph.D. Postdoctoral researcher
Kivinen, Jyri; Ph.D. Postdoctoral researcher
R.-Tavakoli, Hamed; Ph.D. Postdoctoral researcher
Rao, Anwer Muhammad; Ph.D. Postdoctoral researcher

Yang, Zhirong; D.Sc. (Tech.) Postdoctoral researcher
Abbas, Mudassar; M.Sc. (Tech.) Doctoral student
Berglund, Mathias; M.Sc. Doctoral student
Chen, Xi; M.Sc. Doctoral student
Cho, KyungHyung; M.Sc. Doctoral student
Enarvi, Seppo; M.Sc. (Tech.) Doctoral student
Grönroos, Stig-Arne; 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
Lu, Yao; M.Sc. Doctoral student
Luttinen, Jaakko; M.Sc. (Tech.) Doctoral student
Mansikkaniemi, Andre; M.Sc. (Tech.) Doctoral student
Nieminen, Ilari; M.Sc. (Tech.) Doctoral student
Rasmus, Antti; M.Sc. 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
Takala, Pyry; M.Sc. Doctoral student
Varjokallio, Matti; M.Sc. (Tech.) Doctoral student

Group of Professor Petri Myllymäki, UH-CS

Myllymäki, Petri; Ph.D. Professor
Roos, Teemu; Ph.D. Assistant professor
Floréen, Patrik; Ph.D. Senior researcher
Koskela, Markus; D.Sc. (Tech.) Senior researcher
Rissanen, Jorma; Ph.D. Senior researcher
Järvisalo, Matti; D.Sc. (Tech.) Academy Research Fellow
Klami, Arto; D.Sc. (Tech.) Academy Research Fellow
Dikmen, Onur; Ph.D. Postdoctoral researcher
Głowacka, Dorota; Ph.D. Postdoctoral researcher
Hyttinen, Antti; Ph.D. Postdoctoral researcher
Kontkanen, Petri; Ph.D. Postdoctoral researcher
Malone, Brandon; Ph.D. Postdoctoral researcher
Sjöberg, Mats; Ph.D. Postdoctoral researcher
Tasoulis, Sotiris; Ph.D. Postdoctoral researcher
Wallner, Johannes; D.Sc. (Tech.) Postdoctoral researcher
Jitta, Aditya; M.Sc. Doctoral student

Berg, Jeremias; M.Sc. Doctoral student
Hyvönen, Ville; M.Sc. Doctoral student
Leppä-aho, Janne; M.Sc. Doctoral student
Määttä, Jussi; M.Sc. Doctoral student
Perkiö, Jukka; M.Sc. Doctoral student
Pulkkinen, Teemu; M.Sc. Doctoral student
Pyykkö, Joel; M.Sc. Doctoral student
Saikko, Paul; M.Sc. Doctoral student
Sakaya, Joseph; M.Sc. Doctoral student
Wettig, Hannes; M.Sc. Doctoral student
Zou, Yuan; M.Sc. Doctoral student

Group of Professor Ilkka Niemelä, Aalto-CS

Niemelä, Ilkka; D.Sc. (Tech.) Vice-president, professor
Janhunen, Tomi; D.Sc. (Tech.) Senior university lecturer
Junttila, Tommi; D.Sc. (Tech.) University lecturer
Rintanen, Jussi; D.Sc. (Tech.) Senior researcher
Bogaerts, Bart; D.Sc. (Tech.) Postdoctoral researcher
Gebser, Martin; Ph.D. Postdoctoral researcher
Oikarinen, Emilia; D.Sc. (Tech.) Postdoctoral researcher
Tasharrofi, Shahab; D.Sc. (Tech.) Postdoctoral researcher
Bomanson, Jori; M.Sc. Doctoral student
Kindermann, Roland; M.Sc. Doctoral student
Laitinen, Tero; M.Sc. (Tech.) Doctoral student

Support staff, Aalto University

Lindqvist, Maria Research Coordinator, Aalto
Ehrstedt, Stefan HR coordinator, Aalto
Käpylä, Maarit Academic Coordinator, Aalto
Kauppila, Minna Secretary, Aalto
Koivisto, Leila Controller, 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
Moen, Pirjo Research Coordinator, UH

Awards and activities

Prizes and academic awards received by personnel of the unit

Professor Samuel Kaski

- Academy Professor 2016-2020

Professor Jukka Corander

- Cozzarelli Prize for the best publication in PNAS during 2014

Doc. Jorma Laaksonen

- Winner award of Large Scale Movie Description Challenge, awarded by ICCV 2015 Workshop on Describing and Understanding Video & The Large Scale Movie Description Challenge, Chile, 2015

Professor Ilkka Niemelä

- ECCAI Fellow, awarded by European Coordinating Committee for Artificial Intelligence, Belgia

Doc. Matti Järvisalo

- Honorary Mention at ICCMA 2015: 1st International Competition on Computation Models of Argumentation, 2015

Doc. Arto Klami

- Best paper award at the 15th Koli Calling conference on computing education research, 2015
- Senior good researcher award of Department of Computer Science, 2014

Professor Teemu Roos

- Senior Good Teacher award of the Department of Computer Science, University of Helsinki, 2014

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

Professor Erik Aurell:

- Co-Chair of NORDITA program (Sweden), 2015
- Co-Chair of KITPC program (China), 2014
- Member of AISTATS Program Committee, 2014
- invited speaker:
  High-Dimensional Data-Driven Science, (HD3 -2015), Kyoto, Japan, 2015
  Models of Life, Krogerup, Denmark, 2015
  Regulation and Inference in Biological Networks Workshop, Bardonecchia, Italy, 2015
  Strolling on Chaos, Turbulence and Statistical Mechanics, Rome, Italy, 2014
  Physics of Evolution, Regulation and Signaling, LMU Munich, Germany, 2014

Professor Jukka Corander:

- Co-chair of the program committee: 17th Artificial Intelligence and Statistics (AISTATS) conference, Iceland, 2014.
- Organizer of symposium on Molecular Evolution of Microbial Genomes as part of the SMBE annual conference, Puerto Rico (USA), 2014.
- Chair for the annual Permafrost Workshop on modelling bacterial and viral evolution, Italy, 2014-2015.
- Associate editor, Scandinavian Journal of Statistics, 2009–
- National Editor and board member, Scandinavian Journal of Statistics, 2011–
- Senior Editor, Microbial Genomics (Society for General Microbiology), 2015.
- Member of grant review panels for Norwegian Research Council, Norway, 2014–
- Reviewer, ERC StG application
- Opponent at the doctoral defense of Öystein Sörensen, University of Oslo, 2015.
- Opponent at the doctoral defense of Petter Arnesen, Norwegian University of Science and Technology, 2015.

Doc. Tomi Janhunen:

- Program Committee Member:
  The 13th International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR 2015), USA, 2015.
  The 31st International Conference on Logic Programming (ICLP 2015), Ireland, 2015.
  The first global conference on artificial intelligence (GCAI 2015), Georgia, 2015.
  The 8th International Workshop on Answer Set Programming and Other Computing Paradigms (ASPOCP 2015), Ireland, 2015.

The 24th International Joint Conference on Artificial Intelligence (IJCAI 2015), Argentina, 2015.
The 14th International Conference on Principles of Knowledge Representation and Reasoning, Vienna, Austria, 2014.
The 14th European Conference on Logics in Artificial Intelligence, Madeira, Portugal, 2014.
The 14th European Conference on Artificial Intelligence, Prague, Czech Republic, 2014.
The 7th International Workshop on Answer Set Programming and Other Computing Paradigms, Vienna, Austria, 2014.
The 4th International Conference on Logic and Search, Vienna, Austria, 2014.

- Session Chairman:
  The 3rd Workshop on Grounding, Transforming, and Modularizing Theories with Variables (GTTV 2015), USA, 2015.
  The 13th International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR 2015), USA, 2015.
  The 14th International Conference on Principles of Knowledge Representation and Reasoning, Vienna, Austria, 2014.
  The 14th European Conference on Logics in Artificial Intelligence, Madeira, Portugal, 2014.
  The 14th European Conference on Artificial Intelligence, Prague, Czech Republic, 2014.
  The 7th International Workshop on Answer Set Programming and Other Computing Paradigms, Vienna, Austria, 2014.

- Reviewer:
  Artificial Intelligence Journal
  Theory and Practice of Logic Programming
  Journal of Logic and Computation

- Opponent at University of Potsdam, Benjamin Kaufmann, Germany, 2015.

- Pre-examiner of a doctoral thesis:
  University of New South Wales, Drescher, Christian, Australia, 2014.
  University of Potsdam, Kaufmann, Benjamin, Germany, 2014.

Professor Samuel Kaski:

- Program Committee Member:
  IJCAI 2015, International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015
  ICML 2015, Lille, France, 2015
  Machine Learning Summer School MLSS, Reykjavik, 2014
  ICMI 2014, The 16th International Conference on Multimodal Interaction, 2014, Istanbul

- Chairman of programme committee:
  AISTATS, Reykjavik, 2014
  Mathematica and Statistical Aspects of Molecular Biology, 25th Annual MASAMB Workshop, Finland

- Member of Scientific Advisory Board:
  Novel Materials Discovery (NOMAD) Laboratory, a European Center of Excellence
  Center of Excellence BigInsight, Oslo, Norway
  Thematic Research Unit in Systems Biology and Chemical Biology of the GIGA Research center of the Univ. of Liège

- Action editor:
  Journal of Machine Learning Research
  International Journal of Knowledge Discovery in Bioinformatics

- Invited talks:
  Statistical Genomics Workshop at MASAMB 2014, Sheffield, UK
  Information-Based Induction Sciences IBIS2015, Tsukuba, Japan
  4th International Workshop on Cognitive Information Processing (CIP2014), Copenhagen, Denmark
  The 9th Annual IAPR international conference on Pattern Recognition in Bioinformatics (PRIB), Stockholm, 2014
  Big Data Analytics in the Life Sciences Industry, Brussels, Belgium, 2015
  Statistical Genomics Workshop at MASAMB, Sheffield, UK, 2014
  Cognitive Information Processing, Copenhagen, 2014
  Thee 9th IAPR conference on Pattern Recognition in Bioinformatics, Stockholm
  SYMBIOTIC 2014, International Workshop on Symbiotic Interaction, 2014, Helsinki
  MLCB14, NIPS workshop on Machine Learning in Computational Biology, Montreal, 2014

Doc. Arto Klami:

- Program Committee Member:
  AAAI Conference on Artificial Intelligence, 2014
  International Joint Conference on Artificial Intelligence (IJCAI), 2015
  Uncertainty in Artificial Intelligence (UAI), 2015

- Reviewer:
  Journal of Machine Learning Research, 2014
  Machine Learning, 2014
  Journal of Artificial Intelligence Research, 2014
  IEEE Transactions on Neural Networks and Learning Systems, 2014
  International Conference on Machine Learning (ICML), 2015

Professor Mikko Kurimo:

- Program Committee member:
  Eusipco 2015, Nizza, France
  ICASSP 2014, Florence, Italy
  ICASSP 2015, Brisbane, Australia
  EMNLP 2015, Lissabon, Portugal
  Interspeech 2015, Dresden, Germany
  Interspeech 2014, Singapore
  Eusipco 2014, Lissabon, Portugal

- Editorial board member: ACM Transactions of Speech and Language Processing, 2014-2015.

- Opponent at the doctoral dissertation of Kairit Sirts, Tallinn University of Technology, Estonia

Doc. Jorma Laaksonen:

- Editorial Board Member, Pattern Recognition Letters, the Netherlands.
- Editor of scientific book: Advances in Independent Component Analysis and Learning Machines, 1st Edition, Nertherlands, 2015.
- Opponent at the doctoral defence of Heydar Maboudi Afkham, Kungliska tekniska högskolan, Sweden.

Professor Petri Myllymäki:

- NSF Strategic partnership coordinator, NSF Science and Technology Center for Science of Information , 2014-2015
- Program Committee Co-Chair:
  The Seventh Workshop on Information Theoretic Methods in Science and Engineering, Honolulu, USA, 2014
  The Eight Workshop on Information Theoretic Methods in Science and Engineering, 2015
- Senior Program Committee member:
  The 30th Conference on Uncertainty in Artificial Intelligence (UAI-2014)
  The 31st Conference on Uncertainty in Artificial Intelligence (UAI-2015)
  International Joint Conference on Artificial Intelligence (IJCAI-2015)
- Program Committee Member:
  The Seventh European Workshop on Probabilistic Graphical Models (PGM-2014)
  The Third International Symposium on Learning and Data Sciences (SLDS2015)
  Advisor for the Second Workshop on Advanced Methodologies for Bayesian Networks, 2015
- Reviewer for International Journal of Approximate Reasoning, 2015

Professor Ilkka Niemelä:

- Chairman of the session: 30th International Conference on Logic Programming, Wien, 2014
- Position of trust in Deutsche Forschungsgemeinschaft (DFG)
- Editor of scientific journal: Theory and Practice of Logic Programming, UK

Professor Erkki Oja:

- International Neural Network Society, INNS, College of Fellows -steering committee member, USA.
- Euroopan Commission, ERC Starting Grant - evaluation panel member, field of computer science, Belgium.
- Editorial Board Member:
  Natural Computing - An International Journal, the Netherlands
  Neural Computation, USA.
  International Journal of Pattern Recognition and Artificial Intelligence, Singapore.

Dr. Martin Gebser:

- Program Committee Member:
  ICLP Doctoral Consortium, Vienna, 2014.

- Chairman of the session:
  International Conference on Logic Programming, Vienna, 2014.
  Workshop on Answer Set Programming and Other Computing Paradigms, Vienna, 2014.

- Membership of editorial board of scientific journals: Association for Logic Programming Newsletter, USA, 2014.

- Pre-examiner of a doctoral thesis in University of Calabria, Dodaro, Carmine, Italy, 2014.

Doc. Antti Honkela:

- Program Committee Member:
  Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS 2014)
  Conference on Uncertainty in Artificial Intelligence (UAI 2014)
  11th International Workshop on Computational Systems Biology (WCSB 2014)
  Workshop on Machine Learning in Computational Biology (MLCB 2014)
  NIPS 2014 Workshop on Variational Inference, 2014.
  Conference on Uncertainty in Artificial Intelligence (UAI 2015)
  Workshop on Machine Learning in Computational Biology (MLCB 2015)
  NIPS 2015 Workshop on Advances in Approximate Bayesian Inference, 2015.
  International Conference on Machine Learning (ICML 2015)

- Chair:
  25th Annual MASAMB Workshop on Mathematical and Statistical Aspects of Molecular Biology, Finland, 2015.
  Privacy-aware computational genomics 2015 (PRIVAGEN 2015)
  ICML 2015 Workshop on Machine Learning Open Source Software: Open Ecosystems

- Reviewer:
  Funding proposal for European Science Foundation, 2014.
  Funding proposal for Medical Research Council UK, 2014.
  Neural Information Processing Systems conference (NIPS 2015)
  Journal: Machine Learning, 2014-2015.
  Journal: Nucleic Acids Research, 2014-2015.
  Journal: Nature Communications, 2015.
  Journal: BMC Systems Biology, 2015.

- Associate Editor in Statistical Applications in Genetics and Molecular Biology, 2014–

- Speaker: 23rd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2015)

- Action editor: Journal of Machine Learning Research (Machine Learning Open Source Software section)

Doc. Matti Järvisalo:

- Invited Talk at IDA Machine Learning Seminar, Linköping University, 2015
- Organizer of SAT Competition 2014
- Program Committee Member:
  24th International Joint Conference on Artificial Intelligence (IJCAI 2015) Machine Learning Track
  18th International Conference on Theory and Applications of Satisfiability Testing (SAT 2015)
  22nd RCRA International Workshop on Experimental Evaluation of Algorithms for Solving Problems with Combinatorial Explosion (RCRA 2015)
  21st International Conference on Principles and Practice of Constraint Programming (CP 2015) Doctoral Program
  24th International Joint Conference on Artificial Intelligence (IJCAI 2015) Main Track
  17th International Conference on Theory and Applications of Satisfiability Testing (SAT 2014)
  4th International Workshop on Logic and Search (LaSh 2014)
  29th AAAI Conference on Artificial Intelligence (AAAI 2015)
  3rd Workshop on Combining Constraint Solving with Mining and Learning (CoCoMile 2014)
  20th International Conference on Principles and Practice of Constraint Programming Doctoral Program (CP 2014 DP)
  21st European Conference on Artificial Intelligence (ECAI 2014)
- Reviewer:
  Information Sciences, 2015
  Journal of Automated Reasoning, 2015
  Artificial Intelligence Journal, 2014-2015
  Journal of Experimental and Theoretical Artificial Intelligence, 2014
  13th International Symposium on Artificial Intelligence and Mathematics (ISAIM 2014)
  Annals of Mathematics and Artificial Intelligence, 2014
  16th International Conference on Distributed Computing and Networking (ICDCN 2015)
  Journal of Artificial Intelligence Research, 2015
- Guest Editor for Journal of Satisfiability, Boolean Modeling and Computation, 2014-2016
- Invited Tutorial:
  Dagstuhl seminar Theory and Practice of SAT Solving, Germany 2015
  BIRS workshop Theoretical Foundations of Applied SAT Solving, Canada 2014

Dr. Emilia Oikarinen:

- Program Committee Member:
  4th International Workshop on Logic and Search (LaSh 2014), Vienna, Austria, 2014.
  7th Workshop on Answer Set Programming and Other Computing Paradigms (ASPOCP 2014), Vienna, Austria, 2014.

8th Workshop on Answer Set Programming and Other Computing Paradigms (ASPOCP 2015), Cork, Ireland, 2015.
31st International Conference on Logic Programming (ICLP 2015), Cork, Ireland, 2015.
3rd Workshop on Grounding, Transforming, and Modularizing Theories with Variables (GTTV 2015), Lexington, KY, USA, 2015.
12th International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR 2015), Lexington, KY, USA, 2015.

Doc. Jaakko Peltonen:

- Member of the organising committee:
  ESANN 2015, European Symposium on Artificial Neural Networks, Bruges, Belgium, 2015.
  ICANN 2014, International Conference on Artificial Neural Networks, Hamburg, Germany, 2014.
  WSOM 2014, Workshop on Self-Organizing Maps, Mittweida, Germany, 2014.

- Speaker:
  EEE VIS 2014, IEEE Visual Analytics Science and Technology, IEEE Information Visualization, and IEEE Scientific Visualization, Paris, France, 2014.
  CML 2014, International Conference on Machine Learning, Beijing, China, 2014.
  AAAI 2014, AAAI Conference on Artificial Intelligence, Quebec City, Canada, 2014.
  NIPS 2014, International Conference on Neural Information Processing Systems, Montreal, Canada, 2014.
  Statistics for big data meeting of the International Biometric Society, British and Irish Region (IBS-BIR), London, United Kingdom, 2014.

- Member of editorial board of scientific journals: Neural Processing Letters, Germany, 2014–2015.

- Indo-Dutch Joint Research Programme for ICT, Reviewer of funding application, The Netherlands, 2014.

- Opponent at University of California Merced, Maksym Vladymyrov, United States, 2014.

Professor Tapani Raiko:

- Invited speaker, Visual Forum, Gothenburg, Sweden, 2015.

- Opponent at Örebro University, Martin Längkvist, Sweden, 2015.

- Opponen at Chalmers University of Technology, Olof Mogren, Sweden, 2015.

Dr. Jussi Rintanen:

- Program Committee Member:
  International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015.
  AAAI Conference on Artificial Intelligence, Austin, Texas, USA, 2015.
  AAAI Conference on Artificial Intelligence, Canada, 2014.
  European Conference on Artificial Intelligence, Czech Republic, 2014.
  International Conference on Theory and Applications of Satisfiability Testing, Austria, 2014.

International Conference on Principles of Knowledge Representation and Reasoning, Austria, 2014.
European Conference on Logic in Artificial Intelligence, Portugal, 2014.

Professor Teemu Roos:

- Program Committee Chair:
  Eighth Workshop on Information Theoretic Methods in Science and Engineering, Copenhagen, Denmark, 2015
  Seventh Workshop on Information Theoretic Methods in Science and Engineering, Hawaii, USA, 2014

- Program Committee Member:
  IEEE International Conference on Data Mining (ICDM-2014)
  Conference on Uncertainty in Artificial Intelligence (UAI-2014 and 2015)
  Textual Trails: Transmissions of Oral and Written Texts, 2014
  International Joint Conference on Artificial Intelligence (IJCAI-2015)

- Speaker:
  E-Philology Lecture Series, Paris, France, 2015
  Evolutionary Analysis Beyond the Gene, Chicheley, UK, 2014
  2015 Information Theory and Applications Workshop
  Transmission of Texts in Late Medieval and Early Modern Europe. Tools and Techniques for Dealing with Complicated Textual Traditions, Rome, Italy, 2015
  PhD Training School: Atelier Tekstvariatie, Amsterdam, Netherlands, 2015
  Statistics and Operations Research Center (STOR-i) seminar, Lancaster, UK, 2015
  2nd International Conference on Algorithms for Computational Biology, Mexico City, Mexico, 2015
  USC Ming Hsieh Department of Electrical Engineering Seminar. Los Angeles, USA, 2015

- Reviewer:
  the Royal Statistical Society, 2014
  International Conference on Artificial Intelligence and Statistics (AISTATS-2014 and 2015)
  NIPS-2014 and 2015
  The Seventh European Workshop on Probabilistic Graphical Models (PGM-2014)
  Methods in Ecology and Evolution, 2014
  IEEE Transactions on Knowledge and Data Engineering, 2014
  IEEE Transactions on Information Theory
  IEEE Transactions of Pattern Analysis and Machine Intelligence, 2015
  ACM Transactions on Knowledge Discovery from Data, 2015
  Machine Learning, 2015
  Bayesian Analysis, 2015
  Journal of Machine Learning Research

- Book proposal reviewer, Morgan Kaufmann Publishers, 2014

- Guest Editor of Literary and Linguistic Computing, Special Issue on Studia Stemmatologica

- Evaluation of applications for the Swiss National Science Foundation, 2014

Dr. Zhirong Yang:

- Reviewer:
  The International Conference on Machine Learning (ICML) 2015
  The journal of Data Mining and Knowledge Discovery, 2015
  IEEE Signal Processing Letters, 2015
  IEEE Transactions on Image Processing, 2015
  Entropy, 2015
  Artificial Inteliigence, 2015
  The 29th Annual Conference on Neural Information Processing Systems (NIPS), 2015

Important domestic positions of academic service by personnel of the unit

Professor Jukka Corander:

- President, The Finnish Society of Biostatistics, 2013–2016.
- Vice-director, Department of Mathematics and Statistics, University of Helsinki, Finland, 2010–
- Board member of Faculty of Sciences, University of Helsinki, 2014–2017
- Research Program director, Helsinki Institute of Information Technology (HIIT), 2015–
- Board member of the DOMAST Doctoral Programme in Mathematics and Statistics, 2013–
- Board member of Helsinki Institute of Life Science, 2015–
- Docentship application evaluator:
  Marko Laine, Lappeenranta University of Technology, 2014.
  Antti Solonen, Lappeenranta University of Technology, 2014.
- Evaluator of assistant/associate professor candidates in applied mathematics, Tampere University of Technology, 2014.
- Invited talks:
  SMBE Satellite conference on reticulated microbial evolution, Kiel, Germany, 2014.
  14th Geilo Winter School. Geilo, Norway, 2014.
  26th Sigrid Juselius International Symposium: Emerging Infections, Finland, 2015.
  Mathematical and Computational Evolutionary Biology conference, France, 2015.
  8th International Congress on Industrial and Applied Mathematics, Beijing, 2015.
  Genome Science 2015, Birmingham, 2015.

Professor Samuel Kaski:

- Statistics Finland, Member of the Advisory Board
- Chairman of Board of Directors, CSC - IT Center for Science, 2012-2015
- Director, Helsinki Institute for Information Technology HIIT, 2010-2015

Professor Mikko Kurimo:

- Program Committee member: Fonetiikan päivät 2015, Aalto, Finland

Professor Petri Myllymäki:

- Director, Helsinki Institute for Information Technology HIIT, 2015-2020
- Director of the Helsinki Doctoral Education Network in Information and Communications Technology (HICT)
- Member of the National Committee for Research Data Management
- Expert member of a Big Data working group, Academy of Finland, The Computational Science Programme

- Member of the Steering Group of the Doctoral School in Natural Sciences, University of Helsinki
- Academic Coordinator of the Data to Intelligence (D2I) research programme, Digile, 2011-2016
- Member of the Council of the Faculty of Science, University of Helsinki
- Director of the Helsinki Doctoral Programme in Computer Science and Engineering (Hecse), 2011-2015
- The Finnish Academy of Technology, member 2013–
- Invited speaker: Oivalluksia 6 - Data to Intelligence, 2014

Doc. Antti Honkela:

- Deputy Board Member of Helsinki Institute for Information Technology (HIIT), 2014
- Board Member of Helsinki Institute for Information Technology (HIIT), 2015–

Doc. Jorma Laaksonen:

- Opponent at the doctoral dissertation of Ekaterina Riabchenko, Lappeenranta University of Technology, 2015
- Pre-examiner of a doctoral thesis, Lappeenranta University of Technology, 2014

Doc. Jaakko Peltonen:

- Speaker: Sino-Finnish Summer School on Social Media Data Analysis, Tampere, Finland, 2014
- Pre-examiner of a doctoral thesis in University of Turku, Seppo Pulkkinen, Finland, 2014

Professor Tapani Raiko:

- Invited speaker:
  Sodankuva murroksessa? - Hybridisodankäynti ja disruptiiviset teknologiat
  Slush
  Teknologia'15

Professor Teemu Roos:

- Tenure-track assistant professorship position working group, Department of Computer Science, University of Helsinki, 2014
- Chair of Helsinki Distinguished Lecture Series on Future Information Technology, 2013-2015

Research visits abroad by personnel of the unit

- Erik Aurell: Chinese Academy of Sciences, Beijing, 1 month. Research visit, 2015
- Jukka Corander: Sanger Institute, UK, 0.5 months. Research visit, 2015
- Gino Del Ferraro:
  Chinese Academy of Science, Beijing, 1 month. Research visit, 2015
  Gino Del Ferraro, University of Cuba, 1 month. Research visit, 2015
- Martin Gebser: Several research visits to University of Potsdam, Germany, 2014
- Antti Honkela: University of Manchester, UK. Research visit, 2015
- Tomi Janhunen: Several research visits to University of Potsdam, Germany, 2014-2015
- Matti Järvisalo: Linköping University, Sweden. Research visit, 2015
- Samuel Kaski:
  University of Kyoto, Japan, 3 months
  Several research visits to UCL, UK, 2014-2015
- Antti Kangasrääsiö: University of Kyoto, Japan, 2 months. Research visit, 2015
- Arto Klami: Amazon Development Center Berlin, 3 months. Research visit, 2015
- Mikko Kurimo:
  Radboud University, Netherlands. Research visit, 2015
  KTH Tukholma, Sweden. Research visit, 2015
  Johns Hopkins University, USA. Research visit, 2014
  Universidad Politechnica de Madrid, Spain. Research visit, 2014
  University of Antwerpen, Belgium. Research visit, 2015
  Tallinn University of Technology, Estonia. Research visit, 2015
- Janne Leppä-aho: University of California, Berkeley, 3 months. Research visit, 2015
- Pekka Marttinen:
  Harvard University, USA, 6 months. Research visit, 2014
  Imperial College London, UK, 0.5 months. Research visit, 2014
- Pekka Parviainen: Kunliga Tekniska Höskolan, Sweden. Research visit, 2014
- Jaakko Peltonen:
  Several research visits to Biological Research Center of the Hungarian Academy of Sciences, Szeged, Hungary, 2015
  Schloss Dagstuhl Leibniz-Zentrum fur Informatik, Germany. Research visit, 2015
  University of Pittsburgh, Pittsburgh, USA. Research visit, 2015
- Tapani Raiko:
  Several visits to University of Denmark, 2015
  University of Montreal, Canada, 5 months. Research visit, 2014
- Teemu Roos:
  Finnish Institute in Rome (Villa Lante), 10 days. Research visit, 2014
  University of California Berkeley, USA, 6 weeks. Research visit, 2015
  University of Southern California, USA, 3 weeks. Research visit, 2015
- Lu Wei: Harvard University, USA, 5 months. Research visit, 2015

Research visits by foreign researchers to the unit

Long research visits (more than two weeks)

- Tanel Alumäe, Dr., Tallinn University of Technology, Estonia, 2014
- Brian Arnold, Prof., Harvard University, 2015
- Wray Buntine, Prof., Monash University, Australia, 2015
- Eduardo Dominguez Vasquez, Dr., University of Cuba, 2015
- Spiros Georgakopoulos, University of Thessaly, Greece, 2015
- Andrej Gisbrecht, Bielefeld University, Germany, 2014
- Akmal Haidar, Dr., University of Montreal, Canada, 2015
- David Hofmeyer, Dr., Lancaster University, United Kingdom
- Simon King, Dr., University of Edinburgh, UK, 2014
- Alejandro Lage-Castellanos, Dr., University of Cuba, 2015
- Hiroshi Mamitsuka, PhD, University of Kyoto, Japan, 2014
- Nicos Pavlidis, Dr., Lancaster University, UK, 2014
- Jose Moreno Pimentel, Universidad Politechnica de Madrid, Spain, 2014
- Alan Saul, University of Sheffield, UK, 2014
- Alexander Schulz, Bielefeld University, Germany, 2014
- Weitao Sun, Dr., Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing, 2015
- Yingying Xu, Tokyo Institute of Technology, Japan, 2015
- Fan Yang, Xiamen University, China, 2014
- Xiao Yang, Dr., Yale University, USA, 2014
- Ming Yi, Dr., Wuhan Institute of Physics and Mathematics, Chinese Academy of Sciences, China, 2015
- Hai-Jun Zhou, Dr., Chinese Academy of Sciences, China, 2015

Short research visits

- Florence d'Alché-Buc, Télécom ParisTech, 2014
- Cedric Archambeau, Amazon, Germany, 2014
- Armin Biere, Prof., Johannes Kepler University, Linz, Austria, 2014
- KyungHyun Cho, Prof., New York University, USA, 2015
- Corinna Cortes, Dr., Google, USA, 2014

- James Cussens, Dr., University of York, UK, 2015
- Martin Gebser, Prof., University of Potsdam, Germany, 2015
- Anna Goldernberg, Toronto University, Canada, 2014
- Susanne Graf, Dr., CNRS, Greboble, France, 2014
- Gustaf Eje Henter, Dr., University of Edinburgh, UK, 2015
- Kaspar Hornbaek, Copenhagen University, Denmark, 2014
- Vinay Jethava, Dr., Chalmers, Sweden, 2014
- Roland Kaminski, University of Potsdam, Germany, 2014
- Helen McNally, Prof., Purdue University, USA, 2015
- Brandon Malone, Dr., Max Plank Institute for Biology of Ageing, Germany, 2015
- Sara Mostafavi, University of British Columbia, 2014
- Roderick Murray-Smith, Prof., Glasgow, Scotland, 2015
- Ann Nicholson, Prof., Monash University, Australia, 2015
- Philipp Obermeier, University of Potsdam, Germany, 2014
- Max Ostrowski, University of Potsdam, Germany, 2014
- Max Ostrowski, University of Potsdam, Germany, 2015
- Razvan Pascanu, Dr., Google DeepMind, UK, 2015
- Jeroen de Ridder, TU Delft, The Netherlands, 2014
- Torsten Schaub, Prof., University of Potsdam, Germany, 2014-2015
- Sebastian Schellhorn, University of Potsdam, Germany, 2014
- Kana Shimizu, AIST, Japan, 2014
- Evgenia Ternovska, Prof., Simon Fraser University, Canada, 2015
- Ottokar Tilk, Tallinn University of Technology, Estonia, 2014
- Maria Uther, Prof., University of Winchester, UK, 2015
- Philipp Wanko, University of Potsdam, Germany, 2015
- Ole Winther, Dr., DTU, Denmark, 2014
- Yoshihiro Yamanishi, Kyushu University, Japan, 2014
- AIST researchers: Kana Shimizu, Goichiro Hanaoka, Tadanori Teruya, Koji Nuida, Nuttapong Attrapadung and Takahiro Matsuda, Japan, 2015

Doctoral dissertations

Foundations and Advances in Deep Learning

Kyunghyun Cho

Doctoral dissertation for the degree of Doctor of Science in Technology on the 21st of March 2014

External examiners:
Hugo Larochelle, Prof.
James Bergstra, Dr.
Opponent:
Nando de Freitas, Prof.

