The 3rd International Workshop on

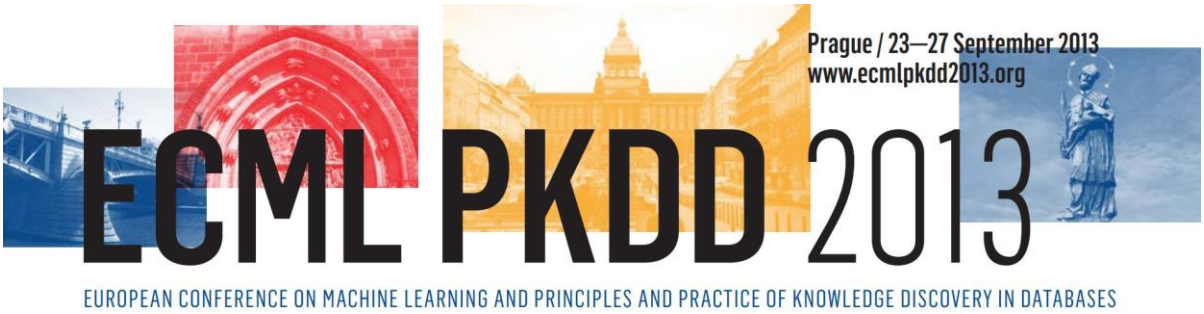
Scalable Decision Making: Uncertainty, Imperfection, Deliberation

held in conjunction with the **European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2013)**

September 23, 2013, Prague, Czech Republic

Printed in the Czech Republic

ISBN 978-80-903834-8-7



The 3rd International Workshop on

Scalable Decision Making: Uncertainty, Imperfection, Deliberation

held in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2013)

September 23, 2013

Prague, Czech Republic

http://www.utia.cz/ECMLHome

Organising and Programme Committee

Henry Brighton, Max Planck Institute for Human Development, Berlin, Germany

Itzhak Gilboa, HEC, Paris, France

Tatiana V.Guy, Institute of Information Theory and Automation, Czech Republic

Robert J. Howlett, KES International, UK

Miroslav Kárný, Institute of Information Theory and Automation, Czech Republic

Adrian Raftery, University of Washington, USA

Fabrizio Ruggeri, Institute of Applied Mathematics and Information Technology, Italy

Václav Šmídl, Institute of Information Theory and Automation, Czech Republic

David H. Wolpert, Santa Fe Institute, USA

The support of Czech Science Foundation GA 13-13502S is acknowledged.

Scope of the Workshop

Machine learning (ML) and knowledge discovery both use and serve to decision making (DM), which has to cope with uncertainty, incomplete knowledge, problem and data complexity and imperfection (limited cognitive and evaluating capabilities) of the involved heterogeneous multiple participants (aka agents, decision makers, components, controllers, classifiers, etc.). Contemporary DM deals with complex systems characterised by heterogeneous components and their goal-motivated dynamic interactions. The individual participants are selfish, i.e. follow their individual goals. There is no well-justified way to influence or describe the resulting collective behaviour of such a system via a well-proved combination of the selfish components. Economic and natural sciences describe concepts governing the functioning of systems of selfish participants as well as ways influencing their behaviour. However, the majority of solutions rely on the human moderator/manager controlling such a system. Sophisticated ML and AI solutions developed consider artificial moderators (for instance, automatic traders used in markets, e-democracy support) as well.

Without moderator, decision making with imperfect selfish decision makers lacks a firm prescriptive basis. This problem emerges repeatedly and has no easy solution. While the theoretical, algorithmic and application achievements are immense, real-life complex problems uncover discrepancies between normative and descriptive theories. This clearly indicates the need for alternative ways, deepening and unifying the current achievements across scientific schools as well as research domains. For instance, i) the consistent theory of incomplete Bayesian games cannot be applied by imperfect participants; ii) a desirable incorporation of "deliberation effort" into the design of decision-making strategies remains unsolved. At the same time, real societal, biological, economical systems efficiently cope with the imperfectness as confirmed by numerous descriptive studies. Driven by complexity, these systems exhibit a kind of (self-organising) behaviour without any intrinsic utility. This can be viewed as a result of an external control towards a common goal.

The *workshop generally aims* to exploit the knowledge and experience of multi-disciplinary scientific community and to extract a set of fundamental concepts describing a phenomenon of dynamic decision making with interacting imperfect selfish participants. Devices (e.g. robots), computer algorithms (e.g. controllers), humans (e.g. experts) and their combination are considered.

List of Invited Talks

- ***Craig Boutilier**, University of Toronto, Canada*

  PREFERENCE ELICITATION FOR SOCIAL CHOICE: A STUDY IN STABLE MATCHING AND VOTING (joint work with Joanna Drummond and Tyler Lu)

- ***David Rios Insua**, King Juan Carlos University, Madrid, Spain*

  DESIGNING SOCIETIES OF ROBOTS (joint work with Pablo G. Esteban)

- ***Miroslav Kárný,** Institute of Information Theory and Automation, Prague*

  A UNIFIED VIEW ON ROOTS OF IMPERFECTION (joint work with Tatiana V. Guy)

