Doctoral dissertations

Learning from environmental data: methods for analysis of forest nutrition time series

Mika Sulkava

Dissertation for the degree of Doctor of Science in Technology on 18 January 2008.

External examiners:
Sašo Džeroski (Jožef Stefan Institute, Ljubljana, Slovenia)
Alfredo Vellido (Universitat Politècnica de Catalunya, Barcelona, Spain)
Opponent:
Thomas Martinez (University of Lübeck, Germany)

Abstract:
Data analysis methods play an important role in increasing our knowledge of the environment as the amount of data measured from the environment increases. This thesis fits under the scope of environmental informatics and environmental statistics. They are fields, in which data analysis methods are developed and applied for the analysis of environmental data.

The environmental data studied in this thesis are time series of nutrient concentration measurements of pine and spruce needles. In addition, there are data of laboratory quality and related environmental factors, such as the weather and atmospheric depositions.

The most important methods used for the analysis of the data are based on the self-organizing map and linear regression models. First, a new clustering algorithm of the self-organizing map is proposed. It is found to provide better results than two other methods for clustering of the self-organizing map. The algorithm is used to divide the nutrient concentration data into clusters, and the result is evaluated by environmental scientists. Based on the clustering, the temporal development of the forest nutrition is modeled and the effect of nitrogen and sulfur deposition on the foliar mineral composition is assessed.

Second, regression models are used for studying how much environmental factors and properties of the needles affect the changes in the nutrient concentrations of the needles between their first and second year of existence. The aim is to build understandable models with good prediction capabilities. Sparse regression models are found to outperform more traditional regression models in this task.

Third, fusion of laboratory quality data from different sources is performed to estimate the precisions of the analytical methods. Weighted regression models are used to quantify how much the precision of observations can affect the time needed to detect a trend in environmental time series. The results of power analysis show that improving the quality may decrease the time needed for detection of the trend by many years.

The data analysis methods developed and applied in this thesis are found to produce results which are understandable for the environmental scientists. They are, therefore, useful for studying the condition of the environment and evaluating the possible causes for changes in it.

Stability and inference in discrete diffusion scale-spaces

Ramunas Girdziusas

Dissertation for the degree of Doctor of Science in Technology on 29 February 2008.

External examiners:
Samuli Siltanen (Tampere University of Technology)
Keijo Ruotsalainen (University of Oulu)
Opponents:
Samuli Siltanen (Tampere University of Technology)
Atanas Gotchev (Tampere University of Technology)

Abstract:
Taking averages of observations is the most basic method to make inferences in the presence of uncertainty. In late 1980's, this simple idea has been extended to the principle of successively average less where the change is faster, and applied to the problem of revealing a signal with jump discontinuities in additive noise.

Successive averaging results in a family of signals with progressively decreasing amount of details, which is called the scale-space and further conveniently formalized by viewing it as a solution to a certain diffusion-inspired evolutionary partial differential equation (PDE). Such a model is known as the diffusion scale-space and it possesses two long-standing problems: (i) model analysis which aims at establishing stability and guarantees that averaging does not distort important information, and (ii) model selection, such as identification of the optimal scale (diffusion stopping time) given an initial noisy signal and an incomplete model.

This thesis studies both problems in the discrete space and time. Such a setting has been strongly advocated by Lindeberg [1991] and Weickert [1996] among others. The focus of the model analysis part is on necessary and sufficient conditions which guarantee that a discrete diffusion possesses the scale-space property in the sense of sign variation diminishing. Connections with the total variation diminishing and the open problem in a multivariate case are discussed too.

Considering the model selection, the thesis unifies two optimal diffusion stopping principles: (i) the time when the Shannon entropy-based Liapunov function of Sporring and Weickert [1999] reaches its steady state, and (ii) the time when the diffusion outcome has the least correlation with the noise estimate, contributed by Mrázek and Navara [2003]. Both ideas are shown to be particular cases of the marginal likelihood inference. Moreover, the suggested formalism provides first principles behind such criteria, and removes a variety of inconsistencies. It is suggested that the outcome of the diffusion should be interpreted as a certain expectation conditioned on the initial signal of observations instead of being treated as a random sample or probabilities. This removes the need to normalize signals in the approach of Sporring and Weickert [1999], and it also better justifies application of the correlation criterion of Mrázek and Navara [2003].