Abstract:
Deep neural networks have become increasingly popular under the name of deep learning recently due to their success in challenging machine learning tasks. Although the popularity is mainly due to recent successes, the history of neural networks goes as far back as 1958 when Rosenblatt presented a perceptron learning algorithm. Since then, various kinds of artificial neural networks have been proposed. They include Hopfield networks, self-organizing maps, neural principal component analysis, Boltzmann machines, multilayer perceptrons, radial-basis function networks, autoencoders, sigmoid belief networks, support vector machines and deep belief networks.

The first part of this thesis investigates shallow and deep neural networks in search of principles that explain why deep neural networks work so well across a range of applications. The thesis starts from some of the earlier ideas and models in the field of artificial neural networks and arrive at autoencoders and Boltzmann machines which are two most widely studied neural networks these days. The author thoroughly discusses how those various neural networks are related to each other and how the principles behind those networks form a foundation for autoencoders and Boltzmann machines.

The second part is the collection of the ten recent publications by the author. These publications mainly focus on learning and inference algorithms of Boltzmann machines and autoencoders. Especially, Boltzmann machines, which are known to be difficult to train, have been in the main focus. Throughout several publications the author and the co-authors have devised and proposed a new set of learning algorithms which includes the enhanced gradient, adaptive learning rate and parallel tempering. These algorithms are further applied to a restricted Boltzmann machine with Gaussian visible units.

In addition to these algorithms for restricted Boltzmann machines the author proposed a two-stage pretraining algorithm that initializes the parameters of a deep Boltzmann machine to match the variational posterior distribution of a similarly structured deep autoencoder. Finally, deep neural networks are applied to image denoising and speech recognition.

Computational Modeling and Simulation of Language and Meaning: Similarity-Based Approaches

Tiina Lindh-Knuutila

Doctoral dissertation for the degree of Doctor of Science in Technology on the 9th of May 2014

External examiners:
Hanna Suominen, Doc.
John A. Bullinaria, Dr.
Opponents:
Fred Karlsson, Prof. (Emeritus)
Ari Visa, Prof.

Abstract:
This dissertation covers various similarity-based, data-driven approaches to model language and lexical semantics. The availability of large amounts of text data in electronic form allows the use of unsupervised, data-driven methodologies. Compared to linguistic models based on expert knowledge, which are often costly or unavailable, the data-driven analysis is faster and more flexible. The same methodologies can be often used regardless of the language. In addition, data-driven analysis may be exploratory and offer a new view on the data.

The complexity of different European languages was analyzed at syntactic and morphological level using unsupervised methods based on compression and unsupervised morphology induction. The results showed that the unsupervised methods are able to produce useful analyses that correspond to linguistic models.

The distributional word vector space models represent the meaning of words in a text context of co-occurring words, collected from a large corpus. The vector space models were evaluated with linguistic models and human semantic similarity judgment data. Two unsupervised methods, Independent Component Analysis and Latent Dirichlet Allocation, were able to find groups of semantically similar words, corresponding reasonably well to the evaluation sets. In addition to validating the results of the unsupervised methods with the evaluation data, the research was also exploratory. The unsupervised methods found semantic word sets not covered by the evaluation set, and the analysis of the categories of the evaluation sets showed quality differences between the categories.

In the agent simulation models, the meaning of words was directly linked to the perceived context of the agent. Each agent had a subjective conceptual memory, in which the associations between words and perceptions were formed. In a population of simulated agents, the emergence of a shared vocabulary was studied through simulated language games. As a result of the simulations, a shared vocabulary emerges in the community.

Retrieval of Gene Expression Measurements with Probabilistic Models

Ali Faisal

Doctoral dissertation for the degree of Doctor of Science in Technology on the 15th of August 2014

External examiners:
Reija Autio, Dr.
Julio Saez-Rodriguez, Dr.
Opponent:
Hiroshi Mamitsuka, Prof.

Abstract:
A crucial problem in current biological and medical research is how to utilize the diverse set of existing biological knowledge and heterogeneous measurement data in order to gain insights on new data. As datasets continue to be deposited in public repositories it is becoming important to develop search engines that can efficiently integrate existing data and search for relevant earlier studies given a new study. The search task is encountered in several biological applications including cancer genomics, pharmacokinetics, personalized medicine and meta-analysis of functional genomics.

Most existing search engines rely on classical keyword or annotation based retrieval which is limited to discovering known information and requires careful downstream annotation of the data. Data-driven model-based methods, that retrieve studies based on similarities in the actual measurement data, have a greater potential for uncovering novel biological insights. In particular, probabilistic modeling provides promising model-based tools due to its ability to encode prior knowledge, represent uncertainty in model parameters and handle noise associated to the data. By introducing latent variables it is further possible to capture relationships in data features in the form of meaningful biological components underlying the data.

This thesis adapts existing and develops new probabilistic models for retrieval of relevant measurement data in three different cases of background repositories. The first case is a background collection of data samples where each sample is represented by a single data type. The second case is a collection of multimodal data samples where each sample is represented by more than one data type. The third case is a background collection of datasets where each dataset, in turn, is a collection of multiple samples. In all three setups the proposed models are evaluated quantitatively and with case studies the models are demonstrated to facilitate interpretable retrieval of relevant data, rigorous integration of diverse information sources and learning of latent components from partly related dataset collections.

Connectivity inference with asynchronously updated kinetic Ising models

Hong-Li Zeng

Doctoral dissertation for the degree of Doctor of Science in Technology on the 15th of August 2014

External examiners:
Manfred Opper, Prof.
Federico Ricci-Tersenghi, Prof.
Opponent:
Reimer Kühn, Prof.

Abstract:
This thesis focuses on the inference of network connections from statistical physics point of view. The reconstruction methods of the asynchronously updated kinetic Ising model with an asymmetric Sherrington-Kirkpatrick (SK) model is studied theoretically. Both approximate and exact learning rules for the couplings from the generated dynamical data are developed. The approximate formulae are based on naive mean field (nMF) and Thouless-Anderson-Palmer (TAP) equations respectively. The exact learning rules are derived for two cases: one in which both the spin history and the update times are known and one in which only the spin history. One can average over all possible choices of update times to obtain an averaged learning rule that depends only on spin correlations. We studied all the learning rules numerically. Good convergence is observed in accordance with the theoretical expectations.

The developed inference learning rules are applied to two data sets. One is spike trains recorded from 20 retinal ganglion cells and the other is generated by transactions of 100 highly traded stocks on the New York Stock Exchange (NYSE).

For the neuron data set, we compared the inferred asynchronous couplings with the equilibrium ones. The results show that the inferred couplings from these two models are very similar. This implies that real dynamical process of the neuron system satisfies the Gibbs equilibrium conditions and that the final distribution of states is the Gibbs stationary distribution.

For the financial data set, three inference methods are applied to reconstruct the coupling matrices between traded stocks. They are equilibrium, synchronous and asynchronous inference formula respectively. All of them are based on mean-field approximation. Synchronous and asynchronous Ising inference methods give results which are coherent with equilibrium case, but more detailed since the obtained interaction networks are directed.

Continuous Context Inference on Mobile Platforms

Sourav Bhattacharya

Doctoral dissertation for the degree of Doctor of Science in Technology on the 25th of August 2014

External examiners:
Nicholas Lane, Dr.
Jörg Ott, Prof. Dr.-Ing.
Opponent:
Antonio Kruüger, Prof.

Abstract:
In this thesis we develop novel methods for continuous and sustained context inference on mobile platforms. We address challenges present in real- world deployment of two popular context recognition tasks within ubiquitous computing and mobile sensing, namely localization and activity recognition. In the first part of the thesis, we provide a new localization algorithm for mobile devices using the existing GSM communication infrastructures, and then propose a solution for energy efficient and robust tracking on mobile devices that are equipped with sensors such as GPS, compass, and accelerometer. In the second part of the thesis we propose a novel sparse-coding-based activity recognition framework that mitigates the time-consuming and costly bootstrapping process of activity recognizers employing supervised learning. The framework uses a vast amount of unlabeled data to automatically learn a sensor data representation through a set of extracted characteristic patterns and generalizes well across activity domains and sensor modalities.

Bayesian latent variable models for learning dependencies between multiple data sources

Seppo Virtanen

Doctoral dissertation for the degree of Doctor of Science in Technology on the 25th of August 2014

External examiners:
Teemu Roos, Asst. Prof.
Guillaume Obozinski, Dr.
Opponent:
Cédric Archambeau, Dr.

Abstract:
Machine learning focuses on automated large-scale data analysis extracting useful information from data collections. The data are frequently high-dimensional and may correspond, for example, to images, text documents, or measurements of neural responses. In many applications data can be collected from multiple data sources, that is, views.

This thesis presents novel machine learning methods for analyzing multiple data sources, especially for understanding relationships between them. The analysis provides a comprehensive summary of the data generating process, which may be used for exploring the relationships and for predicting observations of one or more sources. The methods are based on two assumptions: each view provides complementary information of the data generating process, and each view is corrupted by noise. The methods aim to utilize all available information (views), accumulating partly overlapping information and reducing view-specific noise.

In particular, this thesis presents several Bayesian latent variable models that learn a decomposition of latent variables; some of the variables capture information shared by multiple sources, whereas the remaining variables explain noise in each view. The latent variables may be efficiently inferred based on the observed data by using sparsity assumptions and Bayesian inference. The models are applied for analyzing neural responses to natural stimulation as well as for jointly modeling images and text documents.

Probabilistic components of molecular interactions and drug responses

Juuso Parkkinen

Doctoral dissertation for the degree of Doctor of Science in Technology on the 29th of August 2014

External examiners:
Matti Nykter, Prof.
Motoki Shiga, Prof.
Opponent:
Yoshihiro Yamanishi, Prof.

Abstract:
A fundamental question in medicine is how cancer and other complex diseases operate on the molecular level. Identifying the detailed mechanisms and interactions of how diseases progress and respond to drug treatments is essential for developing effective therapies. High-throughput molecular profiling technologies have provided vast amounts of measurement data of these phenomena. However, making sense of these masses of data is far from straightforward and requires advanced computational analysis methods.

Probabilistic component models have been proven an effective tool in analysing and integrating high-dimensional and noisy molecular profiling data sources, such as gene expression. Such models can identify coherent components from the data, and interpreting these components provides insights about the underlying biological processes, such as disease progression and drug responses. In this thesis, probabilistic component models are applied and extended to identify and analyse molecular interaction and drug response patterns.

Identifying functionally coherent gene modules from high-throughput measurements is a central task in many biomedical applications. In this thesis, an earlier component model for network data is extended for capturing functional modules from combinations of gene expression and protein interaction data. The identified modules provide hypotheses for novel molecular pathways and protein functions.

High-throughput drug treatment measurements have made possible the detailed analysis of molecular drug responses and toxicity. In this thesis, probabilistic component models are applied to detect coherent drug response patterns from gene expression data. These patterns provide detailed insights to drug mechanisms of action and are highly applicable in cancer therapy development. Moreover, by associating the identified drug response components to toxicological outcomes, the first comprehensive view of molecular toxicogenomic responses is constructed with high performance in drug toxicity prediction.

The Effects of Mobility on Mobile Input

Joanna Bergström-Lehtovirta

Doctoral dissertation for the degree of Doctor of Science in Technology on the 30th of August 2014

External examiners:
Enrico Rukzion, Prof.
Roope Raisamo, Prof.
Opponent:
Kasper Hornæk, Prof.

Abstract:
Mobile interfaces are designed for interaction while the user is on the move across mobile contexts. Walking, handling a wallet, visually attending to the environment, and even simply carrying a mobile device, however, can have a negative effect on mobile human-computer interaction (HCI). Understanding the negative effects of mobility is important because potential exists for overcoming them via good interface design. Previous work has shown that mobility decreases input performance with a mobile interface. However, the causes of declines in performance often remain unclear and, with them, possible avenues for compensation. I argue that systematic variation of physical constraints emerging from mobile conditions in controlled experiments can reveal considerable effects of mobility on mobile input. Among these physical constraints are competing allocations of hand function, body movement, and various sensory modalities.

The thesis contributes to mobile HCI research by examining the effects of four causes of physical constraints in mobility: 1) gripping of the device, 2) walking, 3) manipulation of external objects, and 4) sensory feedback. The research includes four studies, isolating one constraint each in controlled experiments. In the first two, the levels of grip position and walking speed are varied systematically, for modeling of their effects on mobile input. The other two vary the presence of external objects and sensory feedback.

The findings highlight the constraints' negative effect on manual input performance. However, all of the studies also reveal unaffected sensory or motor resources of the user. Across the four studies, the following findings were made. First, a model for the functional area of the thumb predicts the reachable interface elements on a mobile touchscreen as a function of grip, hand size, and screen size. Secondly, a function describing the tradeoff between walking speed and input performance demonstrates that while walking hampers input performance, users can adjust to an optimal walking speed for mobile interaction. Thirdly, interface design is shown to significantly affect input performance when the user simultaneously manipulates external objects or when sensory feedback is limited. This work calls for mobile HCI research to consider the operationalization of mobile conditions in controlled experiments that can extend knowledge of the effects of mobility on interaction. It also invites exploitation of the empirical results and the proposed methods and models in practice for interface evaluation and design.

Advances in Nonnegative Matrix Decomposition with Application to Cluster Analysis

He Zhang

Doctoral dissertation for the degree of Doctor of Science in Technology on the 19th of September 2014

External examiners:
Rafal Zdunek, Research Scientist Ph.D.
Morten Mørup, Associate Prof.
Opponent:
Ali Taylan Cemgil, Associate Prof.

Abstract:
Nonnegative Matrix Factorization (NMF) has found a wide variety of applications in machine learning and data mining. NMF seeks to approximate a nonnegative data matrix by a product of several low-rank factorizing matrices, some of which are constrained to be nonnegative. Such additive nature often results in parts-based representation of the data, which is a desired property especially for cluster analysis.

This thesis presents advances in NMF with application in cluster analysis. It reviews a class of higher-order NMF methods called Quadratic Nonnegative Matrix Factorization (QNMF). QNMF differs from most existing NMF methods in that some of its factorizing matrices occur twice in the approximation. The thesis also reviews a structural matrix decomposition method based on Data-Cluster-Data (DCD) random walk. DCD goes beyond matrix factorization and has a solid probabilistic interpretation by forming the approximation with cluster assigning probabilities only. Besides, the Kullback-Leibler divergence adopted by DCD is advantageous in handling sparse similarities for cluster analysis.

Multiplicative update algorithms have been commonly used for optimizing NMF objectives, since they naturally maintain the nonnegativity constraint of the factorizing matrix and require no user-specified parameters. In this work, an adaptive multiplicative update algorithm is proposed to increase the convergence speed of QNMF objectives.

Initialization conditions play a key role in cluster analysis. In this thesis, a comprehensive initialization strategy is proposed to improve the clustering performance by combining a set of base clustering methods. The proposed method can better accommodate clustering methods that need a careful initialization such as the DCD.

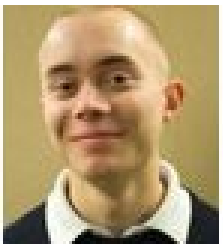
The proposed methods have been tested on various real-world datasets, such as text documents, face images, protein, etc. In particular, the proposed approach has been applied to the cluster analysis of emotional data.

Bayesian Multi-Way Models for Data Translation in Computational Biology

Tommi Suvitaival

Doctoral dissertation for the degree of Doctor of Science in Technology on the 19th of November 2014

External examiners:
Laura Elo-Uhlgren, Dr.
Lukas Käll, Dr.
Opponent:
Anna Goldenberg, Asst. Prof.



Abstract:
The inference of differences between samples is a fundamental problem in computational biology and many other sciences. Hypothesis about a complex system can be studied via a controlled experiment. The design of the controlled experiment sets the conditions, or covariates, for the system in such a way that their effect on the system can be studied through independent measurements. When the number of measured variables is high and the variables are correlated, the assumptions of standard statistical methods are no longer valid. In this thesis, computational methods are presented to this problem and its follow-up problems.

A similar experiment done on different systems, such as multiple biological species, leads to multiple "views" of the experiment outcome, observed in different data spaces or domains. However, cross-domain experimentation brings uncertainty about the similarity of the systems and their outcomes. Thus, a new question emerges: which of the covariate effects generalize across the domains? In this thesis, novel computational methods are presented for the integration of data views, in order to detect weaker covariate effects and to generalize covariate effects to views with unobserved data.

Five main contributions to the inference of covariate effects are presented: (1) When the data are high-dimensional and collinear, the problem of false discovery is curbed by assuming a cluster structure on the observed variables and by handling the uncertainty with Bayesian methods. (2) Prior information about the measurement process can be used to further improve the inference of covariate effects for metabolomic experiments by modeling the multiple layers of uncertainty in the mass spectral data. (3-4) When the data come from multiple measurement sources on the same subjects - that is, from data views with co-occurring samples - it is unknown, whether the covariate effects generalize across the views and whether the outcome of a new intervention can be generalized to a view with no observed data on that intervention. These problems are shown to be possible to solve by assuming a shared generative process for the multiple data views. (5) When the data come from different domains with no co-occurring samples, the inference of between-domain dependencies is not possible in the same way as with co-occurring samples. It is shown that even in this situation, it is possible to identify covariate effects that generalize across the domains, when the experimental design at least weakly binds the domains together. Then, effects that generalize are identified by assuming a shared generative process for the covariate effects.

Probabilistic Modelling of Multiresolution Biological Data

Prem Raj Adhikari

Doctoral dissertation for the degree of Doctor of Science in Technology on the 21th of November 2014

External examiners:
Marko Bohanec, Prof. Dr.
Olli Yli-Harja, Prof.
Opponent:
Jeroen de Ridder, Asst. Prof.

Abstract:
When the measurements from the ever improving measurement technology are accumulated over a period of time, the result is the collection of data in different representations. However, most machine learning and data mining algorithms, in their standard form, are designed to operate on data in single representation.

This thesis proposes machine learning and data mining algorithms to analyze data in different representation with respect to the resolution within a single analysis. The novel algorithms proposed to analyze multiresolution data are in the field of probabilistic modelling and semantic data mining. First, three different deterministic data transformation methods are proposed to transform data across different resolutions. After the data transformation, the resulting data in same resolution are integrated and modeled using mixture models.

Second, similar mixture components in a mixture model are merged one by one repetitively to generate a chain of mixture models. A new fast approximation of the KL-divergence is derived to determine the similarity of the mixture components. The chain of generated mixture models are useful for comparison, for example, in model selection. Third, mixture components in different resolutions are iteratively merged to model multiresolution data generating models in each modeled resolution that incorporate information from data in other resolution.

Fourth, a single multiresolution mixture model with multiresolution mixture components is proposed whose mixture components independently have the capabilities of a Bayesian network. Finally, three-part methodology consisting of clustering using mixture models, rule learning using semantic subgroup discovery, and pattern visualization using banded matrices is developed for comprehensive analysis of multiresolution data.

The multiresolution data analysis methods presented in this thesis improves the performance of the methods in comparison with the their single resolution counterparts. Furthermore, developed methods aims to make the results understandable to the domain experts. Therefore, the developed methods are useful addition in the analysis of chromosomal aberration patterns and the cancer research in general.

Extending SAT Solver with Parity Reasoning

Tero Laitinen

Doctoral dissertation for the degree of Doctor of Science in Technology on the 21st of November 2014

External examiners:
Chu Min Li, Prof.
Roberto Sebastiani, Prof.
Opponent:
Armin Biere, Prof.

Abstract:
Propositional conflict-driven clause-learning (CDCL) satisfy ability (SAT) solvers have been successfully applied in a number of industrial domains. In some application areas such as circuit verification, bounded model checking, logical cryptanalysis, and approximate model counting, some requirements can be succinctly captured with parity (xor) constraints. However, satisfy ability solvers that typically operate in conjunctive normal form (CNF) may perform poorly with straightforward translation of parity constraints to CNF.

This work studies how CDCL SAT solvers can be enhanced to handle problems with parity constraints using the recently introduced DPLL (XOR) framework where the SAT solver is coupled with a parity constraint solver module. Different xor-deduction systems ranging from plain unit propagation through equivalence reasoning to complete incremental Gauss-Jordan elimination are presented. Techniques to analyze xor-deduction system derivations are developed, allowing one to obtain smaller clausal explanations for implied literals and also to learn new parity constraints in the conflict analysis process. It is proven that these techniques can be used to simulate a complete xor-deduction system on a restricted class of instances and allow very short unsatisfiability proofs for some formulas whose CNF translations are hard for resolution. Fast approximating tests to detect whether unit propagation or equivalence reasoning is enough to deduce all implied literals are presented. Methods to decompose sets of parity constraints into sub problems that can be handled separately are developed. The decomposition methods can greatly reduce the size of parity constraint matrices when using Gauss-Jordan elimination on dense matrices and allow one to choose appropriate xor-deduction system for each sub problem. Efficient translations to simulate equivalence reasoning and stronger parity reasoning are developed. It is shown that equivalence reasoning can be simulated by adding a polynomial amount of redundant parity constraints to the problem, but without using additional variables, an exponential number of parity constraints are needed in the worst case. It is proven that resolution simulates equivalence reasoning efficiently. The presented techniques are experimentally evaluated on a variety of challenging problems originating from a number of encryption ciphers and from SAT Competition benchmark instances.

From pixels to semantics: visual concept detection and its applications

Mats Sjöberg

Doctoral dissertation for the degree of Doctor of Science in Technology on the 25th of November 2014

External examiners:
Joni Kämäräinen, Prof.
Georges Quénot, Dr.
Opponent:
Bernard Merialdo, Prof.

Abstract:
The amount of digital visual information available in the world today is enormous, and the rate at which more is continuously generated is simply unbelievable. For example YouTube gets 100 hours of new video every minute, and Facebook more than 350 million new photos every day. At best, this represents the creativity and knowledge of millions or even billions of people, made available to the entire world thanks to the Internet. The problem is of course: how do we find the "needle" that is relevant to us in this enormous "haystack"? Web search engines such as Google and Bing are decent solutions to find textual content, but finding relevant visual content is as yet an unsolved problem. The core issue is the semantic gap between the raw visual data processed by computers, and the abstract concepts and ideas humans use to communicate.

This thesis studies one approach to this problem, namely using mid-level concepts to bridge the semantic gap. These semantic concepts are e.g. objects, locations, persons or events which are relatively concrete and thus comparatively easy to associate with the raw visual data. These can then be used to formulate more abstract queries, or used to index and further organise an image or video database.

An overview of semantic concept detection using machine learning techniques is presented here, together with some applications. A central issue is keeping the computational speed and efficiency at a practical level for huge amounts of visual data, while still producing accurate and relevant results. To this end, this thesis studies several fast approximative versions of the popular Support Vector Machine (SVM) algorithm, and proposes some improvements to the fast Self-Organising Map (SOM) algorithm to improve its accuracy. Several large-scale real-world experimental applications are presented including image retrieval using social network tags, video search, indoor location recognition, and semantic visualisation of large image and video databases.

The empirical evidence presented in this thesis shows that while the semantic gap problem is still not solved, the semantic concept approach produces concrete improvements to real-world applications. The improvements proposed and evaluated contribute to making the machine learning algorithms faster and thus more practically useful for processing huge amounts of visual data.

Advances in Analysis and Exploration in Medical Imaging

Nicolau Gonçalves

Doctoral dissertation for the degree of Doctor of Science in Technology on the 5th of December 2014

External examiners:
Miika Nieminen, Prof.
Nikola Kasabov, Prof.
Opponent:
Kristoffer H. Madsen, Dr.

Abstract:
With an ever increasing life expectancy, we see a concomitant increase in diseases capable of disrupting normal cognitive processes. Their diagnoses are difficult, and occur usually after daily living activities have already been compromised. This dissertation proposes machine learning methods for the study of the neurological implications of brain lesions. It addresses the analysis and exploration of medical imaging data, with particular emphasis to (f)MRI. Two main research directions are proposed. In the first, a brain tissue segmentation approach is detailed. In the second, a document mining framework, applied to reports of neuroscientific studies, is described. Both directions are based on retrieving consistent information from multi-modal data.

A contribution in this dissertation is the application of a semi-supervised method, discriminative clustering, to identify different brain tissues and their partial volume information. The proposed method relies on variations of tissue distributions in multi-spectral MRI, and reduces the need for a priori information. This methodology was successfully applied to the study of multiple sclerosis and age related white matter diseases. It was also showed that early-stage changes of normal-appearing brain tissue can already predict decline in certain cognitive processes.

Another contribution in this dissertation is in neuroscience meta-research. One limitation in neuroimage processing relates to data availability. Through document mining of neuroscientific reports, using images as source of information, one can harvest research results dealing with brain lesions. The context of such results can be extracted from textual information, allowing for an intelligent categorisation of images. This dissertation proposes new principles, and a combination of several techniques to the study of published fMRI reports. These principles are based on a number of distance measures, to compare various brain activity sites. Application to studies of the default mode network validated the proposed approach.

The aforementioned methodologies rely on clustering approaches. When dealing with such strategies, most results depend on the choice of initialisation and parameter settings. By defining distance measures that search for clusters of consistent elements, one can estimate a degree of reliability for each data grouping. In this dissertation, it is shown that such principles can be applied to multiple runs of various clustering algorithms, allowing for a more robust estimation of data agglomeration.

SMT-based Verification of Timed Systems and Software

Roland Kindermann

Doctoral dissertation for the degree of Doctor of Science in Technology on the 5th of December 2014

External examiners:
Joost-Pieter Katoen, Prof.
Fabio Somenzi, Prof.
Opponent:
Susanne Graf, Dr.

Abstract:
Defects in hardware or software can have disastrous consequences. Traditionally, testing has been used to address this threat. While often able to expose bugs, testing can, however, not guarantee that a given system is correct, as has been demonstrated by catastrophic failures of well-tested systems in the past. Verification approaches such as model checking address this shortcoming, not only searching for flaws in a limited set of scenarios, but by trying to prove a system correct, guaranteeing the absence of defects if successful.

The main part of this dissertation discusses various topics in the area of symbolic model checking of timed systems. Unlike finite state systems, most commonly used for verification, timed systems allow to faithfully model timing aspects of the verified system. Symbolic model checking methods attempt to efficiently handle large sets of states using concise representations.

This work contributes to different areas in the field of symbolic verification of timed systems. Firstly, several different symbolic verification methods are explored: bounded model checking, a timed variant of the IC3 algorithm, timed k-induction, and verification by reduction to finite state model checking. Apart from the reduction approach, a common theme among the methods addressed is that they leverage the power of modern SMT solvers to efficiently verify timed systems. Secondly, this work addresses the symbolic verification of quantitative specifications on the timing of events in a system, made in a PSPACE-verifiable subset of the logic MITL. Thirdly, a new representations of timed systems designed to facilitate symbolic verification is introduced.

While not providing the same guarantees as model checking, testing has the advantage that it can usually be performed using the original system instead of a model of the system. The approach of concolic testing aims to combine the advantages of both approaches, executing the system at hand instead of analyzing a model, while at the same time using an SMT solver to guide the executions to maximize coverage. This dissertation evaluates the use of concolic testing for the purpose of differential testing of Java smart card applets.

Real-time Action Recognition for RGB-D and Motion Capture Data

Xi Chen

Doctoral dissertation for the degree of Doctor of Science in Technology on the 16th of January 2015

External examiners:
Guoying Zhao, Prof.
Vassilis Athitsos, Prof.
Opponent:
Ivan Laptev, Dr.

Abstract:
In daily life humans perform a great number of actions continuously. We recognize and interpret these actions unconsciously while interacting and communicating with people and the environment. If the machines and computers could also recognize human gestures as effectively as human beings, a new world would be unfolded, filled with a large number of applications to facilitate our daily life. These significant benefits for the society have motivated the research on machine-based gesture recognition, which has already shown some initial advantages in many applications. For example, gestures can be used as commands to control robots or computer programs instead of using standard input devices such as touch screens or mice.

This thesis proposes a framework for gesture recognition systems based on motion capture and RGB-D data. Motion capture data consists of positions and orientations of the key joints of the human skeleton. RGB-D data contains the RGB image and depth data from which a skeletal model can be learnt. This skeletal model can be seen as a noisy approximation of the more accurate motion capture skeleton model. The modular design of our framework enables convenient recognition using multiple data modalities.

The first part of the thesis introduces various methods used in existing recognition systems in the literature and a brief introduction of the proposed real-time recognition system for both whole body gestures and hand gestures. The second part of the thesis is a collection of eight publications by the author of the thesis. Detailed information about the proposed recognition system can be found in these publications. In general, the framework can be roughly divided into two parts, feature extraction and classification. Both have significant influence on the recognition performance. Multiple features are developed and extracted from the skeletons, images, and depth data for each frame in the motion sequence. These features are combined in the early fusion stage, and classified by a single hidden layer neural network - extreme learning machine. The frame-level classification outputs are then aggregated on the sequence level to obtain the final classification result.