- ***Stephen Roberts,** University of Oxford, United Kingdom*

  SCALABLE INFORMATION AGGREGATION FROM WEAK INFORMATION SOURCES

- ***Naftali Tishby**, The Hebrew University, Israel*

  PREDICTIVE INFORMATION AND THE BRAIN'S INTERNAL TIME

- ***Alessandro E.P. Villa**, University of Lausanne, Switzerland*

  STIMULUS EVALUATION AND RESPONSE SYSTEMS STUDIED BY REACTION TIMES IN DECISION MAKING TASKS

Time	Title	Authors
8:55—9:00	Opening session	Organisers
9:00—9:35	Scalable Information Aggregation from Weak Information Sources	Stephen Roberts
9:35—10:10	Designing Societies of Robots	David Rios Insua, Pablo G. Esteban
10:10—10:30	Cooperative Dimensionality Reduction for Intelligent Feature Selection in Individualised Medicine	Dietlind Zühlke, Gernoth Grunst, Kerstin Röser
10:30—11:00	**Coffee break**	
11:00—11:35	Preference Elicitation for Social Choice: A Study in Stable Matching and Voting	Craig Boutilier
11:35—12:00	**Poster spotlights**	
12:00—12:30	**Posters and Demonstrations**:	
	Granger Lasso Causal Models in High Dimensions - Application to Gene Expression Regulatory Networks	Katerina Hlavackova-Schindler, Hamed Bouzari
	On Approximate Fully Probabilistic Design of Decision Making Strategies	Miroslav Kárný
	Preliminaries of Probabilistic Hierarchical Fault Detection	Ladislav Jirsa, Lenka Pavelková, Kamil Dedecius
	A Note on Weighted Combination Methods for Probability Estimation	Vladimíra Sečkárová
	Estimating Efficiency Offset between Two Groups of Decision-Making Units	Karel Macek
	DEMO: What Lies Beneath Players' Non-Rationalityin Ultimatum Game?	Zuzana Knejflová, Galina Avanesyan, Tatiana V.Guy, Miroslav Kárný
	DEMO: Sparsity in Bayesian Blind Source Separation and Deconvolution	Václav Šmídl, Ondřej Tichý
12:30—14:00	**Lunch Break**	
14:00—14:20	Economic Prediction Using Heterogeneous Data Streams from the World Wide Web	Abby Levenberg, Edwin Simpson, Stephen Roberts, Georg Gottlob
14:20—14:55	A Unified View on Roots of Imperfection	Miroslav Kárný, Tatiana V. Guy
14:55—15:30	Predictive Information and the Brain's Internal Time	Naftali Tishby
15:30—16:00	**Coffee break. Posters & Demos (cont.)**	
16:00—16:20	Belief CSP: A New CSP Framework Under Uncertainty	Aouatef Rouahi, Kais Ben Salah, Khaled Ghédira
16:20—16:55	Stimulus Evaluation and Response Systems Studied by Reaction Times in Decision Making Tasks	Alessandro E.P. Villa
16:55—17:30	**Panel Discussion, Closing Remarks**	All participants

Cooperative Feature Selection in Personalized Medicine

Dietlind Zühlke[1], Gernoth Grunst[2], and Kerstin Röser[3]

[1] Fraunhofer IAIS, Sankt Augustin, Germany
[2] Fraunhofer FIT, Sankt Augustin, Germany
[3] University Hospital Hamburg-Eppendorf, Germany

Abstract. In this paper we present a cooperative workflow to choose a subset of feature groups from representations of biomedical objects containing a large number of multiple typed features. The choice is done in cooperation of a computer based decision support system and a human expert of the application domain breast cancer research for personalized medicine. The iterative procedure tries to solve the Paradox of Intelligent Selection [1, cf. p. 30], i.e. the necessity to choose a suitable set of features for reasonable modelling without having modelled the actual picture with the whole set of features. The selection module is part of an enabling environment that shall support the insight into so far not yet modelled influence factors of disease processes.

1 Motivation

For personalized medicine a *stratification* of individual patients to groups of patients reacting similar to some applied therapy has to be found. In order to become relevant in the clinical routine it is necessary that these patient groups can be characterized by a suitably small and specific set of diagnostic findings. In recent years the new area of *systems medicine* aims to reveal systematic relations between "-omics" analyses (often used in the stratification research in pharmacology) and particular findings in cell morphology as well as the general information about the patient (two realms of findings used in the current clinical diagnostic routine). Together with findings about tissue and organ properties and anamnestic data the informations represent a holistic view on the patients situation. If this view spans different layers (e.g. the organism, specific organs, tissues and cells) this representation is called a *multi-layer model* of the patient [2]. This model should give the optimal orientation for the individualized therapy suggestion in a stable and sensitive diagnostic procedure.

The task to identify suitable patient groups as well as a suitable set of features defining them is not cognitively manageable by the human domain experts without adequate *information-technological support*. Reasons for this restriction are the case-related mental models of human experts, their mental focus shifting to recently handled patient cases and the complexity and heterogeneity of the multi-layer models describing the patients. Computational systems with the objective to identify case types based on multi-layer models that support the

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

cognitive abilities of human experts face several challenges. They have to bridge the cognitive gap between the case-related mental models of the pathological experts and the statistical abilities of automatic systems. The latter are – without the suitable incorporation of domain expert knowledge – often overstrained in biological applications by noise or irrelevant but massive variations. In this sense the *computational decision support system* should induce a constructive interaction [4] between the domain experts and the computational algorithms taking into account the limits of judgement of both.

In the following sections we will introduce a workflow to coordinate the actions of a computer based decision support system and a human domain expert to reveal a suitable choice of feature groups for generating an adequate model of relations e.g. in the systems medicine. The approach was developed during a PhD thesis [5]. The methods and their exemplary testing have been developed in a research project for the improvement of orientation of adjuvant breast cancer therapies.

2 The Application Context – Breast Cancer Research Project Exprimage

2.1 Aim and scope of project

The objective of the Exprimage project was to support *adjuvant therapy suggestions* in breast cancer by incorporating information from several biomedical domains. To achieve this goal, information about the therapies performed (chemotherapy, hormone therapy, radiation therapy) would have been necessary to serve as label for a supervised learning task. This information was not provided. Therefore the documented procedure and achieved results can not be seen as a reasonable proof for a superior mathematical form of feature selection. This could not reasonably be evaluated through comparison with recognition results of other forms of feature selection. Its value can nevertheless be judged by domain experts in so far as it managed to condense vague and so far not biologically validated information into potential relevance patterns that can characterize patient groups. The selected motif of features can be affirmed if pertinent biological explanations are found.

2.2 Data sets

The choice of patient cases and their representing data reflects the research interest of the pathologists in the project. The patient cases were *matched pairs* selected from a larger cohort. This means that the pathological expert chose patient cases that had the *same prognosis* according to the current diagnostic process but that showed a *different progress* of the disease, i.e. one patient survived and the other did not. The cognitive support system should help the pathologists to use additional diagnostic means like automatic image analysis, gen expression analysis, blood parameter analysis that could explain the different courses of the patients.