Throughout this work, the emphasis is given on methods that enable to reduce the problem to that of establishing the positivity of a quadratic form. The necessary and sufficient conditions can then be approached via positivity of matrix minors. A supplementary appendix is provided which summarizes a novel method of evaluating matrix minors. Intuitive examples of difficulties with statistical inference conclude the thesis.

Data analysis methods for cellular network performance optimization

Pasi Lehtimäki

Dissertation for the degree of Doctor of Science in Technology on 3 April 2008.

External examiners:
Jyrki Joutsensalo (University of Jyväskylä)
Ari Hämäläinen (Nokia Research Center)
Opponent:
Tapani Ristaniemi (University of Jyväskylä)

Abstract:
Modern cellular networks including GSM/GPRS and UMTS networks offer faster and more versatile communication services for the network subscribers. As a result, it becomes more and more challenging for the cellular network operators to enhance the usage of available radio resources in order to meet the expectations of the customers.

Cellular networks collect vast amounts of measurement information that can be used to monitor and analyze the network performance as well as the quality of service. In this thesis, the application of various data-analysis methods for the processing of the available measurement information is studied in order to provide more efficient methods for performance optimization.

In this thesis, expert-based methods have been presented for the monitoring and analysis of multivariate cellular network performance data. These methods allow the analysis of performance bottlenecks having an effect in multiple performance indicators.

In addition, methods for more advanced failure diagnosis have been presented aiming in identification of the causes of the performance bottlenecks. This is important in the analysis of failures having effect on multiple performance indicators in several network elements.

Finally, the use of measurement information in selection of most useful optimization action have been studied. In order to obtain good network performance efficiently, the expected performance of the alternative optimization actions must be possible to evaluate. In this thesis, methods to combine measurement information and application domain models are presented in order to build predictive regression models that can be used to select the optimization actions providing the best network performance.

Algorithms for approximate Bayesian inference with applications to astronomical data analysis

Markus Harva

Dissertation for the degree of Doctor of Science in Technology on 9 May 2008.

External examiners:
Petri Myllymäki (University of Helsinki)
Mark Plumbley (University College of London, United Kingdom)
Opponent:
Manfred Opper (Technical University of Berlin)

Abstract:
Bayesian inference is a theoretically well-founded and conceptually simple approach to data analysis. The computations in practical problems are anything but simple though, and thus approximations are almost always a necessity. The topic of this thesis is approximate Bayesian inference and its applications in three intertwined problem domains.

Variational Bayesian learning is one type of approximate inference. Its main advantage is its computational efficiency compared to the much applied sampling based methods. Its main disadvantage, on the other hand, is the large amount of analytical work required to derive the necessary components for the algorithm. One part of this thesis reports on an effort to automate variational Bayesian learning of a certain class of models.

The second part of the thesis is concerned with heteroscedastic modelling which is synonymous to variance modelling. Heteroscedastic models are particularly suitable for the Bayesian treatment as many of the traditional estimation methods do not produce satisfactory results for them. In the thesis, variance models and algorithms for estimating them are studied in two different contexts: in source separation and in regression.

Astronomical applications constitute the third part of the thesis. Two problems are posed. One is concerned with the separation of stellar subpopulation spectra from observed galaxy spectra; the other is concerned with estimating the time-delays in gravitational lensing. Solutions to both of these problems are presented, which heavily rely on the machinery of approximate inference.

Modeling of mutual dependencies

Arto Klami

Dissertation for the degree of Doctor of Science in Technology on 5 September 2008.