The methodologies used in the gesture recognition system are also applied in a proposed image retrieval system. Several image features are extracted and search algorithms are applied to achieve a fast and accurate retrieval. Furthermore, a method is also proposed to align different motion sequences and to evaluate the alignment. The method can be used for gesture retrieval and for skeleton generation algorithm evaluation.

Bayesian Stochastic Partition Models For Markovian Dependence Structures

Väinö Jääskinen

Doctoral dissertation for the degree of Doctor of Science on the 6th of February 2015

External examiners:
Harri Lähdesmäki, Prof.
Jaakko Nevalainen, Prof.
Opponent:
Korbinian Stimmer, Reader

Abstract:
In various fields of knowledge we can observe that the availability of potentially useful data is increasing fast. A prime example is the DNA sequence data. This increase is both an opportunity and a challenge as new methods are needed to benefit from the big data sets. This has sparked a fruitful line of research in statistics and computer science that can be called machine learning. In this thesis, we develop machine learning methods based on the Bayesian approach to statistics. We address a fairly general problem called clustering, i.e. dividing a set of objects to non-overlapping group based on their similarity, and apply it to models with Markovian dependence structures. We consider sequence data in a finite alphabet and present a model class called the Sparse Markov chain (SMC). It is a special case of a Markov chain (MC) model and offers a parsimonious description of the data generating mechanism. A Variable length Markov chain (VLMC) is a popular sparse model presented earlier in the literature and it has a representation as an SMC model. We develop Bayesian clustering methodology for learning the SMC and other Markovian models.

Another problem that we study in this thesis is causal inference. We present a model and an algorithm for learning causal mechanisms from data. The model can be considered as a stochastic extension of the sufficient-component cause model that is popular in epidemiology. In our model there are several causal mechanisms each with its own parameters. A mixture distribution gives a probability that an outcome variable is associated with a mechanism.

Applications that are considered in this thesis come mainly from computational biology. We cluster states of Markovian models estimated from DNA sequences. This gives an efficient description of the sequence data when comparing to methods reported in the literature. We also cluster DNA sequences with Markov chains, which results in a method that can be used for example in the estimation of bacterial community composition in a sample from which DNA is extracted. The causal model and the related learning algorithm are able to estimate mechanisms from fairly challenging data. We have developed the learning algorithms with big data sets in mind. Still, there is a need to develop them further to handle ever larger data sets.

Exact Inference Algorithms and Their Optimization in Bayesian Clustering

Jukka Kohonen

Doctoral dissertation for the degree of Doctor of Science on the 20th of March 2015

External examiners:
Camilla Hollanti, Prof.
Lasse Holmström, Prof.
Opponent:
Jose M. Peña, Prof.

Abstract:
Clustering is a central task in computational statistics. Its aim is to divide observed data into groups of items, based on the similarity of their features. Among various approaches to clustering, Bayesian model-based clustering has recently gained popularity. Many existing works are based on stochastic sampling methods.

This work is concerned with exact, exponential-time algorithms for the Bayesian model-based clustering task. In particular, we consider the exact computation of two summary statistics: the number of clusters, and pairwise incidence of items in the same cluster. We present an implemented algorithm for computing these statistics substantially faster than would be achieved by direct enumeration of the possible partitions. The method is practically applicable to data sets of up to approximately 25 items.

We apply a variant of the exact inference method into graphical models where a given variable may have up to four parent variables. The parent variables can then have up to 16 value combinations, and the task is to cluster them and find combinations that lead to similar conditional probability tables.

Further contributions of this work are related to number theory. We show that a novel combination of addition chains and additive bases provides the optimal arrangement of multiplications, when the task is to use repeated multiplication starting from a given number or entity, but only a certain kind of function of the successive powers is required. This arrangement speeds up the computation of the posterior distribution for the number of clusters. The same arrangement method can be applied to other multiplicative tasks, for example, in matrix multiplication.

We also present new algorithmic results related to finding extremal additive bases. Before this work, the extremal additive bases were known up to length 23. We have computed them up to length 24 in the unrestricted case, and up to length 41 in the restricted case.

Advances in Extreme Learning Machines

Mark van Heeswijk

Doctoral dissertation for the degree of Doctor of Science in Technology on the 17th of April 2015

External examiners:
Guang-Bin Huang, Prof.
Jonathan Tapson, Prof.
Opponent:
Donald C. Wunsch, Prof.

Abstract:
Nowadays, due to advances in technology, data is generated at an incredible pace, resulting in large data sets of ever-increasing size and dimensionality. Therefore, it is important to have efficient computational methods and machine learning algorithms that can handle such large data sets, such that they may be analyzed in reasonable time. One particular approach that has gained popularity in recent years is the Extreme Learning Machine (ELM), which is the name given to neural networks that employ randomization in their hidden layer, and that can be trained efficiently. This dissertation introduces several machine learning methods based on Extreme Learning Machines (ELMs) aimed at dealing with the challenges that modern data sets pose. The contributions follow three main directions.

Firstly, ensemble approaches based on ELM are developed, which adapt to context and can scale to large data. Due to their stochastic nature, different ELMs tend to make different mistakes when modeling data. This independence of their errors makes them good candidates for combining them in an ensemble model, which averages out these errors and results in a more accurate model. Adaptivity to a changing environment is introduced by adapting the linear combination of the models based on accuracy of the individual models over time. Scalability is achieved by exploiting the modularity of the ensemble model, and evaluating the models in parallel on multiple processor cores and graphics processor units. Secondly, the dissertation develops variable selection approaches based on ELM and Delta Test, that result in more accurate and efficient models. Scalability of variable selection using Delta Test is again achieved by accelerating it on GPU. Furthermore, a new variable selection method based on ELM is introduced, and shown to be a competitive alternative to other variable selection methods. Besides explicit variable selection methods, also a new weight scheme based on binary/ternary weights is developed for ELM. This weight scheme is shown to perform implicit variable selection, and results in increased robustness and accuracy at no increase in computational cost. Finally, the dissertation develops training algorithms for ELM that allow for a flexible trade-off between accuracy and computational time. The Compressive ELM is introduced, which allows for training the ELM in a reduced feature space. By selecting the dimension of the feature space, the practitioner can trade off accuracy for speed as required.

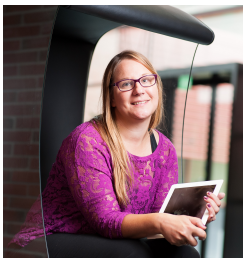
Overall, the resulting collection of proposed methods provides an efficient, accurate and flexible framework for solving large-scale supervised learning problems. The proposed methods are not limited to the particular types of ELMs and contexts in which they have been tested, and can easily be incorporated in new contexts and models.

Developing augmented reality solutions through user involvement

Sanni Siltanen

Doctoral dissertation for the degree of Doctor of Science in Technology on the 22nd of May 2015

External examiners:
Mark Billinghurst
Andreas Dünser
Opponent:
Kaisa Väänänen-Vainio-Mattila, Prof.



Abstract:
Augmented reality (AR) technology merges digital information into the real world. It is an effective visualization method; AR enhances user's spatial perception skills and helps to understand spatial dimensions and relationships. It is beneficial for many professional application areas such as assembly, maintenance and repair. AR visualization helps to concretize building and construction projects and interior design plans - also for non-technically oriented people, who might otherwise have difficulties in understanding what the plans actually mean in the real context. Due to its interactive and immersive nature AR is applied for games and advertising as well.

Although AR is proven to be a valuable visualization method it is not yet commonly used in beneficial consumer level applications. This work first finds out reasons for this and then focuses on developing AR towards wider use. The work is threefold: it considers human factors affecting adoption of the technology, economic factors affecting the viability of AR technology, and development of applications and technical solutions that support these factors.

In this thesis user centric and participatory methods are used to find out reasons that hinder the use of AR, especially in interior design. The outcomes of the studies are manifold: desired features for AR services, bottlenecks preventing the use, user experience (UX) issues and business viability factors. A successful AR solution needs to have a viable business ecosystem besides a reliable technical framework.

The presented application development in assembly guidance and interior design visualization considers UX factors and demonstrates the use of AR in the field of question.

A serious bottleneck for using AR in interior design arises from a typical use situation; a consumer wants to redesign a room. The space where the interior design plan is made is not empty and augmentation does not look realistic when the new furniture is rendered on top of the existing furniture. This problem can be solved by using diminished reality, which means that the old furniture is removed digitally from the view.

This work presents a diminished reality solution for AR interior design. A complete pipeline implementing diminished reality functionality is described. Algorithms and methods are developed to achieve real time high quality diminished reality functionality. The presented practical solution has a great effect for the whole AR interior design field, and enhances it towards real use.

The possibilities of using AR are huge. In order to make beneficial AR solutions,

researchers should be able to reveal the users' needs - both existing and emerging ones - and develop technology to fulfil those needs. This thesis demonstrates that this can be achieved by developing augmented reality solutions through user involvement.

Predictive Classification and Bayesian Inference

Jie Xiong

Doctoral dissertation for the degree of Doctor of Science on the 15th of June 2015

External examiners:
Daniel Thorburn, Prof.
Mattias Villani, Prof.
Opponent:
Arnoldo Frigessi, Prof.

Abstract:
A general inductive probabilistic framework for clustering and classification is introduced using the principles of Bayesian predictive inference, such that all quantities are jointly modelled and the uncertainty is fully acknowledged through the posterior predictive distribution. Several learning rules have been considered and the theoretical results are extended to acknowledge complex dependencies within the datasets. Multiple probabilistic models have been developed for analysing data from a wide variate of fields of application. State-of-art algorithms are introduced and developed for the model optimization.

Advances in Weakly Supervised Learning of Morphology

Oskar Kohonen

Doctoral dissertation for the degree of Doctor of Science in Technology on the 26th of August 2015

External examiners:
Chris Dyer, Prof.
Filip Ginter, Dr.
Opponent:
Lars Borin, Prof.

Abstract:
Morphological analysis provides a decomposition of words into smaller constituents. It is an important problem in natural language processing (NLP), particularly for morphologically rich languages whose large vocabularies make statistical modeling difficult. Morphological analysis has traditionally been approached with rule-based methods that yield accurate results, but are expensive to produce. More recently, unsupervised machine learning methods have been shown to perform sufficiently well to benefit applications such as speech recognition and machine translation. Unsupervised methods, however, do not typically model allomorphy, that is, non-concatenative structure, for example pretty/prettier. Moreover, the accuracy of unsupervised methods remains far behind rule-based methods with the best unsupervised methods yielding between 50-66% F-score in Morpho Challenge 2010.

We examine these problems with two approaches that have not previously attracted much attention in the field. First, we propose a novel extension to the popular unsupervised morphological segmentation method Morfessor Baseline to model allomorphy via the use of string transformations. Second, we examine the effect of weak supervision on accuracy by training on a small annotated data set in addition to a large unannotated data set. We propose two novel semi-supervised morphological segmentation methods, namely a semi-supervised extension of Morfessor Baseline and morphological segmentation with conditional random fields (CRF). The methods are evaluated on several languages with different morphological characteristics, including English, Estonian, Finnish, German and Turkish. The proposed methods are compared empirically to recently proposed weakly supervised methods.

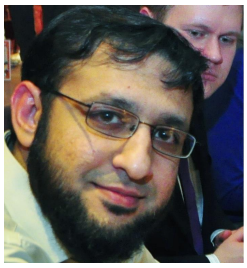
For the non-concatenative extension, we find that, while the string transformations identified by the model have high precision, their recall is low. In the overall evaluation the non-concatenative extension improves accuracy on English, but not on other languages. For the weak supervision we find that the semi-supervised extension of Morfessor Baseline improves the accuracy of segmentation markedly over the unsupervised baseline. We find, however, that the discriminatively trained CRFs perform even better. In the empirical comparison, the CRF approach outperforms all other approaches on all included languages. Error analysis reveals that the CRF excels especially on affix accuracy.

Bayesian multi-view models for data-driven drug response analysis

Suleiman Ali Khan

Doctoral dissertation for the degree of Doctor of Science in Technology on the 7th of September 2015

External examiners:
Sampsa Hautaniemi, Prof.
Manfred Claassen, Prof.
Opponent:
Sara Mostafavi, Asst. Prof.



Abstract:
A central challenge faced by biological and medical research is to understand the impact of chemical entities on living cells. Identifying the relationships between the chemical structures and their cellular responses is valuable for improving drug design and targeted therapies. The chemical structures and their detailed molecular responses need to be combined through a systematic analysis to learn the complex dependencies, which can then assist in improving understanding of the molecular mechanisms of drugs as well as predictions on the effects of unknown molecules. Moreover, with emerging drug-response data sets being profiled over several disease types and phenotypic details, it is pertinent to develop advanced computational methods that can be used to study multiple sets of data together.

In this thesis, a novel multi-disciplinary challenge is undertaken for computationally analyzing interactions between multiple biological responses and chemical properties of drugs, while simultaneously advancing the computational methods to better learn these interactions. Specifically, multi-view dependency modeling of paired data sets is formulated as a means of systematically studying the drug-response relationships. First, the systematic analysis of drug structures and their genome-wide responses is presented as a multi-set dependency modeling problem and established methods are adopted to test the novel hypothesis.

Several novel extensions of the drug-response analysis are then presented that explore responses measured over multiple disease types and multiple levels of phenotypic detail, uncovering novel biological insights of potential impact. These analyses are made possible by novel advancements in multi-view methods. Specifically, the first Bayesian tensor canonical correlation analysis and its extensions are introduced to capture the underlying multi-way structure and applied in analyzing novel toxicogenomic interactions. The results illustrate that modeling the precise multi-view and multi-way formulation of the data is valuable for discovering interpretable latent components as well as for the prediction of unseen responses of drugs.

Therefore, the original contribution to knowledge in this dissertation is two-fold: first, the data-driven identification of relationships between structural properties of drugs and their genome-wide responses in cells and, second, novel advancements of multi-view methods that find dependencies between paired data sets. Open source implementations of the new methods have been released to facilitate further research.

Statistical studies on bacterial transmission and community dynamics : with a special emphasis on the colonization dynamics of Streptococcus pneumoniae during early childhood

Elina Numminen

Doctoral dissertation for the degree of Doctor of Science on the 2nd of October 2015

External examiners:
Kari Auranen, Prof.
Philip O'Neill, Prof.
Opponent:
Gianpaolo Scalia-Tomba, Prof.

Abstract:

A central goal in science is to learn from observations about the process that generated the observations. The principles of statistical inference describe a systematic approach for such learning, in which prior information, knowledge about the underlying mechanisms and the observed data can be combined. In practice, lack of mathematical tractability, huge amounts of missing information, and the sensitivity of the conclusions on the assumptions made represent genuine challenges in the theoretically sound statistical framework. Statistical studies on the dynamics of infectious diseases easily face all these problems at once.

In the thesis we present case-studies in which the datasets on bacterial diversity, mostly on Streptococcus pneumoniae, described in terms of either genotypes or serotypic strains, are analysed. By utilizing the machinery of modern computational statistics different strategies for inference are formulated, which aim to take the special characteristics of each of the studied problem into account, while overcoming the previously mentioned challenges in computational studies. For instance, an approximate Bayesian computation scheme is formulated for analysing cross-sectional strain prevalence data and an importance sampling scheme for analysing transmission trees with a priori known complex features. The obtained results unravel the mechanisms of seasonality in pneumococcal carriage, consequences of the host population structure and the nature of within-host competition between the bacterial strains.

Data Analysis and Next Generation Sequencing: Applications in Microbiology

Nicolas Innocenti

Doctoral dissertation for the degree of Doctor of Science on the 30th of October 2015

Opponent:
Mogens Høgh Jensen, Prof.

Abstract:

Next Generation Sequencing (NGS) is a new technology that has revolutionized the way we study living organisms. Where previously only a few genes could be studied at a time through targeted direct probing, NGS offers the possibility to perform measurements for a whole genome at once. The drawback is that the amount of data generated in the process is large and extracting useful information from it requires new methods to process and analyze it.

The main contribution of this thesis is the development of a novel experimental method coined tagRNA-seq, combining 5?tagRACE, a previously developed technique, with RNA-sequencing technology. Briefly, tagRNA-seq makes it possible to identify the 5? ends of RNAs in bacteria and directly probe for their type, primary or processed, by ligating short RNA sequences, the tags, to the beginnings of RNA molecules. We used the method to directly probe for transcription start and processing sites in two bacterial species, Escherichia coli and Enterococcus faecalis. It was also used to study polyadenylation in E. coli, where the ability to identify processed RNA molecules proved to be useful to separate direct and indirect regulatory effects of this mechanism. We also demonstrate how data from tagRNA-seq experiments can be used to increase confidence on the discovery of anti-sense transcripts in bacteria. A detailed analysis of the data revealed subtle artifacts in the coverage signal towards 3?ends of genes, that we were able to explain and quantify based on Kolmogorov?s broken stick model. We also discovered evidences for circularization of a few RNA transcripts, both in our own data sets and publicly available data.

Designing the tags used in tagRNA-seq led us to the problem of words absent from a text. We focus on a particular subset of these, the minimal absent words (MAWs), and develop a theory providing a complete description of their size distribution in random text. Genomes from viruses and living organisms have MAWs a large fraction of which are well modeled by the theory, but almost always exhibit a behavior different from random texts in the tail of the distribution. MAWs from this tail are closely related to sequences present in the genome that preferentially appear in regions with important regulatory functions. Finally, and independently from tagRNA-seq, we propose a new approach to the problem of bacterial community reconstruction in metagenomic, based on techniques from compressed sensing. We provide a novel algorithm competing with state-of-the-art techniques in the field.

Theses

Master of Science

2014

Airaksinen, Sami
HIP: Model Combination Algorithm for Location Prediction

Ali, Javeria
Exploratory data analysis of microbial communities in food production environment

Amid, Ehsan
Application for alpha-Divergence for Stochastic Neighbor Embedding in Data Visualization

Berg, Jeremias
Cost Optimal Correlation Clustering via Partial Maximum Satisfiability

Bomanson, Jori
Developing Efficient Encodings for Weighted Expressions in Answer Set Programs

Ferreira, Tiago Joel Domingues
Catch the dream Wave. Propagation of Cortical Slow Oscillation to the Striatum in anaesthetised mice

Grönroos, Stig-Arne
Semi-supervised induction of a concatenative morphology with simple morphotactics: A model in the Morfessor family

Hazara, Murtaza
Unsupervised methods in multilingual and multimodal semantic modeling

Karppa, Matti
Estimating Hand Configurations from Sign Language Videos

Kontturi, Matias
Computer vision system for tracking in-app purchases in mobile devices

Konyushkova, Ksenia
Relevance feedback content-based image retrieval

Lebre Magalhães Pereira, João
Supervised Learning for Relationship Extraction from Textual Documents

Leppä-aho, Janne
Pseudo-Likelihood Learning of Gaussian Graphical Models

Lintusaari, Jarno
PCSI-labeled Directed Acyclic Graphs

Luzardo Escandon, Marcos
Eye and Mouth Openness Estimation in Sign Language and News Broadcast Videos

Modig, Arttu
A Metric for Human Motor Capacity

Pantourakis, Michail
Ionic mechanisms in regulation of C-fiber following frequency: Insights from modeling using single and repetitive stimulation

Paul, Debdas
Efficient Parameter Inference for Stochastic Chemical Kinetics

Suvilehto, Jyry
Short Document Information Retrieval for Product Recommendation

2015

Aggarwal, Kunal
T-Cell Receptor Diversity in the Human Immune System

Ali, Mehreen
Survival Modeling Using Factor Analysis Data Integration

Antonova, Evgenia
Applying Answer Set Programming in Game Level Design

Bajpai, Namrata
Cell Lineage Tracking. Implementation of automated lineage tracking on 4D confocal image data in the MorphoGraphX software.

Chan, Yat Hin
Experimentally-based Mathematical Modeling to Analyze T Helper 17 Cell Differentiation in Heterogeneous Cell Populations

Deng, Huining
Implementation and evaluation of a parallel algorithm for structure learning in Bayesian networks

Eraslan, Basak
A probabilistic model for competitive DNA binding modeling using ChIP-seq and MNase-seq data

Hedman, Peter
Sequential Monte Carlo instant radiosity

Heiskanen, Tomas
Ranking extension for kernelized Bayesian matrix factorization

Heuer, Hendrik
Semantic and stylistic text analysis and text summary evaluation

Hore, Sayantan
Designing Interfaces for Exploratory Content Based Image Retrieval Systems

Hyvönen, Ville
Approximate nearest neighbor search using multiple random projection trees

Jahnsson, Niklas
Hashing-based delayed duplicate detection as an approach to improve the scalability of optimal Bayesian network structure learning with external memory frontier breadth-first branch and bound search

Kauppinen, Aki
Datan ja tietämyksen soveltaminen vaativassa tuotteen valinnassa Case: älykäs venttiilin valinta

Kayal, Subhradeep
Audiovisual Speaker Clustering for News Broadcast Videos

Koponen, Laura
Constraint-Based Optimization of Phylogenetic Supertrees

Leino, Katri
Maximum A Posteriori for Acoustic Model Adaptation in Automatic Speech Recognition

Leinonen, Juho
Automatic Speech Recognition for Human-Robot Interaction Using an Under-Resourced Language

Lempiäinen, Tuomas
The PATS problem: search methods and reliability

Leppä-aho, Janne
Pseudo-Likelihood Learning of Gaussian Graphical Models

Mirzaei, Saeideh
Improving Accuracy in Automatic Speech Recognition Systems by Model Adaptation Techniques

Nguyen, Quan
Likelihood-based Phylogenetic Network Inference by Approximate Structural Expectation Maximization

Ni, Shuai
Computational prediction of ETS-regulated elements confer prostate cancer suspectibility

Palon, Preston
Comparison of popular methods for prediction of film ratings

Perello Nieto, Miquel
Merging chrominance and luminance in early, medium, and late fusion using Convolutional Neural Networks

Polvi-Huttunen, Silja
Matrix Factorization for Learning Metagenomic Pathways and Species

Pradhananga, Sailendra
Association studies of exome sequencing data of lung cancer patients undergoing gemcitabine/carboplatin chemotherapy with myelosuppression toxity

Pusa, Taneli
Marginal pseudolikehood in labelled graphical models

Saikko, Paul
Re-implementing and extending a hybrid SAT-IP approach to maximum satisfiability

Sakaya, Joseph
Scalable Bayesian induction of word embeddings

Sankar, Aravind
Identification of bacterial strains from sequencing data using probabilistic modeling

Soppela, Jyri
Nonnegative Matrix Factorization in Text Mining Applications

Sotala, Kaj
Bayes Academy – An Educational Game for Learning Bayesian Networks

Strahl, Jonathan
Patient appointment scheduling system: with supervised learning prediction

Suotsalo, Kimmo
Machine Learning for Structure Discovery in Vector Autoregressive Processes

Wu, Weiming
Design and Implementation of a Shared Task Queue Groupware

Xie, Zhe
From Exploration to Sensemaking: an Interactive Exploratory Search System

Licentiate Theses

2015

Luukkala, Vesa
Rule-Based Metaprogramming for Smart Spaces

Nybo, Kristian
Dimensionality reduction methods for fMRI analysis and visualization

Research Projects

Introduction

Samuel Kaski, Director of COIN

The Finnish Centre of Excellence in Computational Inference Research (COIN), started in January 2012, has rapidly become a creative research environment that bridges three departments and two universities: the Department of Computer Science at Aalto University (coordinator), the Department of Computer Science and the Department of Mathematics and Statistics at the University of Helsinki.

Given the on-going data science revolution made possible by digitalization, the importance of the core technology of computational inference does not need to be emphasized. The overall objective of COIN is to forge and deliver large-scale data-intensive computational modelling and inference techniques which infer what is relevant in vast data masses, and can transform large quantities of raw data from many kinds of sources into useful information.

The ultimate goal of COIN is to develop methodologies for learning more structured models from the combination of data and prior knowledge, and for doing accurate statistical inference rapidly, and to apply the methods to solve key problems in carefully chosen application fields. This is likely to need multi-level modeling where simpler models are used globally to approximate relationships between local more structured models.

COIN works on four core challenges (C1-C4 in Fig. 1) and two flagship applications. The

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

Figure 1: *The organization of the COIN Centre of Excellence. In the matrix, 'o' means that the PI (row) is in charge of the coordination of the challenge or flagship (column), while '×' signifies active collaboration. C1: Learning models from massive data; C2: Learning from multiple data sources; C3: Statistical inference in highly structured stochastic models; C4: Extreme inference; F1: Intelligent information access; F2: Computational molecular biology and medicine.*

first flagship (F1) is Intelligent information access, or contextual interfacing to relevant information in real-world setups. Lack of prior knowledge requires data-driven modeling from massive data (C1) to infer relevance of information from multimodal data (C2) and feedback sources. It is particularly important to get the inferences about relevance immediately (C4) for interactive use.

Data-driven modeling is necessary in much of Computational molecular biology and medicine (F2) as well, since not enough is known about cellular mechanisms. When prior information is available and taken into account in modeling, the models become hierarchically more structured and current methods of statistical inference quickly become intractable as the size of the model increases. This calls for new methods for statistical inference in structured models (C3).

COIN interacts with the rest of the society through multiple channels that include (i) carefully chosen collaboration projects in other domains, with companies and in particular other national Centres of Excellence or equivalent top-level groups in Finland and abroad, (ii) education of highly specialized professionals whose skills match with central challenges of modern information society, and (iii) launching open source software packages and data sets that are widely used also outside academia.

At the midterm evaluation of COIN, the importance of the core technical challenges C1-C4 and flagship applications F1-F2 was noted to be constantly increasing as the progression of data revolution requires inference on larger and more complex problems. During the latter 3-year period, of 2015–2017, COIN focuses in particular on three of the most successful directions: (i) Fostering of fundamental research that combines so-far mostly separately studied aspects of inference. In applications we will focus in particular on two spearheads: (ii) Inference on intractable models, with applications in particular in studies of bacterial evolution. (iii) Interactive intent modelling with the SciNet system for intelligent information access.

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 are strongly emphasized, following the successful existing traditions.

This Biennial Report 2014–2015 details the individual research projects of the six groups during the middlemost two years of the six-year period of COIN. Additional information is available at `www.research.cs.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

The goal of challenge area C1 is to address the methodological problems arising when dealing with massive data sets, and to develop machine learning methods that scale up to real-world "Big Data" settings. It is obvious that this challenge area is very much in parallel with the other challenge areas, and that the same general problems appear in all of our research efforts in various more focused settings. All in all, in our research we consider both unsupervised and supervised machine learning problems. In the unsupervised setting, we have focused our work on using Bayesian networks and other probabilistic graphical models. In challenge area C4, the learning problem is approached from the optimization perspective, and we utilize the expertise of several COIN groups to develop computationally efficient methods for learning graphical models. Below in Section 1.2, we highlight some results that focus on special subclasses of Bayesian networks, where the inference task can be shown to be in fact tractable. An alternative approach for scaling up machine learning is not to restrict the complexity of the model classes under consideration, but to restrict the complexity or the volume of the data. One way to accomplish this is to decrease the dimensionality of the data by reducing the number of features in the data; in Section 1.3, we consider this task in the supervised learning setting, and in Section 1.4, in the unsupervised setting. The latter approach leads to highly efficient methods that can be used for clustering or visualizing very large data sets. It should also be noted that some of the methods developed are sequential in nature, allowing incremental processing of data, which means that the methods are feasible for streaming data, and moreover, as there is no need for iterative batch processing of the whole data set, the size of the data is actually no longer a limiting factor.

1.2 Speeding up Unsupervised Learning

Both the inference and learning tasks in Bayesian networks are NP-hard in general. One approach to deal with this issue has been to investigate special cases where these problems would be tractable. That is, the basic idea is to select models from a restricted class of Bayesian networks that have structural properties enabling fast learning or inference; this way, the computational complexity will not be an issue, though possibly at the cost of accuracy if the true distribution is far from the model family. Most notably, it is known that the inference task can be solved in polynomial time if the network has *bounded tree-width*.

The possibility of tractable inference has motivated several studies also on learning bounded tree-width Bayesian networks. We have attacked this problem by using dynamic programming [2], integer linear programming [4] and weighted partial maximum satisfiability [1]. However, unlike in the case of inference, learning a Bayesian network of bounded tree-width is NP-hard for any fixed tree-width bound at least 2 [2]. Furthermore, it is known that learning many relatively simple classes such as paths and polytrees is also NP-hard. Indeed, until recently the only class of Bayesian networks for which a polynomial time learning algorithm was known were trees, i.e., graphs with tree-width 1.

We have recently proposed *bounded vertex cover number Bayesian networks* [3] as an alternative to the tree-width paradigm. Roughly speaking, we consider Bayesian networks where all pairwise dependencies – i.e., edges in the so-called moralised graph – are covered by having at least one node from the vertex cover incident to each of them; see Figure 1.1

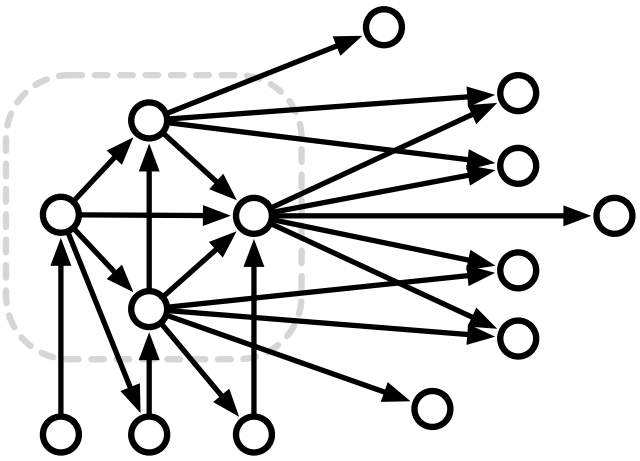


Figure 1.1: A directed acyclic graph. The four vertices circled with dashed line are the minimum vertex cover; the vertex cover number is 4.

for an example. We have shown that learning an optimal Bayesian network structure with vertex cover number at most k can be done in polynomial time for any fixed k. Moreover, vertex cover number provides an upper bound for tree-width, implying that inference is also tractable; thus, we identify a rare example of a class of Bayesian networks where both learning and inference are tractable.

1.3 Feature Selection for Supervised Learning

In order to be able to estimate increasingly complex models from increasingly large data sets, it is often necessary to implement a screening stage to restrict attention to a subset of all the available variable features. An important direction for research is feature selection in streaming data where the algorithms need to be able to process data sequentially as they become available. We have studied a class of information-theoretically justified feature selection techniques for linear regression and provided theoretical guarantees of their consistency in different circumstances [9, 10].

In another study [11], we consider Lasso-based methods for feature selection in a setting where the features are automatically generated by constructing interaction terms of increasingly high order so that the model can represent combinations of any logical functions of the base features. Ongoing research is directed towards combining these feature selection methods to the problem of learning Bayesian network structures by reducing the structure learning problem to a set of regression problems, which links this theme to several other themes mentioned in this report.