The cohort that was available for our studies in the Exprimage project consisted of 93 patient cases. The investigated patients' resectioned tumour tissue was older than five years – the time interval that is the pertinent clinical frame to evaluate the further perspective of the disease. The clinical data for the patients were collected in the clinical routine during diagnosis and therapy. The patients were categorized by their follow-up-status into three outcomes with the following distribution:

- Follow-up status one (alive): 50 patients
- Follow-up status two (relapse): 7 patients
- Follow-up status three (dead): 36 patients

For all analyses described later, patients with follow-up status two were neglected as there were not enough data samples and the pathologists rejected the combination with follow-up status three. As the information about the *therapy response* (the label suitable for the actual research question) was missing, the *follow-up status* of the patients was chosen as surrogate label which we refer to in classification. A naive or random classification according to the prior distribution of the follow-up status would yield a recognition rate of $58,1\%$ (the prior of class one). The classification given by the prognosis from the current diagnostic process – the *grading* of the patient – has a recognition rate of 61.4% on the given cohort.

In our subproject we focused on supporting the current diagnostic process (clinical data) using information derived from digitized tumour images marked with different histological stains (image data).

Clinical data. The clinical data reflect findings for every patient according to the current state of the art in diagnosis and prognosis. We exploited a subset of ten clinical data that were available for all patient samples. We refrained from imputation of missing data as there were not enough complete samples for a valid statistical estimation of values to be imputed. Most of the incorporated features are related to the microscopic diagnosis describing cell morphology. They are a complementary part of a multi-layer model of the patient's situation with respect to information that is gained from the images on the tissue level.

The pathologists handle some of the diagnostic features in groups according to the tumour properties they describe: the characterisation of the hormone receptors status, the invasion of vessels and the general TNM classification (a characterization that was suggested by the Union for International Cancer Control (`http://www.uicc.org/`) for determining the stage of massive tumours in general). We used this grouping of the data in our analysis. The currently established diagnostic standard – the grading – and the age were used in single feature groups.

Image data. For every patient we had two kinds of stained tissue slice images: structural and functional stains. The starting point for image analysis in Exprimage were raw digitized microscopic images of stained tumour tissue slices.

These images show the tumour and surrounding tissue and thus interactions of the tumour with supporting and nourishing structures that are potential indicators of the prognosis that are not consistently used in current diagnostic schemes. Together with the pathologists, we developed a multi-step automatic image analysis [6] building on a basic characterization of the tissue types. It calculated feature groups representing the hallmarks of cancer [7,8]. We selected two main concepts of tumour description – heterogeneity [9] and distribution patterns [6] – and analysed them under structural or functional perspectives. The derived feature groups are shown in an overview in table 2.

3 Conception of feature group selection in biomedical research

In biomedical research a set of medical concepts $m_1, \ldots, m_M$ is used to characterize the situation of the patients for the considered disease. To handle this characterization within a computational decision support system, each medical concept m_m is represented by (possibly several) groups of features $g_1^{m_m}, \ldots, g_G^{m_m}$ where one feature group g_g consists of (possibly several) features $f_1^{g_g}, \ldots, f_F^{g_g}$ that are adequately comparable using a specific dissimilarity measure d^{g_g}. In this paper we introduce a workflow to identify a subset of feature groups $\phi_1, \ldots, \phi_F$ that allows a good separation of given patient groups $p_1, \ldots, p_P$.

Feature group selection is reasonable from two perspectives: in the mathematical perspective it helps to avoid the curse of dimensionality [10] while in the medical perspective the set of feature groups to be handled should be suitable for diagnosis and prognosis in clinical routine. Especially this last condition requires the feature group set not only to be reasonably small but also to be interpretable by the human experts. Thus the identified feature groups shall be biologically evaluated with respect to their potential relevance for the patient prognosis.

3.1 State of the art in feature group selection for biomedical research

Barillot et al. [11] review different feature selection methods in the area of biomedical research, including plain statistic feature selection (see [12] for a comprehensive overview). In the application context these basic methods leave some open questions, e.g. how many features to choose. Furthermore standard feature selection methods are able to handle heterogeneous data types integrally [11]. In the research for biomarkers on the molecular level often documented domain knowledge is used for feature selection in terms of known networks of components [11], e.g. protein-protein interaction networks. In systems medicine that tries to integrate different layers of diagnostic findings, there is (yet) no schematic knowledge available in data bases.

There is a general consent that it is valuable to exploit any form of reliable knowledge in order to orient machine based selection processes [11, section 6], e.g. prior knowledge about the existence of groups of features. Working at the level

of groups of features can induce biological interpretation, reduce the dimension of the statistical problem and increase the stability of the model [11]. Barillot et al. discuss several possibilities to incorporate the domain knowledge with respect to feature groups, in order to detect the importance of groups within given signatures or vice-versa, build signatures from detected important groups.

Furthermore feature groups can be analysed according to their isolated discriminative power by statistical means of correlation with the class partition. On the other hand they can be used in discriminative modelling where the feature groups are used to learn an integral model. According to Barillot et al. group-sparse linear discrimination [11, equation (6.12) p. 190] exploits the groups structure in the features to be coherent with the model structure which directly results in easily interpretable signatures. This method is based on the logistic loss which in turn is based on single differences between numerical feature values.

3.2 Cooperative feature selection

In our application we consider different types of feature groups like numerical, categorical or functional data. Thus we can not apply group-sparse linear discrimination that is only adequate for numerical feature values in our discriminative group modelling. The interpretability of the models is very important in this application. We thus developed a new group modelling algorithm for *automatic contextual feature group selection* that is a descriptive machine learning algorithm based on *modelling class typical representatives (prototypes)*. Before integrating large number of feature groups that is statistically hard to handle into the automatic process for contextual feature groups selection, we performed a pre-selection that applied human expert knowledge.

The human experts in pathology are thinking in case-related contexts. In this respect the pre-selection has to be rendered possible by an automatic feature group ranking that extracts prototypical representatives of the patient case groups is cognitively adequate and supports the human ability to identify potentially relevant feature groups. To test this hypothesis we implemented four selection strategies shown in figure 1.

We start with a short explanation of this algorithm and will postpone the detailed description of the pre-selection strategies.

Contextual feature group selection method. As the technical selection module for discriminative group modelling is linked to corresponding human interpretations, the processing and results have to be transparent for the domain expert. In the group modelling task, we want to determine feature groups that allow the identification and discrimination of significant and potentially relevant patient groups. We developed a method based on Generalized Learning Vector Quantization [16] that models prototypical representatives of the considered classes to achieve a good coupling to the pathological expert' thinking.