External examiners:
Jukka Corander (Åbo Akademi)
Volker Roth (University of Basel, Switzerland)
Opponent:
Tobias Scheffer (Max Planck Institut for Computer Science, Germany)

Abstract:
Data analysis means applying computational models to analyzing large collections of data, such as video signals, text collections, or measurements of gene activities in human cells. Unsupervised or exploratory data analysis refers to a subtask of data analysis, in which the goal is to find novel knowledge based on only the data. A central challenge in unsupervised data analysis is separating relevant and irrelevant information from each other. In this thesis, novel solutions to focusing on more relevant findings are presented.

Measurement noise is one source of irrelevant information. If we have several measurements of the same objects, the noise can be suppressed by averaging over the measurements. Simple averaging is, however, only possible when the measurements share a common representation. In this thesis, we show how irrelevant information can be suppressed or ignored also in cases where the measurements come from different kinds of sensors or sources, such as video and audio recordings of the same scene.

For combining the measurements, we use mutual dependencies between them. Measures of dependency, such as mutual information, characterize commonalities between two sets of measurements. Two measurements can hence be combined to reduce irrelevant variation by finding new representations for the objects so that the representations are maximally dependent. The combination is optimal, given the assumption that what is in common between the measurements is more relevant than information specific to any one of the sources.

Several practical models for the task are introduced. In particular, novel Bayesian generative models, including a Bayesian version of the classical method of canonical correlation analysis, are given. Bayesian modeling is especially justified approach to learning from small data sets. Hence, generative models can be used to extract dependencies in a more reliable manner in, for example, medical applications, where obtaining a large number of samples is difficult. Also, novel non-Bayesian models are presented: Dependent component analysis finds linear projections which capture more general dependencies than earlier methods.

Mutual dependencies can also be used for supervising traditional unsupervised learning methods. The learning metrics principle describes how a new distance metric focusing on relevant information can be derived based on the dependency between the measurements and a supervising signal. In this thesis, the approximations and optimization methods required for using the learning metrics principle are improved.

Discriminative learning with application to interactive facial image retrieval

Zhirong Yang

Dissertation for the degree of Doctor of Science in Technology on 14 November, 2008.

External examiners:
Sami Brandt (University of Malmö)
Joni-Kristian Kämäräinen (Lappeenranta University of Technology)
Opponent:
Irwin King (The Chinese University of Hong Kong)

Abstract:
The amount of digital images is growing drastically and advanced tools for searching in large image collections are therefore becoming urgently needed. Content-based image retrieval is advantageous for such a task in terms of automatic feature extraction and indexing without human labor and subjectivity in image annotations. The semantic gap between high-level semantics and low-level visual features can be reduced by the relevance feedback technique. However, most existing interactive content-based image retrieval (ICBIR) systems require a substantial amount of human evaluation labor, which leads to the evaluation fatigue problem that heavily restricts the application of ICBIR.

In this thesis a solution based on discriminative learning is presented. It extends an existing ICBIR system, PicSOM, towards practical applications. The enhanced ICBIR system allows users to input partial relevance which includes not only relevance extent but also relevance reason. A multi-phase retrieval with partial relevance can adapt to the user's searching intention in a from-coarse-to-fine manner.

The retrieval performance can be improved by employing supervised learning as a preprocessing step before unsupervised content-based indexing. In this work, Parzen Discriminant Analysis (PDA) is proposed to extract discriminative components from images. PDA regularizes the Informative Discriminant Analysis (IDA) objective with a greatly accelerated optimization algorithm. Moreover, discriminative Self-Organizing Maps trained with resulting features can easily handle fuzzy categorizations.

The proposed techniques have been applied to interactive facial image retrieval. Both a query example and a benchmark simulation study are presented, which indicate that the first image depicting the target subject can be retrieved in a small number of rounds.

Inferring relevance from eye movements with wrong models

Jarkko Salojärvi

Dissertation for the degree of Doctor of Science in Technology on 21 November 2008.