1.4 Dimension Reduction and Visualization

One of the most interesting approaches to handling very high dimensional data, where traditional principal component based techniques are intractable, is random projections. We have studied the use of random projections for two problems: clustering and fast

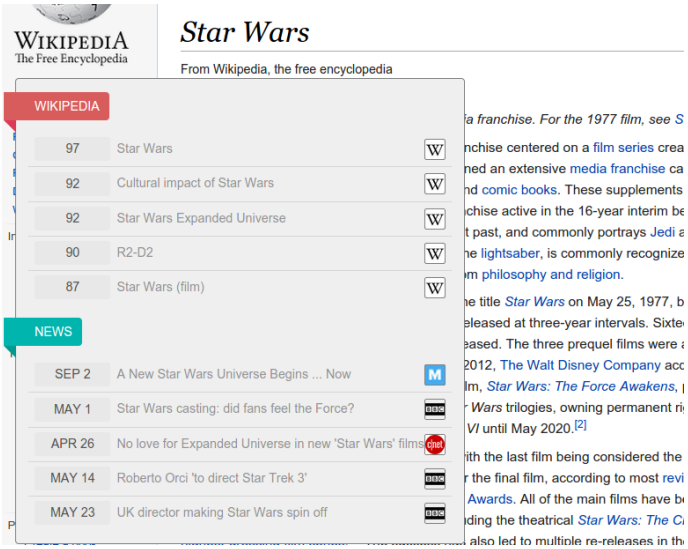


Figure 1.2: A real-time recommendation system Kvasir incorporated into the Chrome browser demonstrates the use of fast approximate nearest neighbor search [14]. Kvasir was a semifinalist in the 2015 Cambridge Postdoc Business Plan Competition.

approximate nearest neighbor queries in large data sets. The proposed clustering methods can be used to carry out certain population genomics analyses several orders of magnitude faster than state-of-the-art Bayesian model-based techniques [13]. The proposed nearest neighbor search method is widely applicable in situations where real-time processing is required and approximate solutions are acceptable [12]. A good example is a web-scale interactive information retrieval application Kvasir [14, 15]; see Fig. 1.2.

Another popular approach to dimension reduction is topic modeling. Efficient estimation procedures can be implemented through Gibbs sampling and related techniques. In [16] we apply topic models to develop efficient Bayesian inference algorithms for supervised regression of multiple count-valued outputs, allowing their variance to be lower or higher than what typical models based on Poisson distribution would assume.

Dimension reduction can also be used for visualization (where the reduced space has two dimensions). Non-linear visualizations are often based on neighborhood graphs, and thus, the fast nearest neighbor solutions mentioned above are a key to fast non-linear visualizations. In [17] we describe a simple and efficient visualization algorithm called Minimap, which is designed to find a balance between the high-level (global) view and the low-level (local) structure.

1.5 Applications

The estimation of the evolutionary relationships between different organisms, i.e., phylogenetics, is a challenging computational problem. The existence of not only vertical gene transfer, which is the cause of dominantly tree-like structures, but also horizontal transfer, gives rise to phylogenetic *networks*. Their estimation is an even more challenging task than the traditional problem of estimating phylogenetic trees. We have proposed a phylogenetic network method that is based on several approximation techniques from probabilistic

graphical models [7]. As a novel application for phylogenetic networks, we have applied the proposed method for the analysis of cultural traditions such as folktales [8].

Another bioinformatics application is the discovery of sequence motifs in transcription factor binding sites. A traditional approach to characterizing motifs is an independence model where each DNA symbol in the motif is independent of the other symbols. By exploiting sequence prediction models we have developed earlier [5], we were able to find intra-motif dependencies that allow a more accurate classification of potential binding sites than the traditional models [6].

References

[1] J. Berg, M. Järvisalo and B. Malone. Learning Optimal Bounded Treewidth Bayesian Networks via Maximum Satisfiability. *17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

[2] J.H. Korhonen and P. Parviainen. Learning bounded tree-width Bayesian networks. *16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013.

[3] J.H. Korhonen and P. Parviainen. Tractable Bayesian Network Structure Learning with Bounded Vertex Cover Number. *Advances in Neural Information PRocessing Systems 28 (NIPS)*, 2015.

[4] P. Parviainen, S.H. Farahani and J.Lagergren. Learning Bounded Tree-width Bayesian Networks using Integer Linear Programming. *17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

[5] R. Eggeling, T. Roos, P. Myllymäki, I. Grosse. Robust learning of inhomogeneous PMMs, in *Proc. 17th Conf on Artificial Intelligence and Statistics (AISTATS), JMLR W&CP 33*, pp. 229–237, 2014.

[6] R. Eggeling, T. Roos, P. Myllymäki, I. Grosse. Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data, *BMC Bioinformatics* 16:375, 2015.

[7] Q. Nguyen and T. Roos. Likelihood-based inference of phylogenetic networks from sequence data by PhyloDAG, in *Proc. 2nd International Conference on Algorithms for Computational Biology (AlCoB), LNBI 9199*, Springer, pp. 126-140, 2015.

[8] J. Tehrani, Q. Nguyen, and T. Roos. Oral fairy tale or literary fake? Investigating the origins of Little Red Riding Hood using phylogenetic network analysis, to appear in *Digital Scholarship in the Humanities*, 2016.

[9] J. Määttä, D.F. Schmidt, and T. Roos. Subset selection in linear regression using sequentially normalized least squares: Asymptotic theory, *Scandinavian Journal of Statistics* 43(2):382–395, 2016.

[10] J. Määttä and T. Roos (2016). Robust sequential prediction in linear regression with Student's t-distribution, in *Proc. 14th International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2016.

[11] Y. Zou and T. Roos (2016). Sparse logistic regression with logical features, in *Proc. 20th Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, Auckland, New Zealand, 2016.

[12] V. Hyvönen, T. Pitkänen, S. Tasoulis, L. Wang, T. Roos, and J. Corander. Fast k-NN search, arXiv:1509.06957, 2015.

[13] S. Tasoulis, L. Cheng, N Valimäki, N.J Croucher, S.R Harris, W.P Hanage, T. Roos, J. Corander. Random projection based clustering for population genomics. In *2014 IEEE International Conference on Big Data*, Oct. 27–30 2014, pp. 675–682, Washington, DC, 2014.

[14] L. Wang, S. Tasoulis, T. Roos and J. Kangasharju. Kvasir: Seamless Integration of Latent Semantic Analysis-Based Content Provision into Web Browsing, In *24th International World Wide Web Conference*, 2015.

[15] L. Wang, S. Tasoulis, T. Roos, and J. Kangasharju. Kvasir: Scalable provision of semantically relevant web content on big data framework, to appear in *IEEE Transactions on Big Data*.

[16] Arto Klami, Abhishek Tripathi, Johannes Sirola, Lauri Väre, and Frederic Roulland. Latent feature regression for multivariate count data. In *P*roceedings of Artificial Intelligence and Statistics, Volume 38 of JMLR C&WP, 2015.

[17] Y. Zhao, S. Tasoulis, and T. Roos Manifold visualization via short walks, to appear in EuroVis-2016.

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. Group Factor Analysis GFA is a generalization of factor analysis to this setting, and additionally generalizes Canonical Correlation Analysis (CCA) in a new way to more than two data sets. While GFA can be seen as an unsupervised method in the sense that no variables are in the special role of "outputs" to be predicted, the Kernelized Bayesian Matrix Factorization is a supervised method in which the goal is to predict missing values of one matrix given other matrices. A new line of work is *data set search*, to find sets useful when analysing a new set.

2.2 Group factor analysis

Group factor analysis (GFA) [11] is a novel factor analysis technique for multi-view settings. Given co-occurring observations from multiple data sources, the task is to provide a latent representation that captures the relationships between the sources and separates them from variation independent of the other sources. For two sources it is sufficient to identify the correlations between the views, and for that special case, GFA is equivalent to canonical correlation analysis (CCA). The core difficulty in solving the GFA problem in the general case is that for more than two sources we also need to identify relationships between all possible subsets of the sources, of which there are exponentially many.

Our solution for the GFA problem builds on Bayesian inference of a joint factor analysis model and it uses group-wise sparsity priors and constraints for identifying the factors. The sparsity-inducing priors relax the combinatorial inference problem into a continuous one, enabling us to identify the factor structure efficiently. For a good overview of the overall concept and illustrations for both artificial and real data collections with tens or hundreds of co-occurring data sources, see [11]. When the approach is extended to allow sparsity across features and samples, it can be used for biclustering of multiple data sources. In a case study on cancer cell line measurements, besides finding interpretable bi-cluster structure in the data, the method also turned out to provides outstandingly accurate predictions on the effect of drugs on the cell lines [4].

The general idea of automatically finding shared and private components extends also beyond the multi-view learning setup. In [10] we applied it for *collective matrix factorization*, the task of finding latent low-dimensional representations for arbitrary collections of matrices. Besides matrices, we showed that the approach is useful for modelling data collections with paired tensors and matrices, outperforming recent work not utilizing this type of group sparsity. In this work, we additionally showed how to relax the strict tensor decomposition assumptions [9].

We have also worked on an alternative solution for the problem of extracting shared information in multiple parallel data sources, building on chains of regressors instead of generative models [12]. The CoCoReg (*Common Components by Regression*) algorithm focuses on extracting only the shared information and can find also nonlinear relationships.

It is computationally efficient and easy to implement using existing regression models as building blocks, yet it finds the shared information more effectively than GFA especially when the relationships are nonlinear. However, CoCoReg is not able to extract the private components and hence does not solve the full GFA problem.

2.3 Kernelized Bayesian matrix factorization

Multiple kernel learning methods have been successful in integrating several data sources in supervised learning tasks, and solving multiple related predictions tasks together with multi-task learning helps "borrow strength" across the tasks. That is particularly important in large p small n domains, of large dimensionality p compared to small sample size n, which are common in computational biology and medicine (F2), where 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 multi-task multiple kernel learning methods [5] that are directly applicable to personalized medicine prediction problems of F2, and turned out to win 43 alternative methods in a public competition. The methods were generalized to matrix factorization given multiple side information sources [8]. The result is effectively a flexible, Bayesian non-linear recommender engine which can use the side information sources to make out-of-matrix predictions.

2.4 Retrieval of experiments

Current technologies enable data from experiments of ever increasing multitude and complexity to be produced at high frequency and in large volumes. Therefore, a main challenge of data-driven sciences is how to make maximal use of the progressively expanding databases of experimental data in order to keep research cumulative. As a complementary task to using textual annotations for retrieving relevant data sets from a database, which relies on manually added meta-data, we develop methods to do retrieval using the measurement values of an entire data set of interest as query input. Our research in this line of work attempts to go beyond purely data-driven retrieval by constructing probabilistic models of the data sets, enabling prior information to be incorporated into each model and to be utilized in the retrieval task. We also develop methods for the subsequent step of doing combined analysis of the retrieved models using novel forms of meta-analysis [6].

In [7], an approach was proposed, which assumes that the query data set can be modelled as a mixture of a fixed set of previously learnt models, each representing one dataset. Here the measure of relevance was given by the inferred mixture weights, with a large weight implying high relevance. The approach is well scalable as it reduces to an optimization problem which has linear time complexity (sublinear for an approximate version). In [13], experiments were retrieved by evaluating the posterior marginal likelihoods of individual models stored for the experiments in the database. This enables the relevance evaluations to be done in parallel, which even further improves scalability. Both of the above approaches are based on maximizing a likelihood for the query data, which has the advantage that the models are not required to belong to the same family for a valid measure of relevance to be defined. However, for very noisy and high-dimensional query data, this may result in the retrieval criterion to become noisy as well.

To reduce noise in the query, retrieval was done in [3] using a model as query input, instead of data, and evaluating the relevance between models using distances between models in the induced model space (see Figure 2.1). Also made explicit in the approach, was the importance of marginalizing out nuisance parameters not of direct relevance for the retrieval task. For example, in a gene expression experiment, one is often more interested in how sets of genes are co-regulated, rather than their exact expression values, which are additionally affected by numerous other influences. Gene expression data sets were modelled using a probabilistic clustering model [2], providing a straightforward way of characterizing each experiment with minimal data preprocessing, while capturing central co-expression patterns.

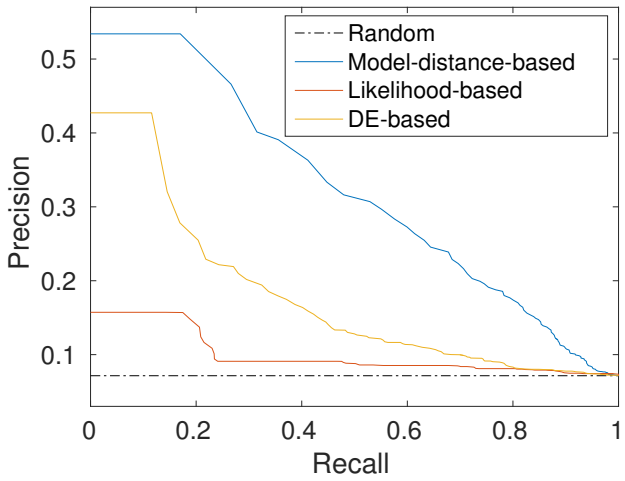


Figure 2.1: Retrieval of gene expression experiments using known cell type as ground truth. For noisy data, using distances between models to evaluate relevance shows an advantage over using marginal likelihoods given the query data, or using correlations between profiles of gene-specific differential expression p-values.

2.5 Learning from multimodal media data

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 that we have addressed in COIN. In particular, a novel unsupervised method for aural feature extraction was proposed in [1] and the results showed that the method is capable of extracting more applicable features for multimedia event detection than those commonly used in speech and audio recognition.

Automatic image captioning is an interesting problem that aims to integrate computer vision and natural language modeling. It has recently been at the center of attention and experienced a rapid growth of research. The growth bas been mostly due to the appearance of large annotated datasets and the availability of the computational power required to handle such large amounts of visual and textual data. For example, the Microsoft Common Objects in Context (COCO) database contains over 200,000 images with at least five human-written captions per image. The recent approaches to image captioning often rely on deep Convolutional Neural Networks (CNN) and Long-Short Term Memory (LSTM) models as key ingredients of their pipeline. The CNN compresses an image into a feature vector to be inputted to an LSTM network that acts as a generative language model. In our recent research, we have attained state-of-the-art captioning results [14] with the COCO database by using explicit scene context features for LSTM language models and developing novel techniques for picking the best available caption from an ensemble of caption generator models.

References

[1] Amid E., Mesaros A., Palomäki K., Laaksonen J., and Kurimo M. 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.

[2] Blomstedt, P., Tang, J., Xiong, J., Granlund, C., and Corander, J. (2015). A Bayesian predictive model for clustering data of mixed discrete and continuous type. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**(3), 489–498.

[3] Blomstedt, P., Dutta, R., Seth, S., Brazma, A., and Kaski, S. (2016). Modelling-based experiment retrieval: A case study with gene expression clustering. *Bioinformatics*, **32**(9), pages 1388–1394.

[4] Kerstin Bunte, Eemeli Leppäaho, Inka Saarinen, and Samuel Kaski Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics*, accepted for publication, 2016.

[5] James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, Muhammad Ammad-ud-din, Petteri

Hintsanen, Suleiman A Khan, John-Patrick Mpindi, Olli Kallioniemi, Antti Honkela, Tero Aittokallio, Krister Wennerberg, NCI DREAM Community, James J Collins, Dan Gallahan, Dinah Singer, Julio Saez-Rodriguez, Samuel Kaski, Joe W Gray, and Gustavo Stolovitzky. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32:1202–1212, 2014.

[6] Dutta, R., Blomstedt, P., and Kaski, S. (2016). Bayesian inference in hierarchical models by combining independent posteriors. *arXiv:1603.09272 [stat.CO]*, submitted to a journal.

[7] Faisal, A., Peltonen, J., Georgii, E., Rung, J., and Kaski, S. (2014). Toward computational cumulative biology by combining models of biological datasets. *PLoS ONE*, **9**(11), e113053.

[8] Mehmet Gönen and Samuel Kaski. Kernelized Bayesian matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:2047–2060, 2014.

[9] Suleiman A. Khan, Eemeli Leppäaho, and Samuel Kaski Bayesian multi-tensor factorization. *Machine learning*, accepted for publication, 2016.

[10] Arto Klami, Guillaume Bouchard, and Abhishek Tripathi. Group-sparse embeddings in collective matrix factorization. In *Proceedings of International Conference on Learning Representations*, 2014.

[11] Arto Klami, Seppo Virtanen, Eemeli Leppäaho, and Samuel Kaski. Group factor analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9):2136–2147, 2015

[12] Jussi Korpela, Andreas Henelius, Lauri Ahonen, Arto Klami, and Kai Puolamäki. Using regression makes extraction of shared variation in multiple datasets easy. *Data Mining and Knowledge Discovery*, doi:10.1007/s10618-016-0465-y, 2016.

[13] Seth, S., Shawe-Taylor, J., and Kaski, S. (2014). Retrieval of experiments by efficient comparison of marginal likelihoods. In C. Loo, K. Yap, K. Wong, A. Teoh, and K. Huang, editors, *Neural Information Processing*, volume 8835 of *Lecture Notes in Computer Science*, pages 135–142. Springer International Publishing.

[14] Shetty R., R.-Tavakoli H., Laaksonen J. Exploiting Scene Context for Image Captioning. In *ACM Multimedia conference workshop on Vision and Language Integration Meets Multimedia Fusion*, 2016.

Chapter 3

C3: Statistical Inference in Structured Stochastic Models

Jukka Corander Michael Gutmann, Luca Martino, Elina Numminen, Ville Parkkinen, Jukka Siren, 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 with 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. Related families introducing partial independence in Bayesian networks (IJAR 2015) and in undirected stratified graphical models (SGMs), also termed as labeled Markov networks (Bayesian Analysis 2014), have 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 further generalized the latter model family by introducing the CSI constrains via a log-linear parametrization (Computational Statistics 2015), which does not necessitate the decomposability restriction. 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 and we developed a highly scalable recursive learning approach for this model class (Bayesian Analysis 2015). To generalize the concept of CSIs to continuous variables, we have also 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 was further proven that SGGMs represent a curved exponential family (CSTM, 2015).

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 experiments MPL was both more accurate and faster than the state-of-the-art approaches based on regularized logistic regression and conditional mutual information.

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, GAPIS) and MCMC algorithms (FUSS, OMCMC) 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 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. The efficiency of the nonreversible stochastic optimization approach was demonstrated e.g. in the IEEE TPAMI 2014 paper introducing a Bayesian clustering model for data of mixed discrete and continuous types.

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. 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 ABC method for estimating transmission trees for bacteria from longitudinal data by combining sequential adaptive Monte Carlo with branching process models. Our highlight work on likelihood-free 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 (JMLR 2015). Our experiments suggest that this approach can be several orders of magnitude faster than the standard sampling methods widely used in ABC. Another highlight paper 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] Numminen E. et al. Two-phase importance sampling for inference about transmission trees. Proc. R. Soc. B (2014)

[2] Gutmann M. and Corander J. Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models. Journal of Machine Learning Research, accepted for publication in August 2015. (2015)

[3] Gutmann M. et al. Classification and Bayesian Optimization for Likelihood-Free Inference. arXiv:1502.05503. (2014)

[4] Martino, L. et al. An Adaptive Population Importance Sampler. In proceedings of the 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP). (2014)

[5] Martino, L. et al. Orthogonal MCMC algorithms. In proceedings of the IEEE Statistical Signal Processing Workshop (SSP14). (2014)

[6] Nyman, H. et al. Stratified Graphical Models - Context-Specific Independence in Graphical Models. Bayesian Analysis (2014)

[7] Nyman H. et al. Marginal and simultaneous predictive classification using stratified graphical models. Advances in Data Analysis and Classification, (2015)

[8] Nyman H. et al. Stratified Gaussian graphical models. Communications in Statistics, Theory and Methods, in press. (2015)

[9] Pensar J. et al. Labeled Directed Acyclic Graphs: a generalization of context-specific independence in directed graphical models. Data Mining and Knowledge Discovery, (2014)

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

[11] Martino, L. et al. MCMC-Driven Adaptive Multiple Importance Sampling. Springer Proceedings in Mathematics and Statistics for the 12th Brazilian Meeting on Bayesian Statistics (EBEB) (2014)

[12] Wei, L. et al. On the outage capacity of orthogonal space-time block codes over multi-cluster scattering MIMO channels. IEEE Transactions on Communications, doi: 10.1109/TCOMM.2015.2419258. (2015)

[13] Martino, L. et al. Smelly Parallel MCMC Chains. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). (2015)

[14] Elvira, V. et al. A Gradient Adaptive Population Importance Sampler. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). (2015)

[15] Martino, L. et al. Interacting Parallel Markov Adaptive Importance Sampling. European Signal Processing Conference (EUSIPCO). (2015)

[16] Martino, L. et al. A Fast Universal Self-tuned Sampler within Gibbs sampling. Digital Signal Processing (2015)

[17] Martino, L. et al. An Adaptive Population Importance Sampler: Learning from Uncertainty. IEEE Transactions on Signal Processing, doi: 10.1109/TSP.2015.2440215. (2015)

[18] Bugallo, M. et al. Adaptive importance sampling in signal processing. Digital Signal Processing (2015)

[19] Blomstedt, P. et al. A Bayesian predictive model for clustering data of mixed discrete and continuous type. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2014)

[20] Pensar, J. et al. The role of local partial independence in learning of Bayesian networks. International Journal of Approximate Reasoning. (2015)

[21] Xiong, J. et al. Recursive learning of sparse Markov models Bayesian Analysis, (2015)

[22] Nyman, H. et al. Context-specific independence in graphical log-linear models Computational Statistics, (2015)

Chapter 4

C4: Extreme Inference

Ilkka Niemelä, Tomi Janhunen, Jukka Corander, Samuel Kaski, Petri Myllymäki, Jeremias Berg, Jori Bomanson, Kerstin Bunte, Martin Gebser, Antti Hyttinen, Matti Järvisalo, Roland Kindermann, Guohua Liu, Brandon Malone, Emilia Oikarinen, Jaakko Peltonen, Jussi Rintanen, Paul Saikko, Johannes Wallner

4.1 Introduction

The goal of challenge area C4 is to develop efficient and scalable learning and reasoning techniques for application problems arising in the context of COIN. Typical reasoning tasks in applications can be recast as combinatorial search and optimization problems, which can then be tackled using constraint-based optimization methods being developed in C4. The application problems of interest include, e.g., learning probabilistic models from data, clustering based on different optimization criteria, as well as fast inference techniques demanded by applications. Thus far, the focus of C4 has been on *exact methods* that aim at finding globally optimal solutions for the problems involved. The results obtained demonstrate the applicability of constraint-based methods to a wide variety of application problems and pinpoint essential features of problems that are crucial for efficiency.

In 2014-2015, the contributions in the area of C4 are manifold. First, there is substantial progress made in the core reasoning techniques and, in particular, the development of new methodology for *answer set programming* (ASP), *Boolean satisfiability checking* (SAT), as well as their extensions. Achievements in these constraint-based paradigms are described in more detail in Sections 4.2 and 4.3, respectively. As regards the main applications in the COIN agenda, the problem of learning *probabilistic graphical models* (PGMs) from data has been addressed extensively, as can be found out in Section 4.4. In addition to COIN-specific applications, also others are emerging due to general applicability of logic-based methods. Further applications are addressed in Section 4.5.

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, constituting the idea of *translation-based* ASP. In 2014-2015, substantial achievements were made on language extensions, applications in knowledge representation, and solver development.

Extended rule types such as *choice*, *cardinality*, and *weight* rules increase the expressive power of ASP. In translation-based ASP, such extensions may have to be translated away and new schemes for the *normalization* of weight rules were developed in [12]. The designs are based on merging and sorting circuits and the number of normal rules required in their normalization is of the order of $n \times (\log_2 k)^2 \times \log_2 W$ where n is the number of literals, k is the bound, and W is the sum of weights assigned to literals. There is ongoing work that extends analogous designs for objective functions used in optimizing variants of ASP. The extended rule types mentioned above are examples of rules involving *aggregate functions* and current implementations rewrite such extensions into simpler forms known as monotone aggregates. In [2], a polynomial, faithful, and modular translation for rewriting common aggregation functions into the simpler form accepted by current solvers. A key (complexity-theoretic) observation is that sometimes proper *disjunctive* rules have to be introduced by such a transformation. In previous research, the ASP paradigm has also been extended by other kinds of constraints, not confining to the rule-based syntax of ASP and potentially involving special variable domains. For instance, the ASPARTAME system [4] enables the use of traditional constraints as understood of constraint programming (CP) within answer set programs. In ASP *modulo acyclicity* approach [13], an answer-set

program is extended by a dynamically varying directed graph and an implicit *acyclicity* constraint is imposed on that graph. Such an extension and its integration in the state-of-the-art ASP solver CLASP provides an alternative way of implementing the unfounded set check used to guarantee the minimality of answer sets.

Many knowledge representation tasks involve trees or similar tree-like structures as abstract datatypes. They are prevalent in may COIN applications and probabilistic graphical models addressed in Section 4.4 form a central example. A systematic study of acyclicity properties is carried out in [28] where the representations of (directed) trees, directed acyclic graphs, and chordal graphs are considered. Several alternative encodings of these properties are developed and experimentally evaluated. The most compact encodings feasible in ASP are linear which contrasts with even exponential encodings found in literature. As further elaborated in follow-up work [30], these encodings can be readily exploited in other logical formalisms via existing translations from ASP. There is a trend in ASP towards modeling domains where the objects under consideration are chancing dynamically (e.g., stream-based reasoning, online reasoning). This sets new requirements from the knowledge representation perspective and, in [26], an approach to successively incorporating new objects in answer-set programs was introduced. The approach enables the incremental addition of new pieces of information in a modular fashion.

The development of contemporary solver technology in COIN is fortified by active cooperation with the developers of state-of-the-art ASP tools, such as GRINGO [25] and CLASP [31] that together form the ASP system CLINGO [32]. As a recent development, the support for *multi-shot solving* [33] has emerged in response to changing information and the dynamic aspects discussed above. In this approach, a reactive procedure may loop on the solving phase while acquiring changes to the problem specification. In addition, an enhanced support for online ASP was established in terms of agent programming [17]. Finally, international solver competitions foster the development of solver technology by providing benchmark problems for the community and regular points of performance evaluation. The COIN researchers have contributed in the organization of the 5th and 6th ASP competitions [16, 34] and submitted a variety of solvers for evaluation.

4.3 Contributions to SAT Methodology

Boolean satisfiability (SAT) solvers provide an efficient implementation of classical propositional logic, assuming the conjunctive normal form (CNF) for propositional formulas in the basic setting. The goal of this research is to develop fundamental techniques for satisfiability checking and to integrate them in state-of-the-art SAT solver technology, hence boosting the efficiency of reasoning. The research was successful on a wide front, spanning from new solving techniques to novel extensions of the SAT paradigm.

Work on solving techniques produced novel preprocessing and inprocessing techniques, and they were also extended beyond NP to the case of Quantified Boolean satisfiability [37] with related theoretical analysis [41, 47]. Major contributions to central international solver competitions were made and reported in [5, 3].

Work on practical state-of-the-art solver technology for maximum satisfiability (MaxSAT), the optimization variant of SAT, reached a level of maturity by the development of the LMHS MaxSAT solver [52]. The solver implements a SAT-ILP hybrid implicit hitting set approach to MaxSAT, and it gained top positions at the 2015 MaxSAT evaluation.

In connection to MaxSAT solver techniques, novel preprocessing techniques, with the promise of speeding up state-of-the-art MaxSAT solvers in general, were developed [10, 11]. Furthermore, a general view to the underlying principles instantiated for MaxSAT in the LMHS solver, motivated by a wide range of Beyond-NP applications, was studied [54], providing a state-of-the-art approach to propositional abduction.

As regards extensions of SAT, the incorporation of specialized acyclicity constraints into SAT was considered in [29]. The resulting extension, called SAT *modulo acyclicity*, is analogous to the one described in connection to ASP (cf. Section 4.2). On the technical side, a constraint propagator for the acyclicity constraint was developed and incorporated in off-the-shelf SAT solvers MINISAT and GLUCOSE. The propagator sanctions stronger inferences compared to certain encodings of the acyclicity constraint in pure SAT. Also, SAT modulo acyclicity offers a viable way to implement ASP via translations. The linear embedding of answer-set programs devised in [27] forms a new basis for translation-based ASP, building on the idea of *cross translation* where the actual output format and the back-end solver to be used are decided in the last phase of translation. A number of translation-based solvers (the LP2ACYC family) participated in the 6th ASP competition, truly challenging native ASP solvers in performance.

4.4 Probabilistic Graphical Models

Work on exact approaches to (optimal) learning of and inference in probabilistic graphical models expanded in scope within 2014–2015. Currently the best performing known approach to learning guaranteed optimal bounded-treewidth Bayesian networks, employing MaxSAT solvers, was developed [9]. In a series of articles [21, 22, 45], new search heuristics and pruning techniques for the A*-based URLEARNING system for optimal Bayesian network structure learning (BNSL) were developed, currently constituting one of the best performing systems for the problem. MaxSAT-based search for cutting planes was developed [53] and shown to be competitive with the ILP-based cutting planes implemented in the GOBNILP system for BNSL. A portfolio approach to BNSL was also developed [44], constituting the most effective exact BNSL approach today, by employing machine learning–based portfolio construction over parameterizations of URLEARNING and GOBNILP. Furthermore, a principled empirical study was performed and reported in [43], providing clear evidence that developing exact approaches to BNSL is important in light of the quality of the learned networks. The case of Markov networks and the respective structure learning problem (MNSL) were considered in [40]. The best performance was obtained using the state-of-the-art ASP solver CLASP and a compact ASP encoding which essentially formalizes perfect elimination orderings of chordal graphs.

Going beyond BNSL/MNSL, novel approaches to wider PGM model classes were developed based on constraint optimization solvers, including a most general approach to cyclic causal structure discovery with latents [38] and a first approach to learning guaranteed-optimal chain graphs [55]. Building on the exact approach to causal structure discovery, the gap between causal structure discovery and the *do*-calculus was bridged in [39] by a method for the identification of causal effects on the basis of arbitrary (equivalence) classes of semi-Markovian causal models, using a general logical representation of the equivalence class of graphs obtained from a causal structure discovery algorithm, the properties of which can then be queried by procedures implementing the *do*-calculus inference for causal effects. By linking the constraint optimization based work to structure discovery, a benchmark

library of structure discovery problems for the constraint solver community was made available [6].

4.5 Further Applications

The development constraint-based search and optimization techniques has led to a wide variety of applications, to be summarized below. Applications in *argumentation* and *temporal planning* are more extensive as detailed in Sections 4.5.1 and 4.5.2, respectively.