To handle the different types of feature groups we extended GLVQ to a dissimilarity adaptation algorithm – Vector based Generalized Learning Vector

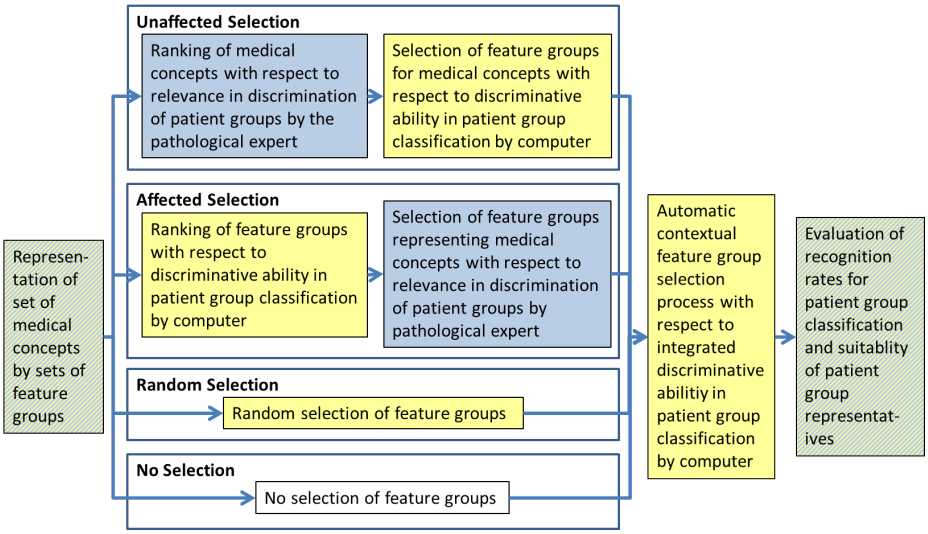


Fig. 1: Schematic workflow of feature group generation, selection and evaluation. Computer based steps are highlighted in yellow, human processing steps are marked blue. Steps incorporating both are given in grey.

Quantization (vb-GLVQ, [5]) – with specialised dissimilarity measures corresponding to the type of features in a feature group. For comparing two patient cases within this algorithm and to adapt the prototypes accordingly, the dissimilarities d^{g_j} in the single feature groups g_j are combined in a weighted sum. The overall dissimilarity between a patient sample v_k and a prototype w_n is given by:

$$D_\alpha\left(v_k, w_n\right) := \sum_{j=1}^{J} \left(\alpha_j^n\right)^2 d^{g_j}\left(\left[v_k\right]_{[j]}, \left[w_n\right]_{[j]}\right) \tag{1}$$

with $\left[v_k\right]_{[j]}$ denoting the feature values in v_k that belong to the feature group g_j and the constraint that $\sum_{j=1}^{J}\left(\alpha_j^n\right)^2 = 1$ for all $n = \{1, \ldots, N\}$ [5]. The weights α_j of the feature groups are adapted within a gradient descent approach that tries to optimize a cost function that represents a maximum margin classifier between different groups of patients. The position of the representatives are adapted in the same manner.

To interpret the dissimilarity parameters identified in this method as relevance values, the dissimilarities within a feature group have to be normalized with respect to their range and variance. In the tests we scaled every dissimilarity within a single feature group by the interquartile range of the pairwise dissimilarities for this feature group in the training data. The interquartile range is the difference between the 75th percentile and the 25th percentile of a sample and is known to be a robust measure of variance, stable against outliers [14]. There

are several further possibilities for dissimilarity normalization that we currently analyse for their influence on a suitable automatic selection of feature groups.

The number of free parameters for this method n_{fp} is given by:

$$n_{\mathrm{fp}} = n_{\mathrm{p}} \cdot n_{\mathrm{d}} + n_{\mathrm{f}} \tag{2}$$

where n_{p} denotes the number of prototypical representatives (one or more per class) and n_{d} is the number of feature dimensions (over all incorporated feature groups) which together form the number of free parameters in the position of the prototypical representatives, while n_{f} denotes the number of incorporated feature groups and accounts for the freedom in the weighting of the dissimilarities in the single feature groups.

Pre-selection of feature groups using cooperative selection strategies According to equation (2) there might be a large number of free parameters to be estimated for a single run of the automatic feature group selection method. In our example that we detail later in section 4 we have 64 feature groups with a total of 175 dimensions, that in the minimum setting of one prototypical representative per class, gives a total number of $2 \cdot 175 + 64 = 414$ free parameters that have to be estimated from 86 patient samples. As this unbalance seems to be statistically challenging we are forced to use a pre-selection of feature groups. We propose different strategies to incorporate the domain knowledge with an automatic evaluation of the discriminative power for the single feature groups.

For the evaluation of the discriminative power of a single feature group, we conducted 20 runs of the vb-GLVQ method introduced in the last section using the single feature group and one prototypical representative per class. For every run, we calculated a new random balanced selection of 30 patient cases for training and 6 patient cases for testing for each class (in total 60 cases for training and 12 cases for testing) and learned for 600 epochs. The discriminative power of a feature group was represented by the average of the test recognition rates that were achieved in the learning task taking into account their variances.

The pre-selection strategies we considered were in detail:

Unaffected selection The pathologists rank the medical concepts represented by feature groups according to the estimated relevance and redundancy. After this medically motivated ranking of the diagnostic features, a prioritization of feature groups is calculated by evaluating the discriminative power of every single feature group. For every concept the chosen number of feature groups that scored highest in the prioritisation is selected for further analysis.

Affected selection The feature groups are automatically ranked according to their single discriminative power. Starting with the best performing feature group, the feature groups are successively selected while skipping redundant diagnostic findings according to the pathological experts.

Random selection To gain a benchmark for the selection process a predefined number of feature groups is randomly drawn from the whole set of feature groups.

No selection A second benchmark is generated by using all available feature groups without any selection.

In the unaffected selection the pathologists judge the relevance of the medical concepts without previous affection or orientation by a computational analysis of the discriminative power of the corresponding feature groups. In the affected selection the results of the discriminative power analysis give a ranking of the feature groups and thus a context of judgement that triggers the selective comment of the pathologists which of the feature groups are actually selected. Both, unaffected and affected selection, incorporate pathological knowledge while neither the random selection nor the incorporation of the whole feature group set (no selection) does.