External examiners:
Petri Myllymäki (University of Helsinki)
Guillaume Bouchard (Xerox Research Centre Europe)
Opponent:
Jan Larsen, (Technical University of Denmark)

Abstract:
Statistical inference forms the backbone of modern science. It is often viewed as giving an objective validation for hypotheses or models. Perhaps for this reason the theory of statistical inference is often derived with the assumption that the "truth" is within the model family. However, in many real-world applications the applied statistical models are incorrect. A more appropriate probabilistic model may be computationally too complex, or the problem to be modelled may be so new that there is little prior information to be incorporated. However, in statistical theory the theoretical and practical implications of the incorrectness of the model family are to a large extent unexplored.

This thesis focusses on conditional statistical inference, that is, modeling of classes of future observations given observed data, under the assumption that the model is incorrect. Conditional inference or prediction is one of the main application areas of statistical models which is still lacking a conclusive theoretical justification of Bayesian inference. The main result of the thesis is an axiomatic derivation where, given an incorrect model and assuming that the utility is conditional likelihood, a discriminative posterior yields a distribution on model parameters which best agrees with the utility. The devised discriminative posterior outperforms the classical Bayesian joint likelihood-based approach in conditional inference. Additionally, a theoretically justified expectation maximization-type algorithm is presented for obtaining conditional maximum likelihood point estimates for conditional inference tasks. The convergence of the algorithm is shown to be more stable than in earlier partly heuristic variants.

The practical application field of the thesis is inference of relevance from eye movement signals in an information retrieval setup. It is shown that relevance can be predicted to some extent, and that this information can be exploited in a new kind of task, proactive information retrieval. Besides making it possible to design new kinds of engineering applications, statistical modeling of eye tracking data can also be applied in basic psychological research to make hypotheses of cognitive processes affecting eye movements, which is the second application area of the thesis.

Input variable selection methods for construction of interpretable regression models

Jarkko Tikka

Dissertation for the degree of Doctor of Science in Technology on 12 December 2008.

External examiners:
Michel Verleysen (Université catholique de Louvain)
Patrik. O. Hoyer (University of Helsinki)
Opponent:
Colin Fyfe (University of the West of Scotland, United Kingdom)

Abstract:
Large data sets are collected and analyzed in a variety of research problems. Modern computers allow to measure ever increasing numbers of samples and variables. Automated methods are required for the analysis, since traditional manual approaches are impractical due to the growing amount of data. In the present thesis, numerous computational methods that are based on observed data with subject to modelling assumptions are presented for producing useful knowledge from the data generating system.

Input variable selection methods in both linear and nonlinear function approximation problems are proposed. Variable selection has gained more and more attention in many applications, because it assists in interpretation of the underlying phenomenon. The selected variables highlight the most relevant characteristics of the problem. In addition, the rejection of irrelevant inputs may reduce the training time and improve the prediction accuracy of the model.

Linear models play an important role in data analysis, since they are computationally efficient and they form the basis for many more complicated models. In this work, the estimation of several response variables simultaneously using the linear combinations of the same subset of inputs is especially considered. Input selection methods that are originally designed for a single response variable are extended to the case of multiple responses. The assumption of linearity is not, however, adequate in all problems. Hence, artificial neural networks are applied in the modeling of unknown nonlinear dependencies between the inputs and the response.

The first set of methods includes efficient stepwise selection strategies that assess usefulness of the inputs in the model. Alternatively, the problem of input selection is formulated as an optimization problem. An objective function is minimized with respect to sparsity constraints that encourage selection of the inputs. The trade-off between the prediction accuracy and the number of input variables is adjusted by continuous-valued sparsity parameters.

Results from extensive experiments on both simulated functions and real benchmark data sets are reported. In comparisons with existing variable selection strategies, the proposed methods typically improve the results either by reducing the prediction error or decreasing the number of selected inputs or with respect to both of the previous criteria. The constructed sparse models are also found to produce more accurate predictions than the models including all the input variables.

Advances in independent component analysis and nonnegative matrix factorization

Zhijian Yuan

Dissertation for the degree of Doctor of Science in Technology on 24 April, 2009.