In 2014-2015, ASP technology was applied in a number of contexts to solve real-life problems. In phylogenetic inference about the origin of species, one central problem is the construction of a supertree out of several phylogenetic trees with possibly conflicting information. In [42], the supertree construction problem was formalized as two ASP encodings, one based on *quartets* and the other on direct *projections*. The encodings were used to compute a genus-level supertree for the family of cats (Felidae). The quality of the supertrees obtained was comparable to the one previously found using heuristic matrix representation with parsimony (MRP) method. Various configuration problems form a traditional application area for ASP. An abstract combination of Siemens product configuration problems was addressed in [35]. It is shown how the performance of CLASP can be improved by incorporating domain-specific heuristic information.

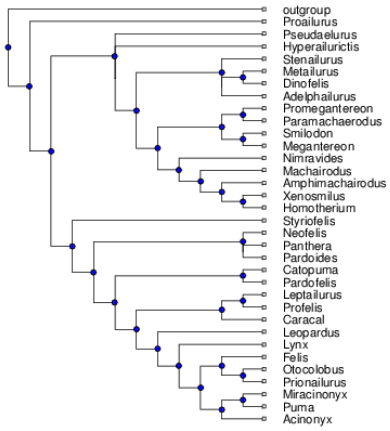


Figure 4.1: Genus Level Supertree for the Felidae Dataset

Yet another task well-suited for ASP is the shift design problem [1]. The goal of the problem being formalized is to align a minimum number of shifts in order to meet required numbers of employees so that over and understaffing is minimized.

In connection to data analysis tasks, MaxSAT-based application studies were reported on various computationally hard problems, including exact treewidth [7], constrained correlation clustering [8], and neighborhood-embedding information visualization [15]. Further studies touched aspects of synthesizing distributed algorithms with SAT-based techniques [18], complexity analysis of various inconsistency measures [56], circuit complexity [23], and connections of modal logics and distributed computing [36].

4.5.1 Abstract Argumentation

In terms of other domain-specific work one strong application focus was on developing and analyzing SAT-based techniques for NP and beyond-NP reasoning problems arising from the study of computational aspects of argumentation — an important area of modern AI research. The SAT-based counterexample-guided abstraction refinement (CEGAR) approach, implemented in the CEGARTIX system [19, 20] for credulous and skeptical acceptance problems over argumentation frameworks. The system gained top positions in the First International Competition on Computation Models for Argumentation (ICCMA'15).

Further advances for these problems were made via developing novel ASP encodings [24] as well as through a study of the fundamental inference rules employed by different argumentation systems [14]. More recently, computational problems to argumentation dynamics were studied [57, 46], providing a first system of its generality for the so-called extension enforcement and status enforcement problems (of NP and beyond-NP complexity), with strong links to belief revision.

4.5.2 Temporal Planning

An important class of sequential decision making problems is based on actions and tasks with a metric duration which can temporally overlap in different ways. An action is executable when its preconditions are true, and the resources it needs are available. When an action is taken, it allocates the required resources, and causes changes to the values of state variables at the specified time points. A plan is a schedule of executable actions that, starting from a specified initial state, reaches a goal state. This type of planning is known as *temporal planning*. In this research, a number of important aspects of solving temporal planning problems with constraint-based methods have been investigated.

In temporal planning, similarly to other reachability problems for timed systems, pruning of the search space with *invariants*, facts that are known to hold in all reachable states, is often critical to the performance of search algorithms. In [48], a general and efficient algorithm for finding a wide class of invariants for temporal planning problems is devised for the first time. Earlier algorithms limit to narrower classes of invariants, typically at-most-one invariants, and are defined only for syntactically very limited temporal modeling languages. The current work lifts both restrictions, and achieves a very high degree of generality both in terms of modeling languages and classes of invariants found.

Further, more efficient translations of temporal planning into constraint-based models were investigated. It was shown that some of the best known existing modeling languages are incompatible with efficient constraint-based models, due to a unnecessarily low-level mechanism for representing resource allocation and deallocation [51], in analogy to other low-level aspects of modeling languages [50]. Then, it was shown in [49], how expressive resource-aware modeling languages for temporal planning can be efficiently translated into the SAT *modulo theories* (SMT) framework, and solved with state-of-the-art SMT solvers. A main innovation in the work is the development of effective methods for discretizing temporal planning models, by replacing real-valued time by integer-valued time, which in many cases leads to constraint models that are far more efficiently solvable than corresponding undiscretized models.

References

[1] Michael Abseher, Martin Gebser, Nysret Musliu, Torsten Schaub, and Stefan Woltran. Shift Design with Answer Set Programming. In *Proceedings of Logic Programming and Non-monotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 32–39. Springer, September 2015.

[2] Mario Alviano, Wolfgang Faber, and Martin Gebser. Rewriting recursive aggregates in answer set programming: back to monotonicity. *Theory and Practice of Logic Programming*, 15(4-5):559–573, 2015.

[3] Adrian Balint, Anton Belov, Matti Järvisalo, and Carsten Sinz. Overview and analysis of the SAT challenge 2012 solver competition. *Artificial Intelligence*, 223:120–155, 2015.

[4] Mutsunori Banbara, Martin Gebser, Katsumi Inoue, Max Ostrowski, Andrea Peano, Torsten Schaub, Takehide Soh, Naoyuki Tamura, and Matthias Weise. ASPARTAME: Solving Constraint Satisfaction Problems with Answer Set Programming. In *Proceedings of Logic Programming and Non-monotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 112–126. Springer, September 2015.

[5] Anton Belov, Daniel Diepold, Marijn Heule, and Matti Järvisalo, editors. *Proceedings of SAT Competition 2014: Solver and Benchmark Descriptions*, volume B-2014-2 of *Department of Computer Science Series of Publications B*. University of Helsinki, 2014. ISBN 978-951-51-0043-6.

[6] Jeremias Berg, Antti Hyttinen, and Matti Järvisalo. Applications of MaxSAT in data analysis. In *Proceedings of the 6th Pragmatics of SAT Workshop*, 2015.

[7] Jeremias Berg and Matti Järvisalo. SAT-based approaches to treewidth computation: An evaluation. In *Proceedings of the 2014 IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI 2014)*, pages 328–335. IEEE Computer Society, 2014.

[8] Jeremias Berg and Matti Järvisalo. Cost-optimal constrained correlation clustering via weighted partial maximum satisfiability. *Artificial Intelligence*, to appear (2015).

[9] 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)*, volume 33 of *JMLR Workshop and Conference Proceedings*, pages 86–95. JMLR, 2014.

[10] Jeremias Berg, Paul Saikko, and Matti Järvisalo. Improving the effectiveness of SAT-based preprocessing for MaxSAT. In Qiang Yang and Michael Wooldridge, editors, *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 239–245. AAAI Press, 2015.

[11] Jeremias Berg, Paul Saikko, and Matti Järvisalo. Re-using auxiliary variables for MaxSAT preprocessing. In *Proceedings of the 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI 2015)*. IEEE Computer Society, 2015.

[12] Jori Bomanson, Martin Gebser, and Tomi Janhunen. Improving the normalization of weight rules in answer set programs. In Eduardo Fermé and João Leite, editors, *Logics in Artificial Intelligence*, volume 8761 of *Lecture Notes in Artificial Intelligence*, pages 166–180. Springer, September 2014.

[13] Jori Bomanson, Martin Gebser, Tomi Janhunen, Benjamin Kaufmann, and Torsten Schaub. Answer Set Programming Modulo Acyclicity. In *Proceedings of Logic Programming and Non-monotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 143–150. Springer, September 2015.

[14] Rémi Brochenin, Thomas Linsbichler, Marco Maratea, Johannes P. Wallner, and Stefan Woltran. Abstract solvers for Dung's argumentation frameworks. In Elizabeth

Black, Sanjay Modgil, and Nir Oren, editors, *Proceedings of the 3rd Workshop on Theory and Applications of Formal Argumentation (TAFA 2015), revised selected papers*, volume 9524 of *Lecture Notes in Computer Science*, pages 40–58. Springer, 2015.

[15] Kerstin Bunte, Matti Järvisalo, Jeremias Berg, Petri Myllymäki, Jaakko Peltonen, and Samuel Kaski. Optimal neighborhood preserving visualization by maximum satisfiability. In Carla E. Brodley and Peter Stone, editors, *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*, pages 1694–1700. AAAI Press, 2014.

[16] Francesco Calimeri, Martin Gebser, Marco Maratea, and Francesco Ricca. The design of the fifth answer set programming competition. In *Technical Communications of the 30th International Conference on Logic Programming, ICLP 2014, 19-22 July, Vienna, Austria*, page Online Supplement, 2014.

[17] Timothy Cerexhe, Martin Gebser, and Michael Thielscher. Online agent logic programming with oClingo. In *Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2014)*, volume 8862 of *Lecture Notes in Artificial Intelligence*, pages 945–957. Springer-Verlag, 2014.

[18] Danny Dolev, Keijo Heljanko, Matti Järvisalo, Janne H. Korhonen, Christoph Lenzen, Joel Rybicki, Jukka Suomela, and Siert Wieringa. Synchronous counting and computational algorithm design. *Journal of Computer and System Sciences*, 82(2):310–332, 2016.

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

[20] Wolfgang Dvořák, Matti Järvisalo, Johannes Peter Wallner, and Stefan Woltran. Complexity-sensitive decision procedures for abstract argumentation (extended abstract). In Qiang Yang and Michael Wooldridge, editors, *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 4073–4077. AAAI Press, 2015.

[21] Xiannian Fan, Brandon Malone, and Changhe Yuan. Finding optimal Bayesian network structures with constraints learned from data. In Jin Tian and Nevin L. Zhang, editors, *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, pages 200–209. AUAI Press, 2014.

[22] Xiannian Fan, Changhe Yuan, and Brandon Malone. Tightening bounds for Bayesian network structure learning. In Carla E. Brodley and Peter Stone, editors, *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*, pages 2439–2445. AAAI Press, 2014.

[23] Magnus Find, Mika Göös, Matti Järvisalo, Petteri Kaski, Mikko Koivisto, and Janne H. Korhonen. Separating OR, SUM, and XOR circuits. *Journal of Computer and System Sciences*, 82(5):793–801, 2016.

[24] Sarah Alice Gaggl, Norbert Manthey, Alessandro Ronca, Johannes Peter Wallner, and Stefan Woltran. Improved answer-set programming encodings for abstract argumentation. *Theory and Practice of Logic Programming*, 15(4–5):434–448, 2015.

[25] Martin Gebser, Amelia Harrison, Roland Kaminski, Vladimir Lifschitz, and Torsten Schaub. Abstract Gringo. *Theory and Practice of Logic Programming*, 15(4-5):449–463, 2015.

[26] Martin Gebser, Tomi Janhunen, Holger Jost, Roland Kaminski, and Torsten Schaub. ASP Solving for Expanding Universes. In *Proceedings of Logic Programming and Nonmonotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 354–367. Springer, September 2015.

[27] Martin Gebser, Tomi Janhunen, and Jussi Rintanen. Answer set programming as SAT modulo acyclicity. In *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI 2014)*, pages 351–356. IOS Press, 2014.

[28] 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)*, pages 634–637. AAAI Press, 2014.

[29] Martin Gebser, Tomi Janhunen, and Jussi Rintanen. SAT modulo graphs: Acyclicity. In *Proceedings of the 14th European Conference on Logics in Artificial Intelligence (JELIA 2014)*, volume 8761 of *Lecture Notes in Artificial Intelligence*, pages 137–151. Springer-Verlag, 2014.

[30] Martin Gebser, Tomi Janhunen, and Jussi Rintanen. Declarative encodings of acyclicity properties. *Journal of Logic and Computation*, Online access, September 2015.

[31] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, Javier Romero, and Torsten Schaub. Progress in clasp Series 3. In *Proceedings of Logic Programming and Nonmonotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 368–383. Springer, September 2015.

[32] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. Clingo = ASP + control: Preliminary report. In *Technical Communications of the 30th International Conference on Logic Programming, ICLP 2014, 19-22 July, Vienna, Austria*, 2014.

[33] Martin Gebser, Roland Kaminski, Philipp Obermeier, and Torsten Schaub. *Ricochet Robots Reloaded: A Case-Study in Multi-shot ASP Solving*, pages 17–32. Springer International Publishing, Switzerland, 2015.

[34] Martin Gebser, Marco Maratea, and Francesco Ricca. The Design of the Sixth Answer Set Programming Competition. In *Proceedings of Logic Programming and Nonmonotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 531–544. Springer, September 2015.

[35] Martin Gebser, Anna Ryabokon, and Gottfried Schenner. Combining Heuristics for Configuration Problems Using Answer Set Programming. In *Proceedings of Logic Programming and Nonmonotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 384–397. Springer, September 2015.

[36] 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*, 28(1):31–53, 2015.

[37] Marijn Heule, Matti Järvisalo, Florian Lonsing, Martina Seidl, and Armin Biere. Clause elimination for SAT and QSAT. *Journal of Artificial Intelligence Research*, 53:127–168, 2015.

[38] Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In Jin Tian and Nevin L. Zhang, editors, *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, pages 340–349. AUAI Press, 2014.

[39] Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Do-calculus when the true graph is unknown. In Tom Heskes and Marina Meila, editors, *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI 2015)*, pages 395–404. AUAI Press, 2015.

[40] Tomi Janhunen, Martin Gebser, Jussi Rintanen, Henrik Nyman, Johan Pensar, and Jukka Corander. Learning discrete decomposable graphical models via constraint optimization. *Statistics and Computation*, Online access, November 2015.

[41] Matti Järvisalo and Janne H. Korhonen. Conditional lower bounds for failed literals and related techniques. In Uwe Egly and Carsten Sinz, editors, *Proceedings of the 17th International Conference on Theory and Applications of Satisfiability Testing (SAT 2014)*, volume 8561 of *Lecture Notes in Computer Science*, pages 75–84. Springer, 2014.

[42] Laura Koponen, Emilia Oikarinen, Tomi Janhunen, and Laura Säilä. Optimizing phylogenetic supertrees using answer set programming. *Theory and Practice of Logic Programming*, 15(4-5):604–619, 2015.

[43] Brandon Malone, Matti Järvisalo, and Petri Myllymäki. Impact of learning strategies on the quality of Bayesian networks: An empirical evaluation. In Tom Heskes and Marina Meila, editors, *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI 2015)*, pages 362–371. AUAI Press, 2015.

[44] Brandon Malone, Kustaa Kangas, Matti Järvisalo, Mikko Koivisto, and Petri Myllymäki. Predicting the hardness of learning Bayesian networks. In Carla E. Brodley and Peter Stone, editors, *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*, pages 2460–2466. AAAI Press, 2014.

[45] Brandon M. Malone and Changhe Yuan. A depth-first branch and bound algorithm for learning optimal Bayesian networks. In Madalina Croitoru, Sebastian Rudolph, Stefan Woltran, and Christophe Gonzales, editors, *Revised Selected Papers of the Third International Workshop on Graph Structures for Knowledge Representation and Reasoning (GKR 2013)*, volume 8323 of *Lecture Notes in Computer Science*, pages 111–122. Springer, 2014.

[46] Andreas Niskanen, Johannes Peter Wallner, and Matti Järvisalo. Optimal status enforcement in abstract argumentation. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*. AAAI Press, 2016.

[47] Emilia Oikarinen and Matti Järvisalo. Answer set solver backdoors. In Eduardo Ferme and Joao Leite, editors, *Proceedings of the 14th European Conference on Logics in Artificial Intelligence (JELIA 2014)*, Lecture Notes in Artificial Intelligence. Springer, 2014.

[48] Jussi Rintanen. Constraint-based algorithm for computing temporal invariants. In E. Fermé and J. Leite, editors, *Logics in Artificial Intelligence, 14th European Conference, JELIA 2014, September 2014, Proceedings*, volume 8761 of *Lecture Notes in Computer Science*, pages 665–673. Springer-Verlag, 2014. pdf not available.

[49] Jussi Rintanen. Discretization of temporal models with application to planning with SMT. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, pages 3349–3355. AAAI Press, 2015.

[50] Jussi Rintanen. Impact of Modeling Languages on the Theory and Practice in Planning Research. In Blai Bonet and Sven Koenig, editors, *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, pages 4052–4056, Palo Alto, California, January 2015. AAAI Press.

[51] Jussi Rintanen. Models of Action Concurrency in Temporal Planning. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 1659–1665. AAAI Press, July 2015.

[52] Paul Saikko, Jeremias Berg, and Matti Järvisalo. LMHS: A SAT-IP hybrid MaxSAT solver. In Nadia Creignou and Daniel Le Berre, editors, *Proceedings of the 19th International Conference on Theory and Applications of Satisfiability Testing (SAT 2016)*, volume 9710 of *Lecture Notes in Computer Science*, pages 539–546. Springer, 2016.

[53] Paul Saikko, Brandon Malone, and Matti Järvisalo. MaxSAT-based cutting planes for learning graphical models. In Laurent Michel, editor, *Proceedings of the 12th International Conference on Integration of Artificial Intelligence and Operations Research Techniques in Constraint Programming (CPAIOR 2015)*, volume 9075 of *Lecture Notes in Computer Science*, pages 345–354. Springer, 2015.

[54] Paul Saikko, Johannes Peter Wallner, and Matti Järvisalo. Implicit hitting set algorithms for reasoning beyond NP. In Chitta Baral, James P. Delgrande, and Frank Wolter, editors, *Proceedings of the 15th International Conference on Principles of Knowledge Representation and Reasoning (KR 2016)*, pages 104–113. AAAI Press, 2016.

[55] Dag Sonntag, Matti Järvisalo, Jose M. Peña, and Antti Hyttinen. Learning optimal chain graphs with answer set programming. In Tom Heskes and Marina Meila, editors, *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI 2015)*, pages 822–831. AUAI Press, 2015.

[56] Matthias Thimm and Johannes P. Wallner. Some complexity results on inconsistency measurement. In Chitta Baral, James P. Delgrande, and Frank Wolter, editors, *Proceedings of the 15th International Conference on Principles of Knowledge Representation and Reasoning (KR 2016)*, pages 114–124. AAAI Press, 2016.

[57] Johannes Peter Wallner, Andreas Niskanen, and Matti Järvisalo. Complexity results and algorithms for extension enforcement in abstract argumentation. In Dale Schuurmans and Michael Wellman, editors, *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016)*, pages 1088–1094. AAAI Press, 2016.

Chapter 5

F1: Intelligent Information Access

Jorma Laaksonen, Erkki Oja, Samuel Kaski, Petri Myllymäki, Mikko Kurimo, Tapani Raiko, Rao Muhammad Anwer, Pedram Daee, Seppo Enarvi, Manuel Eugster, Patrik Floréen, Dorota Głowacka, Cristina Gonzalez-Caro, Dhananjaya Gowda, Stig-Arne Grönroos, Satoru Ishikawa, Heikki Kallasjoki, Antti Kangasrääsiö, Matti Karppa, Reima Karhila, Sami Keronen, Arto Klami, Markus Koskela, Ana Ramírez López, Yao Lu, Petri Luukkonen, Marcos Luzardo, Andre Mansikkaniemi, Kalle Palomäki, Jaakko Peltonen, Hamed R.-Tavakoli, Ulpu Remes, Teemu Roos, Tuukka Ruotsalo Rakshith Shetty, Mats Sjöberg, Peter Smit, Matti Varjokallio, Ville Viitaniemi

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 to the user, and do the inference on-line, learning relevance from the user's responses. 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.

5.2 Contextual information interfaces

A significant fraction of information searches are motivated by the users task, such as documents that the user is reading or writing. An ideal search engine would be able to use contextual information inferred from the task in order to retrieve useful information. We have proposed a method for detecting implicit search intent using the information available from the task that the user is engaged in [1]. Using the intent model, information relevant to the user's task can be proactively retrieved without user initiation. We utilize an upper confidence bound algorithm, which estimates the intent using a multi-armed bandit model via balancing exploration and exploitation.

We study the method in two experimental setups: large-scale simulations and a user study. The simulations used pre-written articles from several standard data sets. In the user study, the user's task was to write an essay on a given topic. The results of the simulations show that a user's search intent can be inferred from the task context and that the model can retrieve information relevant to the task. The user study demonstrated that a higher task-level recall can be achieved with the proposed method, but with the trade-off of sacrificing precision, the latter being a common characteristic of proactive recommendation systems.

References

[1] Markus Koskela, Petri Luukkonen, Mats Sjöberg, Tuukka Ruotsalo, and Patrik Floréen: Detecting Implicit Search Intent in a Writing Task. Submitted to a conference, 2016.

5.3 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 familiar with the information *a priori*, 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.

We have proposed new methodology for interactive information search, namely *Interactive Intent Modeling* [5], 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, an information access system that couples advanced machine learning techniques for interactive intent modeling with advanced information visualization and interaction to boost exploratory 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 5.1. 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.

[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.

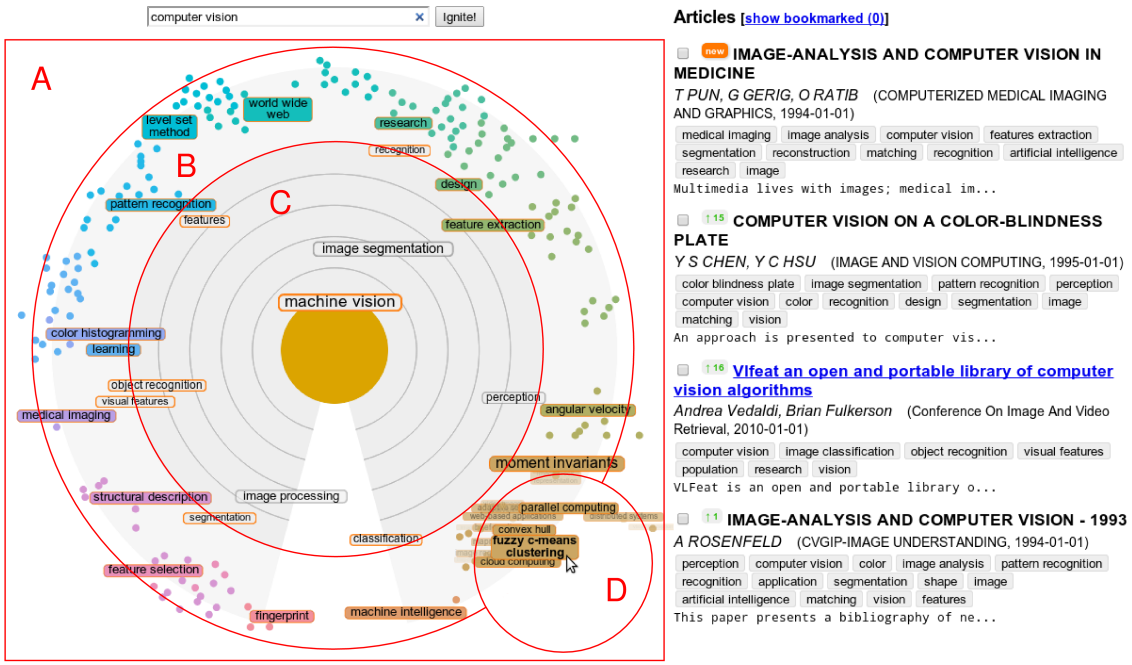


Figure 5.1: A novel information access system based on interactive intent modeling, where users are shown a representation of their estimated present and future intents on an interactive visual interface, and they can give natural feedback on the interface to tune the search towards relevant information. 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). Details of Keywords can be inspected with a fisheye lens (D).

- 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 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 and documents 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.

Based on the foundations of the mentioned system for interactive intent modeling [5], we have proposed and developed several new algorithms and methods that relax the simplifying assumptions of the system. The following, is a summary of the new findings:

Improving controllability and predictability One problem in exploratory search is that the user is often modelled as a passive source of relevance information, instead of an active entity trying to steer the system based on evolving information needs. This may cause the user to feel that the response of the system is inconsistent with her steering. Another problem arises due to the sheer size and complexity of the information space, and hence of the system, as it may be difficult for the user to anticipate the consequences of her actions in this complex environment. These problems can be mitigated by interpreting the user's actions as setting a goal for an optimization problem regarding the system state, instead of passive relevance feedback, and by allowing the user to see the predicted effects of an action before committing to it. In [4], we presented an implementation of these improvements in SciNet system. A user study gave some indication on improvements in task performance, usability, perceived usefulness and user acceptance.

Taking advantage of multiple feedback domains One of the big challenges of exploratory search is that the amount of user feedback is very limited compared to the size of the information space to be explored. To tackle this problem we introduced coupled

multi-armed bandits algorithm [2] that takes into account user feedback on both the retrieved items (documents) and their features (keywords). This new reinforcement learning method employs a probabilistic model of the relationship between the feedback domains to improve the exploratory search. Simulation results and a preliminary user study have shown that the new algorithm improves user satisfaction and quality of retrieved information.

Dealing with concept drift One challenge in intent modeling is that users usually start with considerable uncertainty about their search goals, and so the search intent of the user may be volatile as the user is constantly learning and reformulating her search hypothesis during the search. This may lead to a noticeable concept drift in the relevance feedback given by the user. In [3], We formulated a Bayesian regression model for predicting the accuracy of each individual user feedback and thus find outliers in the feedback data set. To accompany this model, we introduced a timeline interface that visualizes the feedback history to the user and gives her suggestions on which past feedback is likely in need of adjustment. This interface also allows the user to adjust the feedback accuracy inferences made by the model. Simulation experiments demonstrated that the performance of the new user model outperforms a simpler baseline and that the performance approaches that of an oracle, given a small amount of additional user interaction. A user study showed that the proposed modeling technique, combined with the timeline interface, made it easier for the users to notice and correct mistakes in their feedback, resulted in better and more diverse recommendations, allowed users to easier find items they liked, and was more understandable.

Multiple streams of intent modeling We have also introduced a system called IntentStreams which rovides interactive query refinement mechanisms through interactive intent modeling and shows parallel visualization of search streams [1]. The system models each search stream via an intent model allowing rapid user feedback. Streams are shown as columns of search results, with separate interfaces for giving keyword feedback, and interactivity options to transfer keywords across streams. The user interface thus allows swift initiation of alternative and parallel search streams by direct manipulation that does not require typing. A study with 13 participants shows that IntentStreams provides better support for branching behavior compared to a conventional search system.

Transfer of learned user models between systems A user model learned through interactive intent modeling can be useful for systems operating on related domains. We have compared methods of cross-system user model transfer across two large real-life systems [6]: we transfer user models built for information seeking of scientific articles in the SciNet exploratory search system, operating over tens of millions of articles, to perform cold-start recommendation of scientific talks in the CoMeT talk management system, operating over hundreds of talks. Our user study focuses on transfer of novel explicit open user models curated by the user during information seeking. Results show strong improvement in cold-start talk recommendation by transferring open user models, and also reveal why explicit open models work better in cross-domain context than traditional hidden implicit models.

References

[1] Salvatore Andolina, Khalil Klouche, Jaakko Peltonen, Mohammad Hoque, Tuukka Ruotsalo, Diogo Cabral, Arto Klami, Dorota Glowacka, Patrik Floreen, and Giulio Jacucci. IntentStreams: Smart Parallel Search Streams for Branching Exploratory Search. In *Proceedings of ACM IUI 2015, The 20th ACM Conference on Intelligent User Interfaces*, pages 300-305, 2015.

[2] P. Daee, J. Pyykko, D. Głowacka, and S. Kaski. Interactive Intent Modeling from Multiple Feedback Domains. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI)*, pages 71–75, 2016.

[3] A. Kangasraasio, Y. Chen, D. Głowacka, and S. Kaski. IUI'16 Companion. ACM, 2016. Interactive Modeling of Concept Drift and Errors in Relevance Feedback. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP)*, 2016.

[4] A. Kangasraasio, D. Głowacka, and S. Kaski. Improving controllability and predictability of interactive recommendation interfaces for exploratory search. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI)*, pages 247–251, 2015.

[5] T. Ruotsalo, G. Jacucci, P. Myllymäki, and S. Kaski. Interactive intent modeling: Information discovery beyond search. In *Communications of the ACM*, pages 86–92, 2015.

[6] Chirayu Wongchokprasitti, Jaakko Peltonen, Tuukka Ruotsalo, Payel Bandyopadhyay, Giulio Jacucci and Peter Brusilovsky. User Model In a Box: Cross-System User Model Transfer for Resolving Cold Start Problems. In *Proceedings of UMAP'15, The 23rd Conference on User Modelling, Adaptation and Personalization*, pages 289-301, Springer, 2015.

5.4 Biosignal feedback and brain activity

Besides direct interaction patterns, we can exploit various biosignals to learn about the user's attention and relevance of information items. We have strong background in deriving implicit feedback signals from peripheral observations such as eye movements, but the amount of information revealed by such sources is naturally limited. Our current research focuses on more directly estimating the users interests by monitoring the brain activity itself using electroencephalogram (EEG) and magnetoencephalogram (MEG) measurements, which requires advanced machine learning methodologies for extracting the relevant information from noisy measurements.

In [3] we used MEG signals and Gaussian process classifiers for decoding the relevance of images in visual retrieval tasks. We showed that subjective image relevance can be predicted from the brain signal alone and that by fusing the brain signals with eye movements we can further improve the accuracy compared to using either signal alone. While MEG measurements are infeasible for developing practical retrieval tools, these results demonstrate that inferring relevance judgements from brain activity is possible and they help in identifying what kind of activity profiles (temporal and spatial) to look for.

Practical information retrieval systems need to rely on more peripheral measurements, and we have shown that they are also sufficient for detecting relevance judgements. In [1] we used full-scalp EEG to infer the relevance of textual information using a Bayesian multiple kernel learning classifier, demonstrating for the first time that the term relevance can be predicted directly from EEG with high precision. In [2] we showed that textual relevance can be predicted within 6 seconds of the relevance judgement also based on low cost unobtrusive sensors that measure skin conductance and brow muscle activity, but naturally with limited accuracy compared to directly measuring the brain activity. Combining these two sources can lead to further improvements, and these results point towards practical retrieval tools inferring relevance judgements from brain activity and other biosignals.

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. In *Proceedings of thr 37th ACM SIGIR Conference on Research & Development in Information Retrieval*. pages 425–434, 2014. ACM.

[2] Oswald Barral, Manuel J.A. Eugster, Tuukka Ruotsalo, Michiel Sovijärvi-Spapé, Ilkka Kosunen, Niklas Ravaja, Samuel Kaski, and Giulio Jacucci. Exploring peripheral physiology as a predictor of perceived relevance in information retrieval. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 389-399, 2015. ACM.

[3] Jukka-Pekka Kauppi, Melih Kandemir, Veli-Matti Saarinen, Lotta Hirvenkari, Lauri Parkkonen, Arto Klami, Riitta Hari, and Samuel Kaski. Towards brain-activity-controlled information retrieval: Decoding image relevance from MEG signals. *NeuroImage*, 112:288–298, 2015.