4 Tests in the application context

The following section describes the different strategies with respect to the achieved recognition results. We will not describe the pathological results of the selection process, e.g. which features were selected. These details can be found in the PhD thesis of Zühlke [5]. Rather we will provide a technical view of the interactive selection process and suggest generalized interaction strategies or schemes.

4.1 Overview of feature groups for selection

In our application example the complete set of 64 feature groups had a total of 175 dimensions. It is given in table 2 at the end of this article. For every feature group we give the full name and the type. We abbreviate numerical descriptors by N. They are handled using the Euclidean distance. The representatives of Gaussian distributions are marked by type G. They are compared using a special type of Kullback-Leibler-Divergence (cf. [5, equation (3.2.6) p. 29]). We handle representatives of distributions with the Cauchy-Schwarz-Divergence (cf. [15] for details of divergence based vector quantization). We abbreviate this type by D. For the relational feature groups we used dissimilarities based on judgements of the pathological experts (RH, cf. [5] for details of their assignment). In the right most column we show the dimensionality of the corresponding feature group that indicates the number of features within the group.

4.2 Test setting

For the feature group sets selected under the different strategies we iteratively applied the automatic contextual feature selection method (both detailed in section 3.2). In every iteration the feature group selection was based on the accumulated weights determined in 20 runs of the contextual method with one prototypical representative per class and learning for 600 epochs. In the balanced setting, we randomly selected 30 patient cases for training and 6 for testing for both classes with different random initialization for every run. In the unbalanced

setting, we split the whole patient case set randomly into 72 cases for training and 14 cases for testing in a stratified manner changing random seeds for every run. The cut-off-point for the accumulated weights over the 20 runs was chosen by hand with respect to a significant drop of this relevance measure between two consecutive feature groups. Table 1 shows the best average test recognition rates that were achieved during the iterative process using the preliminary feature group selections according to the different cooperative selection strategies.

4.3 Results for different pre-selection strategies.

In the automatic contextual feature selection based on the set extracted by the *unaffected selection strategy* none of the selected feature group sets achieved a test recognition rate that, taking the variation into account, was significantly better than random or naive classification (random: $51,8\%$). The best recognition rate achieved in one iteration of this selection process was 58.2% with a variation of 8.8% There was no clear pattern in the automatic selection of the feature group sets that showed tendencies of an improvement or decline of the recognition rates. In addition the pathological evaluation of the selection process revealed no comprehensible underlying biological concept. The same holds for the selected feature group sets.

The automatic selection of the feature groups based on the *affected selection* yielded the overall best test recognition rate of 66.7%. Taking into account the standard deviation of 7.6% this test recognition rate was higher than random classification (cf. table 1). In this stage the selection comprised

- four clinical feature groups as well as
- one feature group representing a morphometric clustering of the coarse tumour regions and
- the computationally determined ratio of the expression of the oestrogen receptor.

Further reduction of the affected selection that removed the oestrogen receptor feature group decreased the test recognition rate in all pertinent measures. This indicates that information relevant for the discrimination of the disease courses was dropped. In this case no further reduction of the model complexity and feature group set is possible without a loss of predictive power. The pathologists judged the best performing feature group set to be pathologically interesting and worth further investigation. Figure 2 shows a schematic overview over the automatic selection process using the affected feature group pre-selection.

In the contextual feature group selection process for the *random selection* of feature groups none of the results was better than random classification or the trivial classification according to the classes' prior distribution. The best mean recognition rate was 56.3% with a standard variation of 11.7%.

For the feature group set that was *not selected* the overall second best average test recognition rate was achieved for a selection of two feature groups. With a recognition rate of 65.4% and a variation of 12.2% it was better than random

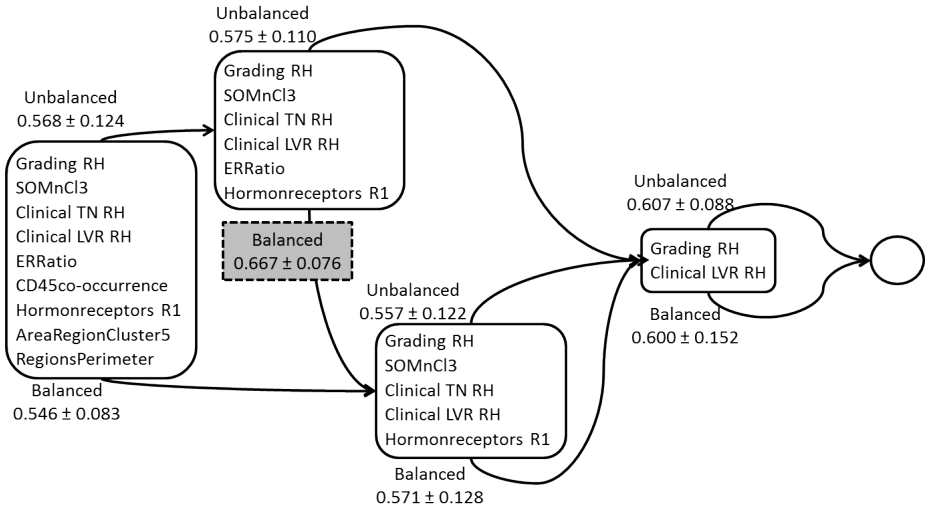


Fig. 2: Schematic overview over the automatic selection process using the affected feature group selection.

classification. However, this combination of two clinical features did not exceed the recognition rate of the current diagnostic process (61.4%) taking its standard variation into account.

The average test recognition rates are shown in an overview in table 1.

Table 1: Best average test recognition rates for the iterative automatic feature group modelling based on different cooperative feature selection strategies

Cooperative selection strategy	Best test recognition rate	Standard variation
Unaffected selection	58.2%	8.8%
Affected selection	66.7%	7.6%
Random selection	56.3%	11.7%
No selection	65.4%	12.2%

4.4 Tentative comparison of recognition rates with other modelling methods

We compared the results of our proposed workflow incorporating the different cooperative pre-selection strategies with a Generalized Learning Vector Quanti-

zation [16] using the squared Euclidean distance. While the GLVQ showed very high recognition rates for the training data set (79.1% with standard deviation of 4.7%) the generalization ability given by the test recognition rate (50.4% with standard deviation 9.9%) was significantly lower than for all tests of the cooperative pre-selection and automatic selection workflow using the vb-GLVQ with suitable dissimilarities. That shows that with the high number of free variables estimated from a small set of patient samples there is a tendency to overfit the model to the training data.