External examiners:
Andrzej Cichocki (RIKEN, Japan)
Patrick O. Hoyer (University of Helsinki)
Opponent:
Fabian Theis (Helmholtz Zentrum München, Germany)

Abstract:
A fundamental problem in machine learning research, as well as in many other disciplines, is finding a suitable representation of multivariate data, i.e. random vectors. For reasons of computational and conceptual simplicity, the representation is often sought as a linear transformation of the original data. In other words, each component of the representation is a linear combination of the original variables. Well-known linear transformation methods include principal component analysis (PCA), factor analysis, and projection pursuit. In this thesis, we consider two popular and widely used techniques: independent component analysis (ICA) and nonnegative matrix factorization (NMF).

ICA is a statistical method in which the goal is to find a linear representation of nongaussian data so that the components are statistically independent, or as independent as possible. Such a representation seems to capture the essential structure of the data in many applications, including feature extraction and signal separation. Starting from ICA, several methods of estimating the latent structure in different problem settings are derived and presented in this thesis. FastICA as one of most efficient and popular ICA algorithms has been reviewed and discussed. Its local and global convergence and statistical behavior have been further studied. A nonnegative FastICA algorithm is also given in this thesis.

Nonnegative matrix factorization is a recently developed technique for finding parts-based, linear representations of non-negative data. It is a method for dimensionality reduction that respects the nonnegativity of the input data while constructing a low-dimensional approximation. The non-negativity constraints make the representation purely additive (allowing no subtractions), in contrast to many other linear representations such as principal component analysis and independent component analysis. A literature survey of Nonnegative matrix factorization is given in this thesis, and a novel method called Projective Nonnegative matrix factorization (P-NMF) and its applications are provided.

Advances in unlimited-vocabulary speech recognition for morphologically rich languages

Teemu Hirsimäki

Dissertation for the degree of Doctor of Science in Technology on 6 August, 2009.

External examiners:
Anssi Klapuri (Tampere University of Technology)
Dilek Hakkani-Tür (International Computer Science Institute, Berkeley, USA)
Opponent:
Steve Renals (University of Edinburgh, United Kingdom)

Abstract:
Automatic speech recognition systems are devices or computer programs that convert human speech into text or make actions based on what is said to the system. Typical applications include dictation, automatic transcription of large audio or video databases, speech-controlled user interfaces, and automated telephone services, for example. If the recognition system is not limited to a certain topic and vocabulary, covering the words in the target languages as well as possible while maintaining a high recognition accuracy becomes an issue.

The conventional way to model the target language, especially in English recognition systems, is to limit the recognition to the most common words of the language. A vocabulary of 60 000 words is usually enough to cover the language adequately for arbitrary topics. On the other hand, in morphologically rich languages, such as Finnish, Estonian and Turkish, long words can be formed by inflecting and compounding, which makes it difficult to cover the language adequately by vocabulary-based approaches.

This thesis deals with methods that can be used to build efficient speech recognition systems for morphologically rich languages. Before training the statistical n-gram language models on a large text corpus, the words in the corpus are automatically segmented into smaller fragments, referred to as morphs. The morphs are then used as modelling units of the n-gram models instead of whole words. This makes it possible to train the model on the whole text corpus without limiting the vocabulary and enables the model to create even unseen words by joining morphs together. Since the segmentation algorithm is unsupervised and data-driven, it can be readily used for many languages.

Speech recognition experiments are made on various Finnish recognition tasks and some of the experiments are also repeated on an Estonian task. It is shown that the morph-based language models reduce recognition errors when compared to word-based models. It seems to be important, however, that the n-gram models are allowed to use long morph contexts, especially if the morphs used by the model are short. This can be achieved by using growing and pruning algorithms to train variable-length n-gram models. The thesis also presents data structures that can be used for representing the variable-length n-gram models efficiently in recognition systems.

By analysing the recognition errors made by Finnish recognition systems it is found out that speaker adaptive training and discriminative training methods help to reduce errors in different situations. The errors are also analysed according to word frequencies and manually defined error classes.