5.5 Visual recognition of human actions

Analysing human actions in images and 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 visual 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 recognition in still images Person description in still images is a challenging problem in computer vision. We investigated in [1] two major aspects of person description: 1) gender and 2) action recognition in still images. Most state-of-the-art approaches for gender and action recognition rely on the description of a single body part, such as face or full-body. However, relying on a single body part is suboptimal due to significant variations in scale, viewpoint, and pose in real-world images, and we proposed a semantic pyramid approach for pose normalization. The experiments clearly demonstrated that the

proposed approach, despite its simplicity, outperforms state-of-the-art methods for gender and action recognition.

In a follow-up study [2], we proposed the use of deep semantic pyramids for human attributes and action recognition. In the context of object recognition and detection, convolutional neural networks (CNNs) or deep features have shown to improve the performance over the conventional shallow features. Our method works by constructing spatial pyramids based on CNNs of different part locations. These pyramids are then combined to obtain a single semantic representation. Again, we were able to show a gain in performance compared to best methods in literature.

Gesture recognition in RGB-D videos Action and gesture recognition from motion capture and RGB-D camera sequences has recently emerged as a renowned and challenging research topic. The general aim is to provide automated analysis of various kinds of human activities. Starting from either video, motion capture, depth data, or some combination of these, many action and gesture recognition methods have been developed for various applications, such as surveillance, human-computer interfaces, gaming, and analysis of sign language. We have proposed a method based on classifying a motion initially on the frame level and then making the final classification decision considering the whole sequence, instead of building a feature representation on the sequence level [3]. Each frame-level feature is classified with an extreme learning machine (ELM) classifier. The method retains high recognition accuracy for up to 40 actions, and is computationally light and can thus be used in online applications.

Video analysis of sign language 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. 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. 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. Also non-manual aspects of the videos would often be of importance.

It is inconceivable to try to understand sign language without recognising also the handshapes. In our study [4], we studied which visual feature extraction methods would be the most useful for the recognition of the handshape and detailed statistical descriptions of the texture within the skin blob areas seemed to perform best. After our earlier studies on head pose estimation, we proposed and evaluated methods for estimating the state of facial elements—eyes, eyebrows and mouth—in the context of sign language in [5]. The applicability of such 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.

We have made publicly available our developed methods in the SLMotion video analysis toolkit [6] and our used video data in the S-pot benchmark dataset [7].

Gesture passwords In addition to usability aspects, human-computer interaction research also has implications on computer security through, e.g., the study of authentication

methods. Building on our earlier work on the maximum achievable information-theoretic capacity of human motion, and thus also human-computer interfaces, we propose a generic index that characterizes the properties of gesture passwords [8]. This research has informed the development of secure, memorable, and easy-to-use authentication approaches especially for mobile devices, see [9].

References

[1] Fahad Shahbaz Khan, Joost van de Weijer, Rao Muhammad Anwer, Michael Felsberg, and Carlo Gatta. Semantic pyramids for gender and action recognition. IEEE Transactions on Image Processing 23(8):3633–3645, 2014.

[2] Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost van de Weijer, Michael Felsberg, and Jorma Laaksonen. Deep Semantic Pyramids for Human Attributes and Action Recognition. In *Proceedings of 19th Scandinavian Conference on Image Analysis (SCIA)*, 2015.

[3] Xi Chen and Markus Koskela. Skeleton-based action recognition with extreme learning machines. Neurocomputing 149:387–396, 2015.

[4] 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.

[5] 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.

[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] 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.

[9] Free-form gesture authentication in the wild, Yang, Y., Clark, G., Lindqvist, J., Oulasvirta, A., Proc. CHI 16, pp. 3722-3735, ACM Press.

5.6 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.

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 [1] 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 [2]. The speaker-adaptive system is also very robust against noise and reverberation in the adaptation data [3,12]. 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 [4] we present an in-depth study of estimators of observation uncertainty to improve the performance of a noisy speech recognition, augmented by new uncertainty heuristics and a channel normalization step to improve the match between signal and the speech and noise dictionaries. The recognition performance is also evaluated using real noisy speech and compared against the SPLICE feature enhancement method.

The missing data methods, which draw inspiration from the human auditory system [5], 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 5.2.

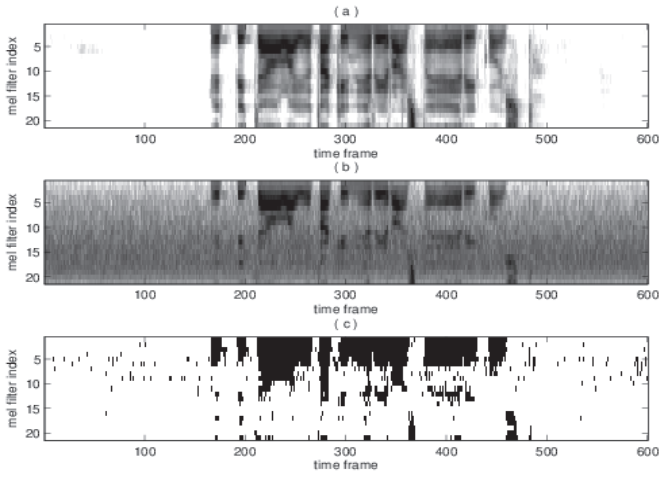


Figure 5.2: Logarithmic Mel spectrogram of (a) an utterance recorded in a quiet environment and (b) the same utterance corrupted with additive noise. The noise mask (c) constructed for the noisy speech signal indicates the speech dominated regions in black and the noise dominated regions in white.

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 [10]. 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.

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 [6] 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 [7]. 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 [8]. 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 [7,8].

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 [9,11]. 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.

References

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

[2] Junichi Yamagishi, Bela Usabaev, Simon King, Oliver Watts, John Dines, Jilei Tian, Rile Hu, Yong Guan, Keiichiro Oura, Keiichi Tokuda, Reima Karhila, and Mikko Kurimo. 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.

[3] Karhila, Reima; Remes, Ulpu; Kurimo, Mikko. Noise in HMM-Based Speech Synthesis Adaptation: Analysis, Evaluation Methods and Experiments. IEEE Journal of Selected Topics in Signal Processing, 8(2): 285-295, 2014

[4] Kallasjoki, Heikki; Gemmeke, Jort; Palomaki, Kalle. Estimating Uncertainty to Improve Exemplar-Based Feature Enhancement for Noise Robust Speech Recognition. IEEE transactions on audio speech and language processing , 2014. Vol. 22, nro 2, pp. 368-380.

[5] Remes U., Ramírez López A., Juvela L., Palomäki K., Brown G. J., Alku P., Kurimo M. (2016) Comparing human and automatic speech recognition in a perceptual restoration experiment, Computer Speech and Language, 35, 14-31, January 2016.

[6] Remes Ulpu; Ramírez López Ana; Palomäki Kalle; Kurimo Mikko. Bounded conditional mean imputation with observation uncertainties and acoustic model adaptation. IEEE Transactions on Audio, Speech and Language Processing, 23(7): 1198-1208, 2015.

[7] Kurimo, Mikko, Enarvi, Seppo, Tilk, Ottokar, Varjokallio, Matti, Mansikkaniemi, Andre, Alumäe, Tanel. Modeling under-resourced languages for speech recognition. Language Resources and Evaluation, pp.1-27, 2016.

[8] Mansikkaniemi, André; Kurimo, Mikko. Adaptation of morph-based speech recognition for foreign names and acronyms. IEEE Transactions on Audio, Speech and Language Processing, 23(5): 941-950, May 2015.

[9] Olli-Philippe Lautenbacher, Liisa Tiittula, Maija Hirvonen, Jorma Laaksonen, Mikko Kurimo. Towards Reliable Automatic Multimodal Content Analysis. 2015 Conference on Empirical Methods for Natural Language Processing, Fourth Workshop on Vision and Language. Lisbon, Portugal 2015, Association for Computational Linguistics, 6-7, 2015.

[10] Ramírez López A., Nobutaka O., Remes U., Palomäki K. and Kurimo M. Designing multichannel source separation based on single-channel source separation, International Conference on Acoustics Speech and Signal Proc. (ICASSP) 2015.

[11] Amid, Ehsan; Mesaros, Annamaria; Palomäki, Kalle J.; Laaksonen, Jorma; Kurimo, Mikko. Unsupervised Feature Extraction For Multimedia Event Detection And Ranking Using Audio Content. IEEE International conference on acoustics, speech and signal processing, ICASSP, May 4-9, Florence, Italy, 2014.

[12] Gowda, Dhananjaya; Kallasjoki, Heikki; Karhila, Reima; Contan, Cristian; Palomäki, Kalle; Giurgiu, Mircea; Kurimo, Mikko. On the role of missing data imputation and NMF feature enhancement in building synthetic voices using reverberant speech. Interspeech-2014, Singapore, September 14-18, 2014. Singapore 2014, ISCA, 2947-2951.

5.7 Video content analysis for intelligent access

Deep Convolutional Neural Network features A recent major development in image and video content classification has been the use of deep convolutional neural networks (CNNs), with excellent results. It has been observed that CNNs trained with one visual dataset can function as highly discriminative features even for considerably different data domains and tasks. One can therefore employ CNNs trained with external data as universal visual feature extractors in a standard concept detection framework. In 2014, our PicSOM image and video analysis system included a total of 24 CNN features extracted with four different CNN networks [1]. Our participation in NIST's annual TRECVID video retrieval evaluation was very successful, and of the total 75 submissions in the Semantic Indexing task, only the MediaMill group of the University of Amsterdam submitted runs that were superior to the two best PicSOM runs in their MXIAP results [2].

Affective content analysis of movies 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. In [3] we performed a set of experiments to predict the affective content for 14 movie clips, taken from popular mainstream movies made between 1955 and 2009 encompassing several genres. 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.

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.

In 2014 and 2015 we organized the MediaEval Affect Task together with an international team of researchers. The task had two tracks: violent scene detection and induced affect detection. The benchmark challenges participants to develop algorithms for finding the relevant scenes in movies provided by the organisers. The task was a great success which engaged the multimedia community, and having a growing participation over the years. The task ended in 2015, but spawned two new continuation tasks for 2016: emotional impact and interestingness detection.

Large-Scale Movie Description Challenge The problem of describing videos using natural language has garnered a lot of interest in the last year, after the great progress recently made in automatic image captioning. This development has been partly driven also by the availability of large public datasets of images and videos with human-annotated captions describing them. A popular recipe for solving the image and video captioning problem has been to use an encoder–decoder model, where the encoder model produces a feature vector representation of the visual input, and the decoder model, usually a Long-Short Term Memory recurrent network model, takes the feature vector as the input and generates a caption as the output. We employed this architecture when participating in

The Large Scale Movie Description Challenge organized in the ICCV 2015 Workshop on describing and understanding video [4]. Among all participants of the challenge, our submission was ranked the best by a human evalution that assessed the correctness, grammar, relevance and helpfulness of the generated captions.

References

[1] Markus Koskela and Jorma Laaksonen. Convolutional Network Features for Scene Recognition. In *Proc. 22nd ACM International Conference on Multimedia*, 2014.

[2] Ville Viitaniemi, Mats Sjöberg, Markus Koskela, Satoru Ishikawa, and Jorma Laaksonen. Advances in Visual Concept Detection: Ten Years of TRECVID. In *Advances in Independent Component Analysis and Learning Machines*, Academic Press, 2015.

[3] Jussi Tarvainen, Mats Sjöberg, Stina Westman, Jorma Laaksonen, and Pirkko Oittinen. Content-Based Prediction of Movie Style, Aesthetics, and Affect: Data Set and Baseline Experiments. IEEE Transactions on Multimedia 16(8):2085–2098, 2014.

[4] Rakshith Shetty and Jorma Laaksonen. Video captioning with recurrent networks based on frame- and video-level features and visual content classification. In *ICCV Workshop on Describing and Understanding Video & The Large Scale Movie Description Challenge (LSMDC).*

5.8 Deep neural network research

In 2014, *NADE-k* [1] was published. This paper extended neural autoregressive density estimator (NADE) with multiple step inference scheme. Such an extension leads to better training performance of NADE. Later in 2014, *Linear State-Space Model with Time-Varying Dynamics* [2] was published. This paper introduced a linear state-space model with time varying dynamics. The model forms the dynamics as a linear combination and the changes can be smooth and more continuous. In 2015, *Techniques for Learning Binary Stochastic Feedforward Neural Networks* [3] was published. Binary stochastic feedforward neural networks have several theoretical appeals compared to continuous valued neural networks. This paper proposed two new estimators for the training the neural networks. Later in 2015, *Bidirectional Recurrent Neural Networks as Generative Models* [4] was published. This paper proposed two probabilistic interpretations of bidirectional RNNs that can be used to reconstruct missing gaps in high-dimensional time series efficiently. At the same time, *Semi-Supervised Learning with Ladder Networks* [5] was published. This paper combined supervised learning with unsupervised learning in deep neural networks. The proposed model is trained to simultaneously minimize the sum of supervised and unsupervised cost functions by backpropagation, avoiding the need for layer-wise pre-training. The resulting model reaches state-of-the-art performance on two common datasets.

References

[1] T. Raiko, Y. Li, K. Cho, and Y. Bengio. Iterative neural autoregressive distribution estimator nade-k. Advances in Neural Information Processing Systems (NIPS), 2014.

[2] J. Luttinen, T. Raiko, and A. Ilin. Linear state-space model with time-varying dynamics. Machine Learning and Knowledge Discovery in Databases (ECML), 2014.

[3] T. Raiko, M. Berglund, G. Alain, and L. Dinh. Techniques for Learning Binary Stochastic Feedforward Neural Networks. International Conference on Learning Representations (ICLR), 2015.

[4] M. Berglund, T. Raiko, M. Honkala, L. Kärkkäinen, A. Vetek, and J. Karhunen. Bidirectional Recurrent Neural Networks as Generative Models. Advances in Neural Information Processing Systems (NIPS), 2015.

[5] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, and T. Raiko. Semi-Supervised Learning with Ladder Networks. Advances in Neural Information Processing Systems (NIPS), 2015.

Chapter 6

F2: Computational Molecular Biology and Medicine

Erik Aurell, Jukka Corander, Samuel Kaski, Antti Honkela, Marcin Skwark, Santeri Puranen, Maiju Pesonen, Yingying Xu, Lu Cheng, Michael Gutmann, Väinö Jääskinen, Luca Martino, Pekka Marttinen, Elina Numminen, Ville Parkkinen, Jukka Siren, Lu Wei, Jie Xiong, Topa H, Brandon Malone, Pekka Marttinen, Niko Välimäki

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. Efforts to bring the group together have been documented in an earlier report. The current reporting period has seen an increased collaboration in the group, as will be documented below. For the current reporting period Marcin Skwark acted as coordinating postdoc from August 2013 until July 2015 after which this function was taken over by the co-PIs acting jointly.

6.2 Protein structure prediction by direct coupling analysis

In a previous report we presented the direct-coupling analysis approach to the prediction of amino acid contacts in a protein structure from many homologous protein sequences. This approach amounts to learning a model in an exponential family over categorical variables (the amino acids), with linear and quadratic terms. By analogy with statistical physics such models are usually referred to as "Potts models"; they differ from the more commonly used "Ising models" in that the variables take values in a discrete set but are not Boolean and in the greater number of interaction parameters. During the reporting period in question we have published a new version of our pseudo-maximum-likelihood-based mathod to plmDCA which is many times faster than the previous method by using a different output routine [1] and one of the largest competitive evaluations to date [2]. An important conceptual point is why these methods work so well at all; on this we recently published (by invitation) a polemic [3].

References

[1] Magnus Ekeberg, Tuomo Hartonen, Erik Aurell Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences Journal of Computational Physics 276 341-356 (2014)

[2] Christoph Feinauer, Marcin J. Skwark, Andrea Pagnani, Erik Aurell Improving contact prediction along three dimensions PLoS Comp Biology (2014) e1003847.

[3] Erik Aurell The Maximum Entropy Fallacy Redux? PLOS Comp. Bio.,1004777 (2016)

[4] MJ Skwark, D Raimondi, M Michel, A Elofsson Improved contact predictions using the recognition of protein like contact patterns PLoS Comput Biol 10 (11), e1003889 24, 2014

6.3 Global epistatic analysis

Aurell and Corander have generalized the methodology of the direct coupling analysis to analyse the co-variation over whole bacterial genomes [1]. For this effort we used a unique data set of more than 3000 full-genome sequences of *Staphylococcus pneumoniae* (the pneumococcus) from patient samples, published by Corander and co-workers from Sanger

Institute (Hinxton, UK) and Imperial College [2]. The main results of [1], paper under review but which have been presented at several international conferences, are that it is possible to find the main known players of antibiotic resistance in the pneumococcus (three variants of the family of penicillin-binding proteins (PBPs)). This is in itself interesting as these key enzymes, which are known to be the target of penecillin and similar drugs, have not previously been identified from large-scale sequencing data. It is also of interest that the analysis allow to predict/identify which pairs of PBPs contribute synergetically to bacterial fitness (epistasis) and moreover which parts of these proteins do so. Furthermore it is possible to identify epistatic effects involving the PBPs and other pneumococcus genes. These results are under further study with our collaborators at Sanger and Imperial. We have also launched an effort to repeat the analysis for several similar data sets which are coming on-line, results which will be reported at a later date.

References

[1] Marcin J. Skwark, Nicholas J Croucher, Santeri Puranen, Maiju Pesonen, Yingying Xu, Claire Chewapreecha, Paul Turner, Simon R. Harris, Julian Parkhill, Stephen D. Bentley, Erik Aurell, Jukka Corander Whole-genome epistasis analysis reveals co-evolving mechanisms of resistance in the pneumococcus Proceedings of the National Academy of Sciences of the United States of America, PNAS

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.

We have introduced several Bayesian methods for microbial metagenomics analysis, that targets to estimate microbiome composition and compare microbiome variation across samples using either high-throughput 16S rRNA gene sequences or massive screening of all genome components. Our method BeBAC represents the most accurate unsupervised method for analyzing high-throughput 16S data. 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 (2014)

[2] Chewapreecha, C. et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nature Genetics (2014)

6.5 Timing of gene expression

Understanding the precise timing of gene transcription is essential for modelling gene regulation and expression. In [1] we reported results of detailed genome-wide dynamical modelling of transcription and expression, highlighting widespread significant delays in RNA production between the completion of Pol-II elongation and accumulation of mature mRNA. To uncover these delays, we fitted a simple model of transcription to RNA-seq and Pol-II ChIP-seq time course data using Bayesian techniques. Production delays of more than 20 min were observed for 11% of genes. Long delays were more common for genes with short pre-mRNAs. They were also associated with late intron retention in premRNA data, suggesting a link to splicing.

Estimation of transcript isoform expression from RNA-sequencing data is one of the most fundamental problems in modern gene expression analysis. The BitSeqVB algorithm [2] based on fast variational inference of a Bayesian model of RNA sequencing data has been shown to deliver state-of-the-art accuracy in this problem in independent evaluations (e.g. Kanitz et al.,). We have also recently extended the method for identification of bacterial strains from sequencing data [3].

References

[1] Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays Honkela A, Peltonen J, Topa H, Charapitsa I, Matarese F, Grote K, Stunnenberg HG, Reid G, Lawrence ND, Rattray M. Proceedings of the National Academy of Sciences of the United States of America, PNAS. 2015;112(42):13115-20.

[2] Fast and accurate approximate inference of transcript expression from RNA-seq data. Hensman J, Papastamoulis P, Glaus P, Honkela A, Rattray M. Bioinformatics. 2015, 31(24):3881-9.

[3] Bayesian identification of bacterial strains from sequencing data Aravind Sankar, Brandon Malone, Sion Bayliss, Ben Pascoe, Guillaume Mèric, Matthew D. Hitchings, Samuel K. Sheppard, Edward J. Feil, Jukka Corander, Antti Honkela Microbial Genomics

Sequence-level analysis of multiple genomes or metagenomes requires a combination of accurate analysis and highly efficient algorithms. In [1] we developed and applied an efficient distributed string mining method for finding discriminative sequence motifs in large metagenome data sets. The same method has formed the basis for the novel alignment-free method for pan-genome-wide association study in [2].

References

[1] Exploration and retrieval of whole-metagenome sequencing samples. Seth S, Välimäki N, Kaski S, Honkela A. Bioinformatics. 2014, 30(17):2471-9.

[2] Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes John A Lees, Minna Vehkala, Niko Välimäki, Simon R Harris, Claire Chewapreecha, Nicholas J Croucher, Pekka Marttinen, Mark R Davies, Andrew C Steer, Stephen Y C Tong, Antti Honkela, Julian Parkhill, Stephen D Bentley, Jukka Corander

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 precision 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 cells. We have addressed the problem of modeling the relationships between chemical structures of drugs causing specific gene expression responses, by applying the novel multi-view data integration method Group Factor Analysis (GFA, see C2), to find relationships between specific structural and chemical properties of drugs with the genome-wide responses they elicit in multiple cancer types (see Figure 6.1), creating testable predictions [2]. We extended this work further, exploiting the natural tensor structure in the data (drugs, genes, cancers) to explore the genome-wide responses that are shared across cancers or specific to individual types of cancer, and which of them are related to drugs effectiveness [3].

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 being building and testing computational models for in-depth analysis of molecular biology of the disease, and to predict most effective treatments. Hence, at the core of personalized medicine there is a computational problem with two interrelated goals: (1) Given molecular profiles of cells and sensitivity measurements on an a-priori fixed set of

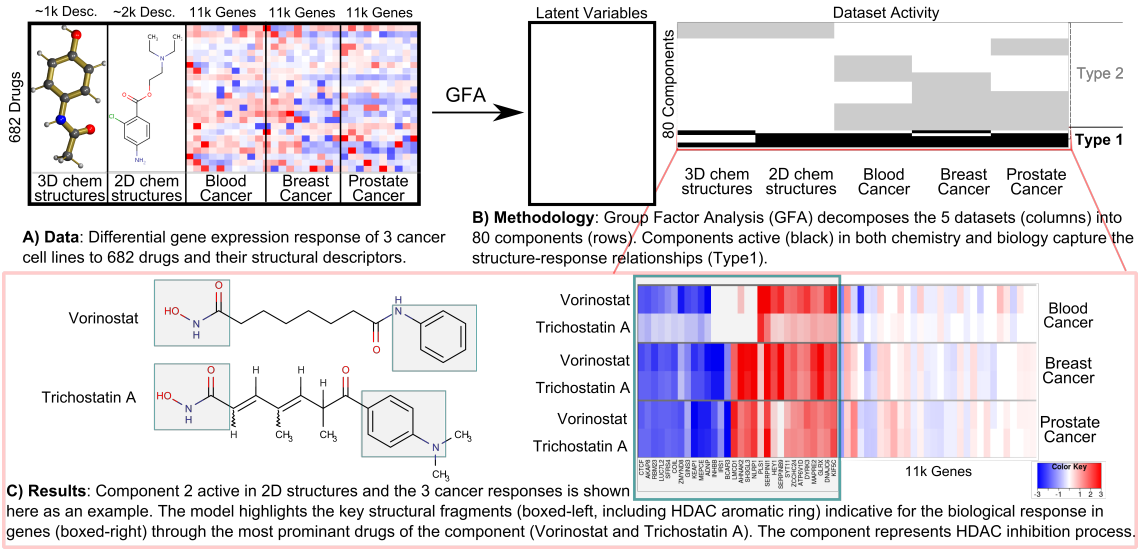


Figure 6.1: Group Factor Analysis (GFA) reveals components of drug effects by simultaneously decomposing data matrices of drug characteristics and genome-wide gene expression induced by the drugs.

drugs, predict sensitivity of a new cell to these drugs. (2) Identify potential biomarkers predictive of drug sensitivity in cancer cells.

We have, in collaboration with the Institute for Molecular Medicine Finland FIMM, developed a novel probabilistic multi-source machine learning method (see Figure 6.2). Our method showed the best predictive performance by outperforming other state-of-art methods proposed by 41 international teams. The specific goal of the crowd-sourced competition was to predict effectiveness of the drugs on new cells based on the genomic and molecular measurements. The key property of our method was to combine multiple sources of information for the cells with the appropriate use of prior biological knowledge [4].

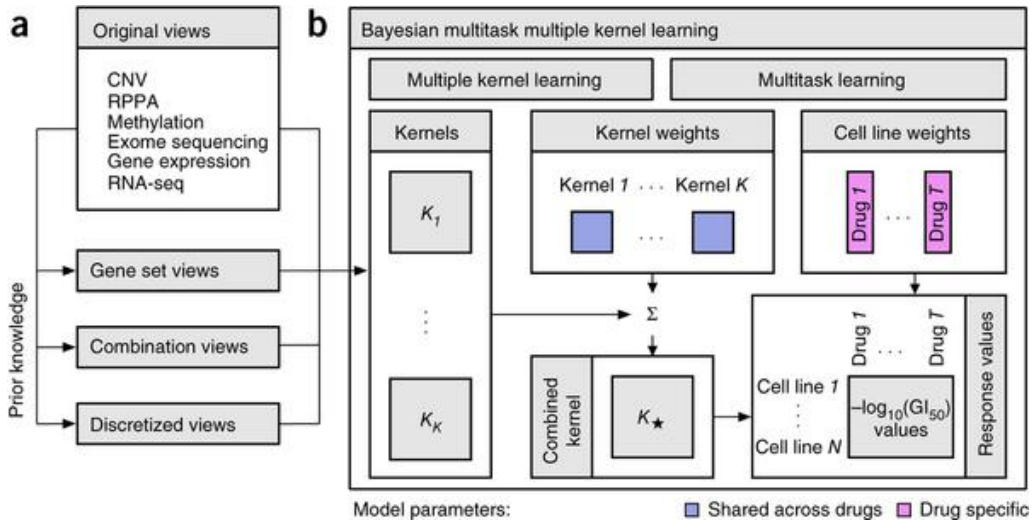


Figure 6.2: Overview of the best performing probabilistic multi-source machine learning method in NCI-DREAM Drug Sensitivity Prediction Challenge [4].

We further extended this line of research by proposing novel multi-view kernelized methods

(see C2) that not only utilize the molecular and genomic features but can additionally incorporate chemical properties of drugs. This is necessary for making predictions for new drugs on existing cells, which is 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. We also addressed a novel task of predicting sensitivity of completely new drugs on new cells. This task is challenging and required several different types of side information sources to enhance the prediction performance [5]. Furthermore, we re-formulated our earlier kernelized multi-view method to be even more effective by doing the data integration selectively [6]. The method uses the existing genomic pathway information in a novel and biologically meaningful fashion to learn associations between groups of genes and the drug sensitivities. The predictive ability of the method was confirmed with independent experimental wet lab validations.

References

[1] Suleiman A. Khan, A Faisal, J.-P. Mpindi, J. A. Parkkinen, T. Kalliokoski, A. Poso, O. P. Kallioniemi, K. Wennerberg, 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) doi: 10.1186/1471-2105-13-112.

[2] Suleiman A. Khan, Seppo Virtanen, Olli P. Kallioniemi, Krister Wennerberg, Antti Poso, and Samuel Kaski. Identification of structural features in chemicals associated with cancer drug response: A systematic data-driven analysis. Bioinformatics 30 (17), i497-i504 (2014), doi: 10.1093/bioinformatics/btu456.

[3] Suleiman A. Khan, Samuel Kaski. Bayesian Multi-view Tensor Factorization. Proceedings of ECML, European Conference on Machine Learning, 656-671 (2014).

[4] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, Muhammad Ammad-ud-din, P. Hintsanen, Suleiman A. Khan, J.P. Mpindi, NCI Dream Community, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, J. J. Collins, D. Gallahan, D. Singer, J. Saez-Rodriguez, S. Kaski, J. W. Gray, and G. Stolovitzky. A Community Effort to Assess and Improve Drug Sensitivity Prediction Algorithms. Nature Biotechnology 32, 120-1212 (2014) doi:10.1038/nbt.2877.

[5] Muhammad Ammad-ud-din, Elisabeth Georgii, Mehmet Gönen, Tuomo Laitinen, Olli Kallioniemi, Krister Wennerberg, Antti Poso, and Samuel Kaski. Integrative and Personalized QSAR Analysis in Cancer by Kernelized Bayesian Matrix Factorization. Journal of Chemical Information and Modeling 54 (8), 2347-2359 (2014) doi: 10.1021/ci500152b.

[6] Muhammad Ammad-ud-din, Suleiman A.Khan, Disha Malani, Astrid Murumägi, Olli Kallioniemi, Tero Aittokallio and Samuel Kaski. Drug response prediction by inferring pathway-response associations with Kernelized Bayesian Matrix Factorization. Bioinformatics (in press).

6.7 Detection and prediction of multivariate associations

Despite intensive investigation, the variance of many clinically relevant phenotypes, such as low- and high-density lipoprotein cholesterol, explained by genome-wide SNP data, falls far below the heritability suggested by twin studies. This motivates the development of new approaches for association detection to help uncover the hidden effects in multivariate data. To address this problem, we developed a non-parametric Bayesian reduced rank regression model that detects multivariate associations between multiple SNPs and a high-dimensional phenotype vector [1] (Fig. 6.3A). Our formulation incorporates prior knowledge about effect sizes and can deal with both common and rare variants. We analyzed the Northern Finland Birth Cohort with 4,702 individuals, for whom genome-wide SNP data and metabolic profiles with 74 traits were available, and discovered two genes, (*XRCC4* and *MTHFD2L*), without previously reported associations, which replicated in a combined analysis of two additional cohorts (Fig. 6.3B). This demonstrates the benefits of multivariate modeling, as the same data had been previously analysed with conventional statistical methods without detecting the associations.

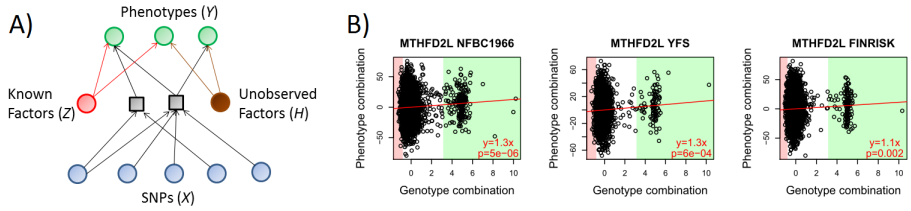


Figure 6.3: A) Graphical illustration of the Bayesian reduced rank regression model for detecting associations between high-dimensional phenotypes and genotypes. The phenotypes are assumed affected by the genotypes (SNPs), known factors such as age or sex, and unknown factors. B) Panels show the identified phenotype combination vs. the genotype combination for one novel association (gene MTHFD2L). The left panel shows results in the NFBC1966 detection data set. The center and right panels show results in two replication data sets: YFS and FINRISK.