For the affected feature group selection we achieved a classification (average test recognition rate 66.7% with a standard variation of 7.6%) that in tendency is better than the classification given by the grading of the patients – the current diagnosis with a recognition rate of 61.4%. The values for the recall and precision of the classes are comparable with the difference that the grading gives slightly higher preference to follow-up status three than our classification.

5 Discussion

We analysed different feature group pre-selection strategies with respect to their suitability to enhance a workflow for feature selection with domain expert knowledge. The quality of the incorporation of domain expert knowledge was judged by the test recognition rates that were achieved using the preliminary feature group selections in an automatic contextual feature selection method as well as by the medical plausibility evaluation of the resulting feature group sets.

Both non-oriented methods of selecting a preliminary feature group set – the *random selection* as well as *no selection* – profit from the automatic contextual feature group selection process. In this process the test recognition rates increased. The whole feature group set was better than random or naive classification with respect to the bias of the classes but not better than the clinical prognosis. The test recognition rate of the random selection did not exceed that of random or naive classification if the standard deviation is taken into account. The random selection did not comprise any of the feature groups identified as relevant in the other tests or by the pathologists.

For the *unaffected selection* the automatic feature selection did not show significant improvements. Possible reasons therefore are:

- The available training data is not representative to derive the free parameters and consequently a selection criterion.
- The normalization of the dissimilarity values in the feature groups is not adequate and therefore the determined weights can not be used as selection criterion.
- The accumulation of the weight values is not adequate.
- The selection of the cut-off in the accumulated weights is not adequate.
- Relevant information was missed.

While the first four reasons are caused by the structure of the feature selection process, the last reason is concerned with the substantial information available

for the process. As only in this case the automatic selection with this structure did not improve recognition results, the most probable reason that this feature group selection failed is that important information is left out or missing. This shows that the pathological knowledge if incorporated too early into the selection process can *miss potentially relevant feature groups*.

The affected selection did profit from the automatic contextual feature selection method up to a certain extent. The method has to be monitored in the achieved test recognition rate in order to avoid oversimplification of the model or the loss of relevant feature groups. The best feature group set identified in the automatic contextual feature selection based on the affected selection (cf. section 4.3) was finally evaluated by the pathologists as showing interesting pathological relations that are worth further research. We expect that with a stable data base and relevant labels the resulting feature group selections reveal pathologically relevant information that is able to adjust adjuvant therapies as it was intended by the research project Exprimage.

6 Summary and conclusion

We described a workflow as well as different interaction strategies to identify relevant contextual feature group selections for the discrimination of disease courses in pathological research for personalized medicine. In the prediction of breast cancer follow-up we could show that using the developed learning and evaluation approaches it is possible to identify so far unknown or rather not considered diagnostic dimensions that are worth further experimental medical research.

Summarizing the discussion of the single strategies for pre-selecting feature groups there is evidence that the unaffected selection strategy is less successful than the affected selection strategy. We think that the domain experts need a context for their relevance evaluation that can be provided by the analysis of the single discriminative power of the feature groups. Furthermore the affected feature selection that incorporates the pathological knowledge is more successful than using the whole feature group set in the automatic contextual feature selection method in terms of higher mean test recognition rate and a lower variance for several runs.

The proposed workflow for feature selection is easily extendible for new feature groups. The challenging part for the incorporation of new data is the determination of a suitable dissimilarity measure in the new feature groups. If they are chosen the workflow should be started anew to account for possible cross-relations between the old and the new feature groups.

References

1. G. Myatt and W. Johnson, *Making Sense of Data III: A Practical Guide to Designing Interactive Data Visualizations*. ITPro collection, Wiley, 2011.
2. E. Klipp, W. Liebermeister, C. Wierling, A. Kowald, and H. Lehrach, *Systems biology: a textbook*. Wiley-VCH, 2009.
3. A. W. M. Smeulders, S. Member, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1349–1380, 2000.
4. N. Miyake, "Constructive interaction and the iterative process of understanding," *Cognitive Science*, vol. 10, no. 2, pp. 151–177, 1986.
5. D. Zühlke, *Vector Quantization based Learning Algorithms for Mixed Data Types and their Application in Cognitive Support Systems for Biomedical Research*. PhD thesis, 2012.
6. J. Bornemeier, "Entwicklung von merkmalen zur bestimmung räumlicher ausbreitungsmuster in histopathologischen gewebeschnitten des mammakarzinoms," Master's thesis, Institut für Computervisualistik, Fachbereich Informatik, Universität Koblenz-Landau, 2011.
7. D. Hanahan and R. A. Weinberg, "The Hallmarks of Cancer," *Cell*, vol. 100, pp. 57–70, Jan. 2000.
8. D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation.," *Cell*, vol. 144, pp. 646–674, Mar. 2011.
9. E. Khabirova, "Image processing descriptors for inner tumor growth patterns," Master's thesis, Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 2011.
10. R. Bellman and R. Corporation, *Dynamic programming*. Rand Corporation research study, Princeton University Press, 1957.
11. E. Barillot, L. Calzone, P. Hupe, J.-P. Vert, and A. Zinovyev, *Computational Systems Biology of Cancer*. Chapman & Hall/CRC Mathematical & Computational Biology, London, UK: CRC Press, 2012.
12. I. Guyon, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
13. T. Kohonen, "Learning vector quantization for pattern recognition," in *Technical Report TKK-F-A601*, 1986.
14. R. R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*. Statistical Modeling and Decision Science, Academic Press, 2nd ed., Dec. 2004.
15. T. Villmann and S. Haase, "Divergence-based vector quantization," *Neural Computation*, vol. 23, no. 5, pp. 1343–1392, 2011.
16. A. Sato and K. Yamada, "Generalized learning vector quantization," in *Advances in Neural Information Processing Systems 8*, (Cambridge, MA, USA), pp. 423–429, MIT Press, 1996.