When modeling multivariate phenotype data, reliable results depend on being able to explain away unknown confounders affecting multiple measurement variables simultaneously. For the challenging task of predicting weak effects of covariates on multivariate response variables, we extended the non-parameteric reduced rank regression formulation of [1] by assuming that the structured noise observed in the response vector is a consequence of noise affecting the latent layer of the model [2]. This assumption improves identifiability, computational tractability and allows for interpretation in terms of commonly used signal-to-noise ratio, thus simplifying the problem. Using the new model, prediction performance was improved over a wide range of multivariate real-world data sets.

References

[1] 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*, btu140, 2014.

[2] Jussi Gillberg, Pekka Marttinen, Matti Pirinen, Antti J. Kangas, Pasi Soininen, Mehreen Ali, Aki S. Havulinna, Marjo-Riitta Järvelin, Mika Ala-Korpela, Samuel Kaski Multiple Output Regression with Latent Noise. *Journal of Machine Learning Research*, 2016, Accepted for publication, JMLR W&CP

Chapter 7

Publications of COIN 2014-2015

References

[1] Michael Abseher, Martin Gebser, Nysret Musliu, Torsten Schaub, and Stefan Woltran. Shift Design with Answer Set Programming. In *Proceedings of Logic Programming and Non-monotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 32–39. Springer, 2015.

[2] Mario Alviano, Wolfgang Faber, and Martin Gebser. Rewriting recursive aggregates in answer set programming: back to monotonicity. *Theory and Practice of Logic Programming*, 15(4-5):559–573, 2015.

[3] 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 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, Florence, Italy, 2014.

[4] Muhammad Ammad-ud din, Elisabeth Georgii, Mehmet Gönen, Tuomo Laitinen, Olli Kallioniemi, Krister Wennerberg, Antti Poso, and Samuel Kaski. Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. *Journal of Chemical Information and Modeling*, 54(8):2347–2359, 2014.

[5] Jan Antfolk, Benny Salo, Katarina Alanko, Emilia Bergen, Jukka Corander, N. Kenneth Sandnabba, and Pekka Santtila. Women's and men's sexual preferences and activities with respect to the partner's age: evidence for female choice. *Evolution and Human Behavior*, 36(1):73–79, 2015.

[6] Kumaripaba Athukorala, Dorota Glowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. Is Exploratory Search Different? A Comparison of Information Search Behavior for Exploratory and Lookup Tasks. *American Society for Information Science and Technology. Journal*, 2015.

[7] Kumaripaba Athukorala, Alan Medlar, Kalle Ilves, and Dorota Glowacka. Balancing Exploration and Exploitation: Empirical Parameterization of Exploratory Search Systems. In *CIKM '15 Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, number 7, pages 1703–1706, 2015.

[8] Kumaripaba Athukorala, Alan John Medlar, Giulio Jacucci, Antti Oulasvirta, and Dorota Glowacka. Beyond Relevance: Adapting Exploration/Exploitation in Information Retrieval. In *IUI '16 Proceedings of the 21st International Conference on Intelligent User Interfaces*, 2015.

[9] Kumaripaba Athukorala, Antti Oulasvirta, Dorota Glowacka, Jilles Vreeken, and Giulio Jacucci. Narrow or Broad? Estimating Subjective Specificity in Exploratory Search. In *CIKM14*, 2014.

[10] Erik Aurell and Ralf Eichhorn. On the von Neumann entropy of a bath linearly coupled to a driven quantum system. *New Journal of Physics*, 17(8 June 2015):065007 15, 2015.

[11] Erik Aurell, Jakub Zakrzewski, and Karol Zyczkowski. Time reversals of irreversible quantum maps. *Journal of Physics A: Mathematical and Theoretical*, 48(38):0000038FT01/1–10, 2015.

[12] Abiyad Baig, Alan McNally, Steven Dunn, Konrad H. Paszkiewicz, Jukka Corander, and Georgina Manning. Genetic import and phenotype specific alleles associated with hyper-invasion in Campylobacter jejuni. *BMC Genomics*, 16, 2015.

[13] Adrian Balint, Anton Belov, Matti Järvisalo, and Carsten Sinz. Overview and analysis of the sat challenge 2012 solver competition. *Artificial Intelligence*, 223:120–155, 2015.

[14] Mutsunori Banbara, Martin Gebser, Katsumi Inoue, Max Ostrowski, Andrea Peano, Torsten Schaub, Takehide Soh, Naoyuki Tamura, and Matthias Weise. aspartame: Solving Constraint Satisfaction Problems with Answer Set Programming. In *Proceedings of Logic Programming and Non-monotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 112–126. Springer, 2015.

[15] Mukesh Bansal, Jichen Yang, Charles Karan, Michael P. Menden, James C. Costello, Hao Tang, Guanghua Xiao, Yajuan Li, Jeffrey Allen, Rui Zhong, Beibei Chen, Minsoo Kim, Tao Wang, Laura M. Heiser, Ronald Realubit, Michela Mattioli, Mariano J. Alvarez, Yao Shen, NCI-DREAM Community, Daniel Gallahan, Dinah Singer, Julio Saez-Rodriguez, Yang Xie, Gustavo Stolovitzky, and Andrea Califano. A community computational challenge to predict the activity of pairs of compounds. *Nature Biotechnology*, 32:1213–1222, 2014.

[16] Oswald Barral, Manuel J. A. Eugster, Tuukka Ruotsalo, Michiel Sovijärvi-Spape, Ilkka Kosunen, Niklas Ravaja, Samuel Kaski, and Giulio Jacucci. Exploring Peripheral Physiology as a Predictor of Perceived Relevance in Information Retrieval. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 389–399, New York, NY, 2015. ACM.

[17] Andrew Barron, Teemu Roos, and Kazuho Watanabe. Bayesian Properties of Normalized Maximum Likelihood and its Fast Computation. In *the IEEE International Symposium on Information Theory (ISIT-2014)*, 2014.

[18] Anton Belov, Daniel Diepold, Marijn Heule, and Matti Järvisalo, editors. *Proceedings of SAT Competition 2014: Solver and Benchmark Descriptions*, volume B-2014-2 of *Department of Computer Science Series of Publications B*. University of Helsinki, 2014. ISBN 978-951-51-0043-6.

[19] Anton Belov, Daniel Diepold, Marijn J.H. Heule, and Matti Järvisalo. Generating the Uniform Random Benchmarks. In *Proceedings of SAT Competition 2014: Solver and Benchmark Descriptions, volume B-2014-2 of Department of Computer Science Series of Publications B*, page 80, 2014.

[20] Anton Belov, Daniel Diepold, Marijn J.H. Heule, and Matti Järvisalo. *Proceedings of SAT Competition 2014: Solver and Benchmark Descriptions*. Department of Computer Science Series of Publications B. University of Helsinki, 2014.

[21] Anton Belov, Daniel Diepold, Marijn J.H. Heule, and Matti Järvisalo. The Application and the Hard Combinatorial Benchmarks in SAT Competition 2014. In *Proceedings of SAT Competition 2014: Solver and Benchmark Descriptions, volume B-2014-2 of Department of Computer Science Series of Publications B*, pages 81–82, 2014.

[22] Jeremias Berg, Antti Hyttinen, and Matti Järvisalo. Applications of MaxSAT in Data Analysis. In *Easychair Proceedings in Computing*, 2015.

[23] Jeremias Berg, Antti Hyttinen, and Matti Järvisalo. Applications of MaxSAT in data analysis. In *Proceedings of the 6th Pragmatics of SAT Workshop*, 2015.

[24] Jeremias Berg and Matti Järvisalo. SAT-based approaches to treewidth computation: An evaluation. In *Proceedings of the 2014 IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI 2014)*, pages 328–335. IEEE Computer Society, 2014.

[25] Jeremias Berg and Matti Järvisalo. Cost-optimal constrained correlation clustering via weighted partial maximum satisfiability. *Artificial Intelligence*, (2015).

[26] 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)*, volume 33 of *JMLR Workshop and Conference Proceedings*, pages 86–95. JMLR, 2014.

[27] Jeremias Berg, Paul Saikko, and Matti Järvisalo. Improving the effectiveness of SAT-based preprocessing for MaxSAT. In Qiang Yang and Michael Wooldridge, editors, *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 239–245. AAAI Press, 2015.

[28] Jeremias Berg, Paul Saikko, and Matti Järvisalo. Re-using auxiliary variables for MaxSAT preprocessing. In *Proceedings of the 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI 2015)*. IEEE Computer Society, 2015.

[29] Otto Berg, Paul Saikko, and Matti Järvisalo. Re-using Auxiliary Variables for MaxSAT Preprocessing. In *International Conference on Tools with Artificial Intelligence*, Proceedings, pages 813–820. International Conference on Tools with Artificial Intelligence, 2015.

[30] M. Berglund, T. Raiko, M. Honkala, L. Kärkkäinen, A. Vetek, and J. Karhunen. Bidirectional Recurrent Neural Networks as Generative Models - Reconstructing Gaps in Time Series. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 856–864, 2015.

[31] Mathias Berglund and Tapani Raiko. Stochastic gradient estimate variance in contrastive divergence and persistent contrastive divergence. In *Workshop track of the International Conference on Learning Representations (ICLR 2014), Banff, Canada*, 2014.

[32] Paul Blomstedt and Jukka Corander. Posterior Predictive Comparisons for the Two-sample Problem. *Communications in Statistics: Theory and Methods*, 44(2):376–389, 2015.

[33] Paul Blomstedt, Romain Gauriot, Niina Viitala, Tapani Reinikainen, and Jukka Corander. Bayesian predictive modeling and comparison of oil samples. *Journal of Chemometrics*, 28(1):52–59, 2014.

[34] Paul Blomstedt, Jing Tang, Jie Xiong, Christian Granlund, and Jukka Corander. A bayesian predictive model for clustering data of mixed discrete and continuous type. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[35] Paul Blomstedt, Jing Tang, Jie Xiong, Christian Granlund, and Jukka Corander. A Bayesian Predictive Model for Clustering Data of Mixed Discrete and Continuous Type. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):489–498, 2015.

[36] Jori Bomanson, Martin Gebser, and Tomi Janhunen. Improving the normalization of weight rules in answer set programs. In Eduardo Fermé and João Leite, editors, *Logics in Artificial Intelligence*, volume 8761 of *Lecture Notes in Artificial Intelligence*, pages 166–180. Springer, 2014.

[37] Jori Bomanson, Martin Gebser, Tomi Janhunen, Benjamin Kaufmann, and Torsten Schaub. Answer Set Programming Modulo Acyclicity. In *Proceedings of Logic Programming and Non-monotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 143–150. Springer, 2015.

[38] Rémi Brochenin, Thomas Linsbichler, Marco Maratea, Johannes P. Wallner, and Stefan Woltran. Abstract solvers for dung's argumentation frameworks. In Elizabeth Black, Sanjay Modgil, and Nir Oren, editors, *Proceedings of the 3rd Workshop on Theory and Applications of Formal Argumentation (TAFA 2015), revised selected papers*, volume 9524 of *Lecture Notes in Computer Science*, pages 40–58. Springer, 2015.

[39] Monica F. Bugallo, Luca Martino, and Jukka Corander. Adaptive importance sampling in signal processing. *Digital Signal Processing*, 47:36–49, 2015.

[40] Kerstin Bunte, Matti Järvisalo, Jeremias Berg, Petri Myllymäki, Jaakko Peltonen, and Samuel Kaski. Optimal neighborhood preserving visualization by maximum satisfiability. In Carla E. Brodley and Peter Stone, editors, *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*, pages 1694–1700. AAAI Press, 2014.

[41] Francesco Calimeri, Martin Gebser, Marco Maratea, and Francesco Ricca. The design of the fifth answer set programming competition. In *Technical Communications of the 30th International Conference on Logic Programming, ICLP 2014, 19-22 July, Vienna, Austria*, page Online Supplement, 2014.

[42] Ana Maria Catrina (Ene), Ioana Borze, Mohamed Guled, Mariana Costache, Gayle Leen, Maria Sajin, Elena Ionica, Aura Chitu, and Sakari Knuutila. Microrna expression profiles in kaposi's sarcoma. *Pathology and Oncology Research*, 20(1):153–159, 2014.

[43] Timothy Cerexhe, Martin Gebser, and Michael Thielscher. Online agent logic programming with oClingo. In *Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2014)*, volume 8862 of *Lecture Notes in Artificial Intelligence*, pages 945–957. Springer-Verlag, 2014.

[44] Saikat Chatterjee, David Koslicki, Siyuan Dong, Nicolas Innocenti, Lu Cheng, Yueheng Lan, Mikko Vehkapera, Mikael Skoglund, Lars K. Rasmussen, Erik Aurell, and Jukka Corander. SEK: sparsity exploiting k-mer-based estimation of bacterial community composition. *Bioinformatics*, 30(17):2423–2431, 2014.

[45] Xi Chen. *Real-time Action Recognition for RGB-D and Motion Capture Data*. Doctoral dissertation, Aalto University School of Science, 2014.

[46] Xi Chen and Markus Koskela. Skeleton-based action recognition with extreme learning machines. *Neurocomputing*, 149(Part A):397–396, 2015.

[47] C. Chewapreecha, S.R. Harris, N.J. Croucher, C. Turner, P. Marttinen, L. Cheng, A. Pessia, D.M. Aanensen, A.E. Mather, A.J. Page, S. Salter, D. Harris, F. Nosten, D. Goldblatt, J. Corander, J. Parkhill, P. Turner, and S.D. Bentley. Dense genomic sampling identifies highways of pneumococcal recombination. *Nature Genetics*, 46(3):305–309, 2014.

[48] Claire Chewapreecha, Pekka Marttinen, Nicholas J Croucher, Susannah J Salter, Simon R Harris, Alison E Mather, William P Hanage, David Goldblatt, Francois H Nosten, Claudia Turner, Paul Turner, Stephen D Bentley, and Julian Parkhill. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genetics*, 10(8):e1004547, 2014.

[49] KyungHyun Cho and Xi Chen. Classifying and visualizing motion capture sequences using deep neural networks. In *Proceedings of the International Conference on Vision Theory and Applications - VISAPP 2014*, Lisbon, Portugal, 2014.

[50] KyungHyun Cho, Tapani Raiko, Alexander Ilin, and Juha Karhunen. *How to Pretrain Deep Boltzmann Machines in Two Stages*, pages 201–219. Springer, Switzerland, 2015.

[51] Andrea Pagnani Erik Aurell Christoph Feinauer, Marcin J. Skwark. Improving contact prediction along three dimensions. *PLOS Computational Biology*, 10(10):e1003847, 2014.

[52] James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, Muhammad Ammad-ud din, Petteri Hintsanen, Suleiman A Khan, John-Patrick Mpindi, Olli Kallioniemi, Antti Honkela, Tero Aittokallio, Krister Wennerberg, NCI DREAM Community, James J Collins, Dan Gallahan, Dinah Singer, Julio Saez-Rodriguez, Samuel Kaski, Joe W Gray, and Gustavo Stolovitzky. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32:1202–1212, 2014.

[53] Sophie E. Darch, Alan McNally, Freya Harrison, Jukka Corander, Helen L. Barr, Konrad Paszkiewicz, Stephen Holden, Andrew Fogarty, Shanika A. Crusz, and Stephen P. Diggle. Recombination is a key driver of genomic and phenotypic diversity in a Pseudomonas aeruginosa population during cystic fibrosis infection. *Scientific Reports*, 5, 2015.

[54] Gino Del Ferraro and Erik Aurell. Dynamic message-passing approach for kinetic spin models with reversible dynamics. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 92(1):010102(R), 2015.

[55] Onur Dikmen and A. Taylan Cemgil. Sequential monte carlo samplers for marginal likelihood computation in multiplicative exponential noise models. In *Proceedings, IEEE Statistical Signal Processing Workshop (SSP'14)*, Gold Coast, Australia, 2014.

[56] Wolfgang Dvorak, Thomas Linsbichler, Emilia Oikarinen, and Stefan Woltran. Resolution-based grounded semantics revisited. In Simon Parsons, Nir Oren, Chris Reed, and Federico Cerutti, editors, *Computational Models of Argument, Proceedings of COMMA 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, pages 269–280. IOS Press, 2014.

[57] Wolfgang Dvorak, Matti Järvisalo, Johannes Peter Wallner, and Stefan Woltran. CEGARTIX v0.4: A SAT-Based Counter-Example Guided Argumentation Reasoning Tool. In *System Descriptions of the First International Competition on Computational Models of Argumentation (ICCMA'15)*, pages 12–14, 2015.

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

[59] Wolfgang Dvořák, Matti Järvisalo, Johannes Peter Wallner, and Stefan Woltran. Complexity-sensitive decision procedures for abstract argumentation (extended abstract). In Qiang Yang and Michael Wooldridge, editors, *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 4073–4077. AAAI Press, 2015.

[60] Ralf Eggeling, Teemu Roos, Petri Myllymäki, and Ivo Grosse. Robust learning of inhomogeneous PMMs. In *Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS-2014)*, pages 229–237, 2014.

[61] Ralf Eichhorn and Erik Aurell. Stochastic thermodynamics. *Physica Scripta*, 89(4):article id. 048001, 2014.

[62] Manuel J. A. Eugster, Tuukka Ruotsalo, Michiel M. Spapé, Ilkka Kosunen, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. Predicting term-relevance from brain signals. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 425–434, New York, NY, 2014. ACM.

[63] Ali Faisal, Jaakko Peltonen, Elisabeth Georgii, Johan Rung, and Samuel Kaski. Toward computational cumulative biology by combining models of biological datasets. *PLOS ONE*, 9(11):e113053, 2014.

[64] Xiannian Fan, Brandon Malone, and Changhe Yuan. Finding optimal Bayesian network structures with constraints learned from data. In Jin Tian and Nevin L. Zhang,

editors, *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, pages 200–209. AUAI Press, 2014.

[65] Xiannian Fan, Changhe Yuan, and Brandon Malone. Tightening bounds for bayesian network structure learning. In Carla E. Brodley and Peter Stone, editors, *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*, pages 2439–2445. AAAI Press, 2014.

[66] Nanna Fyhrquist, Lasse Ruokolainen, Alina Suomalainen, Sari Lehtimaki, Ville Veckman, Johanna Vendelin, Piia Karisola, Maili Lehto, Terhi Savinko, Hanna Jarva, Timo U. Kosunen, Jukka Corander, Petri Auvinen, Lars Paulin, Leena von Hertzen, Tiina Laatikainen, Mika Mäkelä, Tari Haahtela, Dario Greco, Ilkka Hanski, and Harri Alenius. Acinetobacter species in the skin microbiota protect against allergic sensitization and inflammation. *Journal of Allergy and Clinical Immunology*, 134(6):1301–+, 2014.

[67] Sarah Alice Gaggl, Norbert Manthey, Alessandro Ronca, Johannes Peter Wallner, and Stefan Woltran. Improved answer-set programming encodings for abstract argumentation. *Theory and Practice of Logic Programming*, 15(4–5):434–448, 2015.

[68] Luciano Gamberini, Anna Spagnolli, Benjamin Blankertz, Samuel Kaski, Jonathan Freeman, Laura Acqualagna, Oswald Barral, Maura Bellio, Luca Chech, Manuel Eugster, Eva Ferrari, Paolo Negri, Valeria Orso, Patrik Pluchino, Filippo Minelle, Baria Serim, Markus Wenzel, and Giulio Jacucci. Developing a Symbiotic System for Scientific Information Seeking: The MindSee Project. In Benjamin Blankertz, Giulio Jacucci, Luciano Gamberini, Anna Spagnolli, and Jonathan Freeman, editors, *Lecture Notes in Computer Science*, pages 68–80. Springer, 2015.

[69] Changrong Ge, Jordi Gomez-Llobregat, Marcin J. Skwark, Jean-Marie Ruysschaert, Ake Wieslander, and Martin Linden. Membrane remodeling capacity of a vesicle-inducing glycosyltransferase. *FEBS Journal*, 281(16):3667–3684, 2014.

[70] Martin Gebser, Amelia Harrison, Roland Kaminski, Vladimir Lifschitz, and Torsten Schaub. Abstract Gringo. *Theory and Practice of Logic Programming*, 15(4-5):449–463, 2015.

[71] Martin Gebser, Tomi Janhunen, Holger Jost, Roland Kaminski, and Torsten Schaub. ASP Solving for Expanding Universes. In *Proceedings of Logic Programming and Nonmonotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 354–367. Springer, 2015.

[72] Martin Gebser, Tomi Janhunen, and Jussi Rintanen. Answer set programming as SAT modulo acyclicity. In *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI 2014)*, pages 351–356. IOS Press, 2014.

[73] 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)*, pages 634–637. AAAI Press, 2014.

[74] Martin Gebser, Tomi Janhunen, and Jussi Rintanen. SAT modulo graphs: Acyclicity. In *Proceedings of the 14th European Conference on Logics in Artificial Intelligence (JELIA 2014)*, volume 8761 of *Lecture Notes in Artificial Intelligence*, pages 137–151. Springer-Verlag, 2014.

[75] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, Javier Romero, and Torsten Schaub. Progress in clasp Series 3. In *Proceedings of Logic Programming and Nonmonotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 368–383. Springer, 2015.

[76] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. Clingo = ASP + control: Preliminary report. In *Technical Communications of the 30th International Conference on Logic Programming, ICLP 2014, 19-22 July, Vienna, Austria*, page Online Supplement, 2014.

[77] Martin Gebser, Roland Kaminski, Philipp Obermeier, and Torsten Schaub. *Ricochet Robots Reloaded: A Case-Study in Multi-shot ASP Solving*, pages 17–32. Springer International Publishing, Switzerland, 2015.

[78] Martin Gebser, Marco Maratea, and Francesco Ricca. The Design of the Sixth Answer Set Programming Competition. In *Proceedings of Logic Programming and Nonmonotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 531–544. Springer, 2015.

[79] Martin Gebser, Anna Ryabokon, and Gottfried Schenner. Combining Heuristics for Configuration Problems Using Answer Set Programming. In *Proceedings of Logic Programming and Nonmonotonic Reasoning, Volume 9345 of the series Lecture Notes in Computer Science*, pages 384–397. Springer, 2015.

[80] Erik Aurell Gino Del Ferraro. Perturbative large deviation analysis of non-equilibrium dynamics. *J. Phys. Soc. Japan*, 83:084001, 2014.

[81] Mehmet Gönen and Samuel Kaski. Kernelized Bayesian matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):2047–2060, 2014.

[82] Roland Grafström, Penny Nymark, Vesa Hongisto, Ola Spjuth, Rebecca Ceder, Egon Willighagen, Barry Hardy, Samuel Kaski, and Pekka Kohonen. Toward the replacement of animal experiments through the bioinformatics-driven analysis of 'omics' data from human cell cultures. *ATLA: Alternatives to Laboratory Animals*, 43(5):325–332, 2015.

[83] Stig-Arne Grönroos, Kristiina Jokinen, Katri Hiovain, Mikko Kurimo, and Sami Virpioja. Low-Resource Active Learning of North Sámi Morphological Segmentation. In *1st International Workshop on Computational Linguistics for Uralic Languages*, pages 20–33, 2015.

[84] Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. Tuning Phrase-Based Segmented Translation for a Morphologically Complex Target Language. pages 105–111. Association for Computational Linguistics, 2015.

[85] Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185, Dublin, Ireland, 2014. Dublin City University and Association for Computational Linguistics.

[86] 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*, 28(1):31–53, 2015.

[87] James Hensman, Panagiotis Papastamoulis, Peter Glaus, Antti Honkela, and Magnus Rattray. Fast and accurate approximate inference of transcript expression from RNA-seq data. *Bioinformatics*, 31(24):3881–3889, 2015.

[88] Marijn Heule, Matti Järvisalo, Florian Lonsing, Martina Seidl, and Armin Biere. Clause elimination for SAT and QSAT. *Journal of Artificial Intelligence Research*, 53:127–168, 2015.

[89] Antti Honkela, Jaakko Peltonen, Hande Topa, Iryna Charapitsa, Filomena Matarese, Korbinian Grote, Hendrik G. Stunnenberg, George Reid, Neil D. Lawrence, and Magnus Rattray. Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays. *Proceedings of the National Academy of Sciences*, 112(42):13115–13120, 2015.

[90] Ilkka Huopaniemi and Samuel Kaski. Computational statistics approaches to study metabolic syndrome. In Matej Orešič and Antonio Vidal-Puig, editors, *A Systems Biology Approach to Study Metabolic Syndrome*, pages 319–340. Springer, Berlin, 2014.

[91] Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In Jin Tian and Nevin L. Zhang, editors, *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, pages 340–349. AUAI Press, 2014.

[92] Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Do-calculus when the True Graph is Unknown. In Marina Meila and Tom Heskes, editors, *Uncertainty in artificial intelligence*, Uncertainty in artificial intelligence, pages 395–404, 2015.

[93] Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Do-calculus when the true graph is unknown. In Tom Heskes and Marina Meila, editors, *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI 2015)*, pages 395–404. AUAI Press, 2015.

[94] Nicolas Innocenti, Monica Golumbeanu, Aymeric Fouquier D'Herouel, Caroline Lacoux, Remy A. Bonnin, Sean P. Kennedy, Francoise Wessner, Pascale Serror, Philippe Bouloc, Francis Repoila, and Erik Aurell. Whole-genome mapping of 5 ' RNA ends in bacteria by tagged sequencing: a comprehensive view in Enterococcus faecalis. *RNA*, 21(5):1018–1030, 2015.

[95] Nicolas Innocenti, Francis Repoila, and Erik Aurell. Detection and quantitative estimation of spurious double stranded DNA formation during reverse transcription in bacteria using tagRNA-seq. *RNA Biology*, 12(9):1067–1069, 2015.

[96] Satoru Ishikawa, Markus Koskela, Mats Sjöberg, Rao Muhammad Anwer, Jorma Laaksonen, and Erkki Oja. PicSOM experiments in TRECVID 2014. In *Proceedings of the TRECVID 2014 Workshop*, Orlando, FL, USA, 2014.

[97] Tomi Janhunen and Ilkka Niemelä. Cumulativity Tailored for Non-Monotonic Reasoning. In Thomas Eiter, Hannes Strass, Miroslaw Truszczynski, and Stefan Woltran, editors, *Advances in Knowledge Representation, Logic Programming, and Abstract Argumentation - Essays Dedicated to Gerhard Brewka on the Occassion of his 60th Birthday*, pages 96–111, Switzerland, 2015. Springer.

[98] Matti Järvisalo and Janne H. Korhonen. Conditional lower bounds for failed literals and related techniques. In Uwe Egly and Carsten Sinz, editors, *Proceedings of the 17th International Conference on Theory and Applications of Satisfiability Testing (SAT 2014)*, volume 8561 of *Lecture Notes in Computer Science*, pages 75–84. Springer, 2014.

[99] Emma Jokinen, Ulpu Remes, and Paavo Alku. Comparison of Gaussian process regression and Gaussian mixture models in spectral tilt modelling for intelligibility enhancement of telephone speech. In *Interspeech 15*, 2015.

[100] Jussi Määttä, Samuli Siltanen, and Teemu Roos. A Fixed-Point Image Denoising Algorithm with Automatic Window Selection. In *IEEE*, European Workshop on Visual Information Processing, 2014.

[101] Melih Kandemir, Akos Vetek, Mehmet Gönen, Arto Klami, and Samuel Kaski. Multi-task and multi-view learning of user state. *Neurocomputing*, 139:97–106, 2014.

[102] Antti Kangasrääsiö, Dorota Glowacka, and Samuel Kaski. Improving Controllability and Predictability of Interactive Recommendation Interfaces for Exploratory Search. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 247–251, Atlanta, GA, 2015. ACM.

[103] Reima Karhila, Ulpu Remes, and Mikko Kurimo. Noise in HMM-based speech synthesis adaptation: analysis, evaluation methods and experiments. *IEEE J. STSP*, 8(2):285–295, 2014.

[104] J. Karhunen, T. Raiko, and K. Cho. *Unsupervised Deep Learning: A Short Review*, pages 125–142. Academic Press, The Netherlands, 2015.

[105] 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)*, pages 1886–1891, Reykjavík, Iceland, 2014. European Language Resources Association.

[106] Nadav Kashtan, Sara E Roggensack, Sébastien Rodrigue, Jessie W Thompson, Steven J Biller, Allison Coe, Huiming Ding, Pekka Marttinen, Rex R Malmstrom, Roman Stocker, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild prochlorococcus. *Science*, 344(6182):416–420, 2014.

[107] Samuel Kaski and Jukka Corander. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics: AISTATS 2014*. JMLR, 2014.

[108] Jukka-Pekka Kauppi, Melih Kandemir, Veli-Matti Saarinen, Lotta Hirvenkari, Lauri Parkkonen, Arto Klami, Riitta Hari, and Samuel Kaski. Towards brain-activity-controlled information retrieval: Decoding image relevance from MEG signals. *NeuroImage*, 112(1):288–298, 2015.

[109] Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost van de Weijer, Michael Felsberg, and Jorma Laaksonen. Compact color-texture description for texture classification. *Pattern recognition letters*, 51(1):16–22, 2015.

[110] Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost van de Weijer, Michael Felsberg, and Jorma Laaksonen. Deep Semantic Pyramids for Human Attributes and

Action Recognition. In R.R. Paulsen and K.S. Pedersen, editors, *Image Analysis. Proceedings of 19th Scandinavian Conference on Image Analysis (SCIA 2015), Copenhagen, Denmark, June 2015, Springer LNCS 9127*, pages 341–353, Switzerland, 2015. Springer International Publishing.

[111] F.S. Khan, Jiaolong Xu, J. van de Weijer, A.D. Bagdanov, R.M. Anwer, and A.M. Lopez. Recognizing Actions Through Action-Specific Person Detection. *IEEE transactions on image processing*, 24(11):4422–4432, 2015.

[112] Suleiman A. Khan and Samuel Kaski. Bayesian multi-view tensor factorization. In T. Calders et al., editor, *Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2014*, volume I, pages 656–671, Berlin, 2014. Springer.

[113] Suleiman A. Khan, Seppo Virtanen, Olli P. Kallioniemi, Krister Wennerberg, Antti Poso, and Samuel Kaski. Identification of structural features in chemicals associated with cancer drug response: A systematic data-driven analysis. *Bioinformatics*, 30(17):i497–i504, 2014.

[114] Suleiman Ali Khan. Bayesian multi-view models for data-driven drug response analysis. Technical Report 105/2015, Aalto University School of Science, Helsinki, 2015.

[115] Roland Kindermann. *SMT-based Verification of Timed Systems and Software*. Doctoral dissertation, Aalto University School of Science, 2014.

[116] Arto Klami, Guillaume Bouchard, and Abhishek Tripathi. Group-sparse Embeddings in Collective Matrix Factorization. In *International Conference on Learning Representations*, 2014.