Clinical data		
Feature group full name	Type	Dim
TN characterization of the tumour	RH	7
LVR characterization of the tumour	RH	5
Hormone receptor characterization	RH	8
Age of the patient at surgery	N	1
Grading	RH	3
Quantification of tissues types		
Feature group full name	Type	Dim
Absolute tissue area	N	2
Relative area stroma to overall tumour	N	1
Size variation of tumour regions	G	2
Perimeter variation of tumour regions	G	2
Number of AE1AE3 tumour regions	N	1
Mean area of AE1AE3 tumour regions to tumour area	N	1
Structural heterogeneity characterization		
Feature group full name	Type	Dim
Distribution of inner tumour structure	D	5
Number of regions of different inner tumour structures	N	5
Area distribution for solid homogeneous structures	G	2
Area distribution for half-homogeneous structures	G	2
Area distribution for heterogeneous structures	G	2
Area distribution for sparse heterogeneous structures	G	2
Area distribution for traces of tumour	G	2
Functional heterogeneity characterization		
Feature group full name	Type	Dim
Relative area of functional marker to tumour parenchyma	N	3
CD45 distribution in tissue types	D	4
ER distribution in tissue types	D	4
PR distribution in tissue types	D	4
CD45 co-occurrence with other functional markers	D	3
ER co-occurrence with other functional markers	D	3
PR co-occurrence with other functional markers	D	3
Number of regions	N	3
Area distribution for tumour regions	G	2
Area distribution for ER positive regions	G	2
Area distribution for PR positive regions	G	2
Number of tumour regions covered by hormone receptors	N	2
Spatial distribution of CD45 in tumour regions	D	3
Spatial distribution of ER in tumour regions	D	3

Spatial distribution of PR in tumour regions	D	3
Structural tumour distribution pattern characterization		
Feature group full name	Type	Dim
Mean and std dev of edge lengths in Min. Spanning Tree (MST)	G	2
Variation coefficient and min to max for edge lengths in MST	N	2
Average weighted node degree in MST	N	1
Number of nodes in MST	N	1
Randić index in MST	N	1
Distribution of the node degrees in MST	G	2
Variation coefficient and min to max for node degrees in MST	N	2
Mean and std dev of edge lengths in Delaunay Graph (DG)	G	2
Variation coefficient and min to max for edge lengths in DG	N	2
Average weighted node degree in DG	N	1
Number of nodes in DG	N	1
Cyclomatic number in DG	N	1
Randić index in DG	N	1
Distribution of the node degrees in DG	G	2
Variation coefficient and min to max for node degrees in DG	N	2
Morphometry clustering of coarse tumour regions (two clusters)	D	2
Morphometry clustering of coarse tumour regions (three clusters)	D	3
Morphometry clustering of coarse tumour regions (four clusters)	D	4
Morphometry clustering of coarse tumour regions (seven clusters)	D	7
Morphometry clustering of fine tumour regions (two clusters)	D	2
Morphometry clustering of fine tumour regions (three clusters)	D	3
Morphometry clustering of fine tumour regions (four clusters)	D	4
Functional tumour distribution pattern characterization		
Feature group full name	Type	Dim
Ratio CD45 to AE1AE3	N	1
Ratio ER to AE1AE3	N	1
Ratio PR to AE1AE3	N	1
Distribution of RCC8 relations for CD45	D	7
Distribution of RCC8 relations for ER	D	7
Distribution of RCC8 relations for PR	D	7
Linear Distance Quantification for CD45	D	2
Linear Distance Quantification for ER	D	2
Linear Distance Quantification for PR	D	2

Table 2: Overview over all feature groups that we considered for the development of a multi-layer model for breast cancer in Exprimage

Granger Lasso Causal Models in Higher Dimensions - Application to Gene Expression Regulatory Networks

Kateřina Hlaváčková-Schindler and Hamed Bouzari

The Gregor Mendel Institute of Molecular Plant Biology,
Austrian Academy of Sciences, Doktor-Bohr-Gasse 3, 1030 Vienna, Austria
katerina.schindler@gmail.com, hamed.bouzari@oeaw.ac.at

Abstract. Granger causality (GC), based on a vector autoregressive model, is one of the most popular methods in uncovering the temporal dependencies among time series. The original Granger model is able to detect only linear causal dependencies and many approaches were recently developed to extend it to the non-linear modeling. The method Copula-Granger from Bahadori and Liu in 2012 introduces non-linearity into the causality modeling by representing the data distribution by copulas. The detection of causality of gene regulatory networks (GRN) from experimental data, such as gene expression measurements, is a challenging problem, being solved by various computational methods with various success. We applied the Granger Lasso method, the Copula Granger method and the combination of dynamic Bayesian Networks with ordinary differential equation method (ODE-DBN) to cell division cycle gene expression data from the human cancer cell line (HeLa) for a regulatory network of 19 selected genes. We tested the causal detection ability of the methods with respect to the selected benchmark network. We compared the performance of the mentioned methods or various statistical measures. All three methods are scalable and can be easily extended to higher dimensions. The results of both Granger Lasso and Copula Granger outperformed the ODE-DBN both in terms of precision and the computational time. We conclude that the DBN combined with ODE method are not feasible for large GRN because of the computational intensity of the methods and surprisingly low precision. This type of methods is more feasible for modeling of local dynamics within a small genetic regulatory networks, rather than for detection of causal relationships in a large genetic regulatory network. We believe that the assumption of Gaussian processes, on which are DBN based, is in larger genetic regulatory networks violated.

Keywords: Granger causality, graphical Granger Lasso method, Copula Granger method, gene expression data, gene regulatory network.

1 Introduction

Granger causality (GC), based on a vector autoregressive model, is one of the most popular methods in uncovering the temporal dependencies among time se-

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

ries. The original Granger model is able to detect only linear causal dependencies and many approaches were recently developed to extend it to the non-linear modeling. The method Copula-Granger from Bahadori and Liu in 2012 introduces non-linearity into the causality modeling by representing the data distribution by copulas. It is a computationally fast method with respect to the size of the network.