[117] Arto Klami, Seppo Virtanen, Eemeli Leppäaho, and Samuel Kaski. Group Factor Analysis. *IEEE transactions on neural networks and learning systems*, 26(9):2136–2147, 2015.

[118] Jukka Kohonen and Jukka Corander. Addition chains meet postage stamps: reducing the number of multiplications. *Journal of Integer Sequences*, 17(3), 2014.

[119] Jukka Kohonen and Jukka Corander. Computing exact clustering posteriors with subset convolution. *Communications in Statistics: Theory and Methods*, 2015.

[120] Laura Koponen, Emilia Oikarinen, Tomi Janhunen, and Laura Säilä. Optimizing phylogenetic supertrees using answer set programming. *Theory and Practice of Logic Programming*, 15(4-5):604–619, 2015.

[121] Markus Koskela and Jorma Laaksonen. Convolutional network features for scene recognition. In *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, FL, USA, 2014.

[122] David Koslicki, Saikat Chatterjee, Damon Shahrivar, Alan W. Walker, Suzanna C. Francis, Louise J. Fraser, Mikko Vehkapera, Yueheng Lan, and Jukka Corander. ARK: Aggregation of Reads by K-Means for Estimation of Bacterial Community Composition. *PLOS ONE*, 10(10), 2015.

[123] Sakari Kuikka, Jarno Vanhatalo, Samu Mäntyniemi and Henni Pulkkinen, and Jukka Corander. Experiences in Bayesian Inference in Baltic Salmon Management. *Statistical Science*, 29(1):42–49, 2014.

[124] Jorma Laaksonen, Markus Koskela, Mats Sjöberg, and Viitaniemi Ville. PicSOM Content-Based Information Retrieval System, 2015.

[125] Yuehang Lan and Erik Aurell. The stochastic thermodynamics of a rotating Brownian particle in a gradient flow. *Scientific Reports*, 5(5):12266, 2015.

[126] Gemma C. Langridge, Maria Fookes, Thomas R. Connor, Theresa Feltwell, Nicholas Feasey, Bryony N. Parsons, Helena M. B. Seth-Smith, Lars Barquist, Anna Stedman, Tom Humphrey, Paul Wigley, Sarah E. Peters, Duncan J. Maskell, Jukka Corander, Jose A. Chabalgoity, Paul Barrow, Julian Parkhill, Gordon Dougan, and Nicholas R. Thomson. Patterns of genome evolution that have accompanied host adaptation in Salmonella. *Proceedings of the National Academy of Sciences of the United States of America*, 112(3):863–868, 2015.

[127] Gemma C. Langridgea, Maria Fookesa, Thomas R. Connora, Theresa Feltwella, Nicholas Feaseya, Bryony N. Parsons, Helena M. B. Seth-Smith, Lars Barquist, Anna Stedman, Tom Humphrey, Paul Wigley, Sarah E. Peters, Duncan J. Maskell, Jukka Corander, Jose A. Chabalgoity, Paul Barrow, Julian Parkhill, Gordon Dougan, and Nicholas R. Thomson. Patterns of Genome Evolution That Have Accompanied Host Adaptation in Salmonella. In *Proceedings of the National Academy of Sciences, 2014*, pages 863–868, 2014.

[128] Olli-Philippe Lautenbacher, Liisa Tiittula, Maija Hirvonen, Jorma Laaksonen, and Mikko Kurimo. Towards Reliable Automatic Multimodal Content Analysis. In *Conference on Empirical Methods for Natural Language Processing, Fourth Workshop on Vision and Language*, pages 6–7, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[129] Remi Lemoy, Mikko Alava, and Erik Aurell. Local search methods based on variable focusing for random K-satisfiability. *Physical Review E*, 91(1):013305/1–6, 2015.

[130] Guohua Liu, Tomi Janhunen, and Ilkka Niemelä. Introducing real variables and integer objective functions to answer set programming. In Michael Hanus and Ricardo Rocha, editors, *Declarative Programming and Knowledge Management*, volume 8439 of *Lecture Notes in Artificial Intelligence*, pages 118–135. Springer, 2014.

[131] Jaakko Luttinen, Tapani Raiko, and Alexander Ilin. Linear state-space model with time-varying dynamics. In Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8725 of *Lecture Notes in Computer Science*, pages 338–353, Berlin Heidelberg, 2014. Springer.

[132] 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, 2014. European Language Resources Association.

[133] Jussi Määttä, Daniel F. Schmidt, and Teemu Roos. Subset Selection in Linear Regression using Sequentially Normalized Least Squares: Asymptotic Theory. *Scandinavian Journal of Statistics*, 2015.

[134] Erik Aurell Magnus Ekeberg, Tuomo Hartonen. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics*, 276:341–356, 2014.

[135] Brandon Malone, Matti Järvisalo, and Petri Myllymäki. Impact of Learning Strategies on the Quality of Bayesian Networks: An Empirical Evaluation. In Marina Meila and Tom Heskes, editors, *IJCAI 2015*, pages 562–571, 2015.

[136] Brandon Malone, Matti Järvisalo, and Petri Myllymäki. Impact of learning strategies on the quality of Bayesian networks: An empirical evaluation. In Tom Heskes and Marina Meila, editors, *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI 2015)*, pages 362–371. AUAI Press, 2015.

[137] Brandon Malone, Kustaa Kangas, Matti Järvisalo, Mikko Koivisto, and Petri Myllymäki. Predicting the hardness of learning Bayesian networks. In Carla E. Brodley and Peter Stone, editors, *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*, pages 2460–2466. AAAI Press, 2014.

[138] Brandon M. Malone and Changhe Yuan. A depth-first branch and bound algorithm for learning optimal bayesian networks. In Madalina Croitoru, Sebastian Rudolph, Stefan Woltran, and Christophe Gonzales, editors, *Revised Selected Papers of the Third International Workshop on Graph Structures for Knowledge Representation and Reasoning (GKR 2013)*, volume 8323 of *Lecture Notes in Computer Science*, pages 111–122. Springer, 2014.

[139] L. Martino, H. Yang, D. Luengo, J. Kanniainen, and J. Corander. A fast universal self-tuned sampler within Gibbs sampling. *Digital Signal Processing*, 47:68–83, 2015.

[140] Luca Martino, Victor Elvira, David Luengo, and Jukka Corander. An adaptive population importance sampler. In *International Conference on Acoustics Speech and Signal Processing ICASSP*. IEEE, 2014.

[141] Luca Martino, Victor Elvira, David Luengo, and Jukka Corander. An Adaptive Population Importance Sampler: Learning From Uncertainty. *IEEE Transactions on Signal Processing*, 63(16):4422–4437, 2015.

[142] M. Marttinen, A.-M. Pajari, E. Päivärinta, M. Storvik, P. Marttinen, T. Nurmi, M. Niku, V. Piironen, and M. Mutanen. Plant sterol feeding induces tumor formation and alters sterol metabolism in the intestine of apcmin mice. *Nutrition and Cancer: An International Journal*, 66(2):259–269, 2014.

[143] Pekka Marttinen, Nicholas J. Croucher, Michael U. Gutmann, Jukka Corander, and William P. Hanage. Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microbial Genomics*, 1(1):5, 2015.

[144] 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*, 30(14):2026–2034, 2014.

[145] Guillaume Meric, Maria Miragaia, Mark de Been, Koji Yahara, Ben Pascoe, Leonardos Mageiros, Jane Mikhail, Llinos G. Harris, Thomas S. Wilkinson, Joana Rolo, Sarah Lamble, James E. Bray, Keith A. Jolley, William P. Hanage, Rory Bowden, Martin C. J. Maiden, Dietrich Mack, Herminia de Lencastre, Edward J. Feil, Jukka Corander, and Samuel K. Sheppard. Ecological Overlap and Horizontal Gene Transfer in Staphylococcus aureus and Staphylococcus epidermidis. *Genome Biology and Evolution*, 7(5):1313–1328, 2015.

[146] Mirco Michel, Sikander Hayat, Marcin J. Skwark, Chris Sander, Debora S. Marks, and Arne Elofsson. Pconsfold: improved contact predictions improve protein models. *Bioinformatics*, 30(17):pp. i482–i488, 2014.

[147] Laura Morley, Alan McNally, Konrad Paszkiewicz, Jukka Corander, Guillaume Meric, Samuel K. Sheppard, Jochen Blom, and Georgina Manning. Gene Loss and Lineage-Specific Restriction-Modification Systems Associated with Niche Differentiation in the Campylobacter jejuni Sequence Type 403 Clonal Complex. *Applied and Environmental Microbiology*, 81:3641–3647, 2015.

[148] Alexander Mozeika, Onur Dikmen, and Joonas Piili. Consistent inference of a general model using the pseudolikelihood method. *Physical Review E*, 90:010101, 2014.

[149] Quan Nguyen and Teemu Roos. Likelihood-Based Inference of Phylogenetic Networks from Sequence Data by PhyloDAG. In Adrian Horia Dediu, Francisco Hernández Quiroz, Carlos Martín-Vide, and David A. Rosenblueth, editors, *Lecture Notes in Computer Science - Lecture Notes in Bioinformatics*, Lecture Notes in Computer Science - Lecture Notes in Bioinformatics, pages 126–140, 2015.

[150] Elina Numminen, Claire Chewapreecha, Claudia Turner, David Goldblatt, Francois Nosten, Stephen D. Bentley, Paul Turner, and Jukka Corander. Climate induces seasonality in pneumococcal transmission. *Scientific Reports*, 5, 2015.

[151] Suvi Elina Numminen, Claire Chewapreecha, Jukka Siren, Claudia Turner, Paul Turner, Stephen Bentley, and Jukka Corander. Two-phase importance sampling for inference about transmission trees. *Proceedings of the Royal Society B. Biological Sciences*, 2014.

[152] Henrik Nyman, Johan Pensar, Timo Koski, and Jukka Corander. Stratified Graphical Models - Context-Specific Independence in Graphical Models. *Bayesian Analysis*, 9(4):883–908, 2014.

[153] Henrik Nyman, Jie Xiong, Johan Pensar, and Jukka Corander. Marginal and simultaneous predictive classification using stratified graphical models. *Advances in Data Analysis and Classification*, 2015.

[154] Emilia Oikarinen and Matti Järvisalo. Answer set solver backdoors. In Eduardo Ferme and Joao Leite, editors, *Proceedings of the 14th European Conference on Logics in Artificial Intelligence (JELIA 2014)*, Lecture Notes in Artificial Intelligence. Springer, 2014.

[155] Zhirong Yang Onur Dikmen and Erkki Oja. Learning the information divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7):1442–1454, 2015.

[156] Yaara Oren, Mark B. Smith, Nathan I. Johns, Millie Kaplan Zeevi, Dvora Biran, Eliora Z. Ron, Jukka Corander, Harris H. Wang, Eric J. Alm, and Tal Pupko. Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 111(45):16112–16117, 2014.

[157] 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.

[158] Juuso Parkkinen and Samuel Kaski. Probabilistic drug connectivity mapping. *BMC Bioinformatics*, 15:113, 2014.

[159] Jaakko Peltonen, Ali Faisal, Elisabeth Georgii, Johan Rung, and Samuel Kaski. Toward computational cumulative biology by combining models of biological datasets. In *NIPS 2014 workshop on Machine Learning in Computational Biology*, 2014.

[160] Johan Pensar, Henrik Nyman, Timo Koski, and Jukka Corander. Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models. *Data Mining and Knowledge Discovery*, 29(2):503–533, 2015.

[161] Alberto Pessia, Yonatan Grad, Sarah Cobey, Juha Puranen, and Jukka Corander. K-Pax2: Bayesian identification of cluster-defining amino acid positions in large sequence datasets. *Microbial Genomics*, 1(1), 2015.

[162] Hamed R.-Tavakoli, Adham Atyabi, Seppo J. Rantanen. Antti, Laukka, Samia Nefti-Meziani, and Janne Heikkilä. Predicting the Valence of a Scene from Observers' Eye Movements. *PLOS ONE*, 10(9):0138198, 2015.

[163] Tapani Raiko, Mathias Berglund, Guillaume Alain, and Laurent Dinh. Techniques for learning binary stochastic feedforward neural networks. In *NIPS 2014 Workshop on Deep Learning and Representation Learning, Montreal, Canada*, ArXiv preprint, [stat.ML], 2014.

[164] Tapani Raiko, Mathias Berglund, Guillaume Alain, and Laurent Dinh. Techniques for Learning Binary Stochastic Feedforward Neural Networks. In *ICLR 2015*, 2015.

[165] Tapani Raiko, Li Yao, KyungHyun Cho, and Yoshua Bengio. Iterative Neural Autoregressive Distribution Estimator (NADE-k). In M. Welling Z. Ghahramani, N.D. Lawrence C. Cortes, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2015.

[166] Ana Ramirez Lopez, Nobutaka Ono, Ulpu Remes, Kalle Palomäki, and Mikko Kurimo. Designing multichannel source separation based on single-channel source separation. In *IEEE International Conference on Acoustics*, unknown, 2015.

[167] Antti Rasmus, Harri Valpola, Mathias Berglund, Mikko Honkala, and Tapani Raiko. Semi-Supervised Learning with Ladder Networks. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, Garnett R., and R. Garnett, editors, *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, page 3540, 2015.

[168] Ulpu Remes, Ana Ramirez Lopez, Kalle Palomäki, and Mikko Kurimo. Bounded conditional mean imputation with observation uncertainties and acoustic model adaptation. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(7):1198–1208, 2015.

[169] Sandra Reuter, Thomas R. Connor, Lars Barquist, Danielle Walker, Theresa Feltwell, Simon R. Harris, Maria Fookes, Miquette E. Hall, Nicola K. Petty, Thilo M. Fuchs, Jukka Corander, Muriel Dufour, Tamara Ringwood, Cyril Savin, Christiane Bouchier, Liliane Martin, Minna Miettinen, Mikhail Shubin, Julia M. Riehm, Riikka Laukkanen-Ninios, Leila M. Sihvonen, Anja Siitonen, Mikael Skurnik, Juliana Pfrimer Falcao, Hiroshi Fukushima, Holger C. Scholz, Michael B. Prentice, Brendan W. Wren, Julian Parkhill, Elisabeth Carniel, Mark Achtman, Alan McNally, and Nicholas R. Thomson. Parallel independent evolution of pathogenicity within the

genus Yersinia. *Proceedings of the National Academy of Sciences of the United States of America*, 111(18):6768–6773, 2014.

[170] Jussi Rintanen. Constraint-based algorithm for computing temporal invariants. In E. Fermé and J. Leite, editors, *Logics in Artificial Intelligence, 14th European Conference, JELIA 2014, September 2014, Proceedings*, volume 8761 of *Lecture Notes in Computer Science*, pages 665–673. Springer-Verlag, 2014.

[171] Jussi Rintanen. Discretization of temporal models with application to planning with SMT. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, pages 3349–3355. AAAI Press, 2015.

[172] Jussi Rintanen. Impact of Modeling Languages on the Theory and Practice in Planning Research. In Blai Bonet and Sven Koenig, editors, *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, pages 4052–4056, Palo Alto, California, 2015. AAAI Press.

[173] Jussi Rintanen. Models of Action Concurrency in Temporal Planning. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 1659–1665. AAAI Press, 2015.

[174] Jorma Rissanen, Peter Harremoes, S. Forchhammer, Teemu Roos, and Petri Myllymäki. *Proceedings of the Eighth Workshop on Information Theoretic Methods in Science and Engineering*. Series of Publications B. University of Helsinki, Department of Computer Science, Finland, 2015.

[175] Tuukka Ruotsalo, Jaakko Peltonen, Manuel J.A. Eugster, Dorota Glowacka, Aki Reijonen, Giulio Jacucci, Petri Myllymäki, and Samuel Kaski. Intentradar: Search user interface that anticipates user's search intents. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, pages 455–458, New York, NY, USA, 2014. ACM.

[176] Tuukka Ruotsalo, Jaakko Peltonen, Manuel JA Eugster, Dorota Glowacka, Aki Reijonen, Giulio Jacucci, Petri Myllymäki, and Samuel Kaski. SciNet: Interactive Intent Modeling for Information Discovery. In *Communications of the ACM, Vol. 58 No. 1*, pages 1043–1044, 2015.

[177] Paul Saikko and Matti Järvisalo. LMHS: A SAT-IP Hybrid MaxSAT solver, 2015.

[178] Paul Saikko, Brandon Malone, and Matti Järvisalo. MaxSAT-based cutting planes for learning graphical models. In Laurent Michel, editor, *Proceedings of the 12th International Conference on Integration of Artificial Intelligence and Operations Research Techniques in Constraint Programming (CPAIOR 2015)*, volume 9075 of *Lecture Notes in Computer Science*, pages 345–354. Springer, 2015.

[179] Hiroaki Sasaki, Michael U. Gutmann, H. Shouno, and Aapo Hyvärinen. Estimating Dependency Structures for non-Gaussian Components with Linear and Energy Correlations. In *Journal of Machine Learning Research*, JMLR: Workshop and Conference Proceedings, pages 868–876, 2014.

[180] Hannes Schulz, Kyunghyun Cho, Tapani Raiko, and Sven Behnke. Two-layer contractive encodings for learning stable nonlinear features. *Neural networks*, 64:4–11, 2015.

[181] Sohan Seth and Manuel Eugster. Archetypal Analysis for Nominal Observations. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):849 –861, 2015.

[182] Sohan Seth and Manuel. Eugster. Probabilistic Archetypal Analysis. *Machine Learning*, 102(1):85–113, 2015.

[183] Sohan Seth, John Shawe-Taylor, and Samuel Kaski. Retrieval of experiments by efficient comparison of marginal likelihoods. In Chu Kiong Loo, Kemm Siah Yap, Kok Wai Wong, Andrew Teon, and Kaizhu Huang, editors, *Neural Information Processing, Proceedings of ICONIP 2014*, volume Part II of *Lecture Notes in Computer Science Volume 8835*, pages 135–142, Switzerland, 2014. Springer.

[184] Sohan Seth, Niko Välimäki, Samuel Kaski, and Antti Honkela. Exploration and retrieval of whole-metagenome sequencing samples. *Bioinformatics*, 30(17):2471–2479, 2014.

[185] Samuel K Sheppard, Lu Cheng, Guillaume Méric, Caroline Haan, Ann-Katrin Llarena, Pekka Marttinen, Ana Vidal, Anne Ridley, Felicity Clifton-Hadley, Thomas R Connor, et al. Cryptic ecology among host generalist campylobacter jejuni in domestic animals. *Molecular Ecology*, 23(10):2442–2451, 2014.

[186] Michael Sherman, Gradeigh Clark, Yulong Yang, Shridatt Sugrim, Arttu Modig, Janne Lindqvist, Antti Oulasvirta, and Teemu Roos. User-Generated Free-Form Gestures for Authentication: Security and Memorability. In *MobiSys '14 Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, 2014.

[187] Rakshith Shetty and Jorma Laaksonen. Video captioning with recurrent networks based on frame- and video-level features and visual content classification. Technical report, ICCV 2015, Santiago, Chile, 2015.

[188] Mats Sjöberg, Bogdan Ionescu, Yu-Gang Jiang, Vu Lam Quang, Markus Schedl, and Claire-Hélène Demarty. The MediaEval 2014 Affect Task: Violent Scenes Detection. In *MediaEval 2014 Workshop*, Barcelona, Spain, 2014. CEUR Workshop Proceedings.

[189] Mats Sjöberg and Jorma Laaksonen. Using semantic features to detect novel visual concepts. In *Proceedings of the 12th International Workshop on Content Based Multimedia Indexing (CBMI 2014)*, pages 1–6, Klagenfurt, Austria, 2014. IEEE.

[190] Marcin J. Skwark, Daniele Raimondi, Mirco Michel, and Arne Elofsson. Improved contact predictions using the recognition of protein like contact patterns. *PLOS Computational Biology*, 10(11):pp. e1003889, 2014.

[191] Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

[192] Panu Somervuo, Jouni Kvist, Suvi Ikonen, Petri Auvinen, Lars Paulin, Patrik Koskinen, Liisa Holm, Minna Taipale, Anne Duplouy, Annukka Ruokolainen, Suvi Saarnio, Jukka Siren, Jukka Kohonen, Jukka Corander, Mikko J. Frilander, Virpi

Ahola, and Ilkka Hanski. Transcriptome Analysis Reveals Signature of Adaptation to Landscape Fragmentation. *PLOS ONE*, 9(7), 2014.

[193] Dag Sonntag, Matti Järvisalo, Jose Pena, and Antti Hyttinen. Learning Optimal Chain Graphs with Answer Set Programming. In Tom Heskes and Marina Meila, editors, *IJCAI 2015*, pages 822–831, 2015.

[194] Dag Sonntag, Matti Järvisalo, Jose M. Peña, and Antti Hyttinen. Learning optimal chain graphs with answer set programming. In Tom Heskes and Marina Meila, editors, *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI 2015)*, pages 822–831. AUAI Press, 2015.

[195] Tommi Suvitaival, Juuso Parkkinen, Seppo Virtanen, and Samuel Kaski. Cross-organism toxicogenomics with group factor analysis. *Systems Biomedicine*, 2:e29291, 2014.

[196] Tommi Suvitaival, Simon Rogers, and Samuel Kaski. Stronger findings for metabolomics through Bayesian modeling of multiple peaks and compound correlations. *Bioinformatics*, 30(17):i461–i467, 2014.

[197] Tommi Suvitaival, Simon Rogers, and Samuel Kaski. Stronger findings from mass spectral data through multi-peak modeling. *BMC Bioinformatics*, 15:208, 2014.

[198] Jussi Tarvainen, Sjöberg, Mats, Stina Westman, Jorma Laaksonen, and Pirkko Oittinen. Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments. *IEEE Transactions on Multimedia*, 16(8):2085–2098, 2014.

[199] Sotirios Tasoulis, Lu Cheng, Niko Välimäki, Nicholas Croucher, Simon Harris, William Hanage, Teemu Roos, and Jukka Corander. Random Projection Based Clustering for Population Genomics. In *IEEE International Conference on Big Data 2014 (IEEE Big Data 2014)*, pages 675 – 682, 2014.

[200] Ana P. Tedim, Patricia Ruiz-Garbajosa, Jukka Corander, Concepcion M. Rodriguez, Rafael Canton, Rob J. Willems, Fernando Baquero, and Teresa M. Coque. Population Biology of Intestinal Enterococcus Isolates from Hospitalized and Nonhospitalized Individuals in Different Age Groups. *Applied and Environmental Microbiology*, 81(5):1820–1831, 2015.

[201] Aino Tietäväinen, Jukka Corander, and Edward Hæggström. Baseline adjustment increases accurate interpretation of posturographic sway scores. *Gait & Posture*, 42(3):285–288, 2015.

[202] Jukka Toivanen, Matti Järvisalo, Olli Alm, Dan Ventura, Martti Vainio, and Hannu Toivonen. Transformational Creation of Novel Songs. *Journal of Artificial Intelligence Research*, 2015.

[203] Steven Y. C. Tong, Frieder Schaumburg, Matthew J. Ellington, Jukka Corander, Bruno Pichon, Fabian Leendertz, Stephen D. Bentley, Julian Parkhill, Deborah C. Holt, Georg Peters, and Philip M. Giffard. Novel staphylococcal species that form part of a Staphylococcus aureus-related complex: the non-pigmented Staphylococcus argenteus sp nov and the non-human primate-associated Staphylococcus schweitzeri sp nov. *International Journal of Systematic and Evolutionary Microbiology*, 65:15–22, 2015.

[204] Hande Topa and Antti Honkela. Gaussian process modelling of multiple short time series. In *Proceedings of ESANN 2015*, pages 83–88, Belgium, 2015.

[205] Hande Topa, Agnes Jonas, Robert Kofler, Carolin Kosiol, and Antti Honkela. Gaussian process test for high-throughput sequencing time series: application to experimental evolution. *Bioinformatics*, 31(11):1762–1770, 2015.

[206] Ikram Ullah, Pekka Parviainen, and Jens Lagergren. Species Tree Inference Using a Mixture Model. *Molecular biology and evolution*, 32(9):2469–2482, 2015.

[207] Karolis Uziela and Antti Honkela. Probe Region Expression Estimation for RNA-Seq Data for Improved Microarray Comparability. *PLOS ONE*, 10(5), 2015.

[208] Väinö Jääskinen, Jie Xiong, Jukka Corander, and Timo Koski. Sparse Markov chains for sequence data. *Scandinavian Journal of Statistics*, 41(3):639–655, 2014.

[209] Tommi Vatanen, Maria Osmala, Tapani Raiko, Krista Lagus, Marko Sysi-Aho, Matej Oresic, Timo Honkela, and Harri Lähdesmäki. Self-organization and missing values in SOM and GTM. *Neurocomputing*, 147:60–70, 2015.

[210] Minna Vehkala, Mikhail Shubin, Thomas R. Connor, Nicholas R. Thomson, and Jukka Corander. Novel R Pipeline for Analyzing Biolog Phenotypic Microarray Data. *PLOS ONE*, 10(3):1–14, 2015.

[211] 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)*, pages 1892–1897, Reykjavík, Iceland, 2014. European Language Resources Association.

[212] Ville Viitaniemi, Matti Karppa, and Jorma Laaksonen. 2D appearance based techniques for tracking the signer configuration in sign language video recordings. In *Proceedings of 11th International Conference on Image Analysis and Recognition (ICIAR 2014)*, volume 2, pages 29–38, Vilamoura, Portugal, 2014.

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

[214] Ville Viitaniemi, Mats Sjöberg, Markus Koskela, Satoru Ishikawa, and Jorma Laaksonen. *Advances in Visual Concept Detection: Ten Years of TRECVID*, pages 249–278. Academic Press, Amsterdam, The Netherlands, 2015.

[215] Sami Virpioja and Stig-Arne Grönroos. LeBLEU: N-gram-based Translation Evaluation Score for Morphologically Complex Languages. In *The Tenth Workshop on Statistical Machine Translation (WMT15)*, pages 411–416. Association for Computational Linguistics, 2015.

[216] Astrid von Mentzer, Thomas R. Connor, Lothar H. Wieler, Torsten Semmler, Atsushi Iguchi, Nicholas R. Thomson, David A. Rasko, Enrique Joffre, Jukka Corander, Derek Pickard, Gudrun Wiklund, Ann-Mari Svennerholm, Asa Sjoling, and Gordon Dougan. Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution. *Nature Genetics*, 46(12):1321–1326, 2014.

[217] Liang Wang, Sotirios Tasoulis, Teemu Roos, and Jussi Kangasharju. Kvasir: Seamless Integration of Latent Semantic Analysis-Based Content Provision into Web Browsing. In *International World Wide Web Conference Committee (IW3C2)*, pages 251–254, 2015.

[218] Kazuho Watanabe and Teemu Roos. Achievability of asymptotic minimax regret by horizon-dependent and horizon-independent strategies. *Journal of Machine Learning Research*, 16:2357–2375, 2015.

[219] Lu Wei, Zhong Zheng, Jukka Corander, and Giorgio Taricco. Outage capacity of OSTBCs over pico-cellular MIMO channels. In *Information Theory (ISIT), 2014 IEEE International Symposium*, pages 616–620, 2014.

[220] Lu Wei, Zhong Zheng, Jukka Corander, and Giorgio Taricco. On the Outage Capacity of Orthogonal Space-Time Block Codes Over Multi-Cluster Scattering MIMO Channels. *IEEE Transactions on Communications*, 63(5):1700–1711, 2015.

[221] Lucy A. Weinert, Roy R. Chaudhuri, Jinhong Wang, Sarah E. Peters, Jukka Corander, Thibaut Jombart, Abiyad Baig, Kate J. Howell, Minna Vehkala, Niko Valimaki, David Harris, Tran Thi Bich Chieu, Nguyen Van Vinh Chau, James Campbell, Constance Schultsz, Julian Parkhill, Stephen D. Bentley, Paul R. Langford, Andrew N. Rycroft, Brendan W. Wren, Jeremy Farrar, Stephen Baker, Ngo Thi Hoa, Matthew T. G. Holden, Alexander W. Tucker, Duncan J. Maskell, and BRaDP1T Consortium. Genomic signatures of human and animal disease in the zoonotic pathogen Streptococcus suis. *Nature Communications*, 6, 2015.

[222] Helga Westerlind, Väinö Jääskinen, Jukka Corander, Jan Hillert, and Timo Koski. The learning for mixtures of multicausal interaction networks. *Statistics in Medicine*, 2014.

[223] Chirayu Wongchokprasitti, Jaakko Peltonen, Tuukka Ruotsalo, Payel Bandyopadhyay, Giulio Jacucci, and Peter Brusilovsky. User Model in a Box: Cross-System User Model Transfer for Resolving Cold Start Problems. In Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Seamus Lawless, editors, *Lecture Notes in Computer Science, Vol. 9146*, pages 289–301, Switzerland, 2015. Springer International Publishing.

[224] Jie Xiong, Väinö Jääskinen, and Jukka Corander. Recursive learning for sparse Markov models. *Bayesian Analysis*, 2015.

[225] Zhirong Yang, Jaakko Peltonen, and Samuel Kaski. Optimization equivalence of divergences improves neighbor embedding. In Eric P. Xing and Tony Jebara, editors, *Proceedings of ICML 2014, The 31st International Conference on Machine Learning*, pages 460–468. JMLR, 2014.

[226] Zhirong Yang, Jaakko Peltonen, and Samuel Kaski. Majorization-Minimization for Manifold Embedding. In Guy Lebanon and S.V.N. Vishwanathan, editors, *The 18th International Conference on Artificial Intellgence and Statistics (AISTATS'15)*, pages 1088–1097, USA, 2015. JMLR W&CP.

[227] He Zhang, Mehmet Gonen, Zhirong Yang, and Erkki Oja. Understanding emotional impact of images using bayesian multiple kernel learning. *Neurocomputing*, 165(00):3–13, 2015.

[228] He Zhang, Zhirong Yang, and Erkki Oja. Adaptive multiplicative updates for quadratic nonnegative matrix factorization. *Neurocomputing*, 134:206–213, 2014.

[229] He Zhang, Zhirong Yang, and Erkki Oja. Improving cluster analysis by co-initializations. *Pattern Recognition Letters*, 45:71–77, 2014.

[230] Zhong Zheng, Lu Wei, Roland Speicher, Ralf Müller and Jyri Hämäläinen , and Jukka Corander. On the Fluctuation of Mutual Information in Double Scattering MIMO Channels. In *2014 International Zurich Seminar on Communications*, pages 104–107, 2014.

[231] Yuan Zou and Teemu Roos. On Model Selection, Bayesian Networks, and the Fisher Information Integral. In Joe Suzuki and Maomi Ueno, editors, *Proc. 2nd Workshop on Advanced Methodologies for Bayesian Networks (AMBN-2015)*, Lecture Notes in Computer Science 9505, Springer, pages 122–135, 2015.