Transcriptional regulation in a cell is a process with a complex non-linear dynamics. Models of transcriptional regulation are commonly depicted in the form of a network, where directed connections between nodes represent the regulatory interactions. The goal of these models is to infer the structure of gene regulatory networks (GRN) from experimental data. Biological samples are usually profiled using gene expression microarrays (GE) and the measured microRNA (mRNA) levels provide a quantitative information to assess molecular control mechanisms. A gene can be computationally represented by a single data value (row) consisting of d measurements $x^i = (x_1^i, \ldots, x_d^i)$. An experiment (sample) y is a single microarray experiment corresponding to a single column in the GE matrix, $y = (x_j^1, \ldots, x_j^n)^T$ where n is the number of genes in the data set. A gene expression profile from microarrays has typically 5000 to 100000 variables and just $15 - 100$ measurements. The detection of inference (causality, in other words) of GRN from experimental data, such as gene expression measurements, is a challenging problem, being solved by various computational methods with various success. The most applied method to model for causal relationships in gene regulatory networks from experimental data are the so called dynamic Bayesian networks (DBN), for example [23]. The exact models for small regulatory networks are commonly approximated by ODE, which can be obtained as the expectation of the chemical master equation under certain assumptions. A number of different modeling approaches using ODE with Bayesian modeling have been proposed, including, among others, Cao and Zhao, [9], Bansal et al., [6] and Zou and Conzen [25]. In our paper we consider the model from Äijö and Lähdesmäki [1]. The authors have in [1] shown that the combination of the DBN and ODE methods outperforms the causality detection in small gene regulatory networks. The drawback of these two models considered separately as well as of their combination is their exponential computational time with respect to the size of the networks, making the computation costly for large networks. Several other methods modeling causal relationships have been recently proposed and applied to gene expression data, such structural equation models, probabilistic Boolean networks, fuzzy controls and differential equations. These methods are mainly applied to small genetic networks and will not be discussed in this paper.

For these reasons, we focused on the class of GC methods which have shown to have a high precision and fast computation even for causality detection in large networks. We applied the Granger Lasso method, the Copula Granger method and the combination of dynamic Bayesian Networks with ordinary differential equation method (ODE-DBN) to cell division cycle gene expression data from the human cancer cell line (HeLa) for a regulatory network of 19 selected genes. We tested their causal detection ability with respect to a benchmark network.

We compared the performance of the mentioned methods on various statistical measures. The results of both Granger Lasso and Copula Granger Lasso outperformed the ODE-DBN both in terms of precision and the computational time. The computation of both Granger methods was a few seconds on a common PC workstation, while the DBN-ODE method needed for each gene two minutes of real time. Concerning the robustness of the methods with respect to noise, the precision of the results was tested with various levels of noise on data. The precision of the inference results was the best for Copula Granger Lasso method, while the DBN-ODE method issued into the over fitting and spurious results already with low levels of noise on data. Both Copula Granger Lasso and Granger Lasso methods are scalable methods and can be easily extended to higher dimensions. We conclude that the DBN combined with ODE method are not feasible for large GRN because of the computational intensity of the methods and surprisingly low precision. This type of methods is more feasible for modeling of local dynamics within a small genetic regulatory networks, rather than for detection of causal relationships in a large genetic regulatory network. We believe that the assumption of Gaussian processes, on which are DBN based, is in larger genetic regulatory networks violated.

2 Granger Causality

Causality has been in the literature defined in many ways. We consider the concept of causality as a time-dependent relationship among time series, as time is relevant in the biological experiments. The most used conception of time dependent causality is the Granger causality GC [12] which is based on the probabilistic notion of causality, and is defined as follows: An event X is a cause to the event Y if (i) X occurs before Y, (ii) likelihood of X is non zero, and (iii) likelihood of occurring Y given X is more than the likelihood of Y occurring alone.

Granger developed this conception of causality into the mathematical scheme based on a VAR. As Granger put it, a consequence of statements (i) and (ii) is that the causal variable can help to forecast the effect variable after other data has been first used [12]. This restricted sense of causality, referred to as Granger causality, characterizes the extent to which a process X_t is leading another process Y_t, and builds upon the notion of incremental predictability. It is said that the process X_t Granger causes another process Y_t if future values of Y_t can be better predicted using the past values of X_t and Y_t rather then only past values of Y_t. The standard test of GC developed by Granger is based on a linear vector auto-regressive model (VAR)

$$Y_t = a_o + \sum_{k=1}^{L} b_{1k} Y_{t-k} + \sum_{k=1}^{L} b_{2k} X_{t-k} + \xi_t, \tag{1}$$

where ξ_t are uncorrelated random variables with zero mean and variance σ^2, L is the specified number of time lags, and time $t = L+1, \ldots, N$. The null hypothesis

that X_t does not Granger cause Y_t is supported when $b_{2k} = 0$ for $k = 1, \ldots, L$, reducing Eq. (1) to

$$Y_t = a_o + \sum_{k=1}^{L} b_{1k} Y_{t-k} + \tilde{\xi}_t. \quad (2)$$

This model leads to two well-known alternative test statistics, the Granger-Sargent and the Granger-Wald test. The equations 1 and 2 express the bivariate causality among time series X and Y. This linear framework for measuring and testing causality has been widely applied not only in economy and finance, but also in diverse fields of natural sciences such as climatology or neurophysiology. The concept of GC can be extended to more than two time series so that the vector autoregressive model VAR is replaced by a multivariate vector autoregressive model MVAR [13]. These models are called graphical Granger models and will be discussed in the following.

Since the conception of Granger causality can detect only linear causal relationships, various nonlinear extensions of GC were proposed, for example the nonlinear predictors based on so-called radial basis functions [8]. Ancona end Marinazzo applied this idea to GC and introduced the so called kernel Granger methods, [2] and [18]. Other extension of GC are from Chen et al. and can be found in [10]. In this paper we will in the following deal with the extension of Granger method from [4].

3 Dynamic Bayesian Networks and Ordinary Differential Equations

A Bayesian network [14] or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclig graph (DAG). A dynamic Bayesain network is a Bayesian network which relates variables to each other over adjacent time steps. A causal Bayesian network is a network with an explicit requirement of causal relationships. The additional semantics of the causal networks specify that if a node X is actively caused to be in a given state x, then the probability density function changes to the one of the network obtained by cutting the links from $X's$ parents to X, and setting X to the caused value x. Using these semantics, one can predict the impact of external interventions from data obtained prior to intervention.

Zou et al. in [24] recently compared the (multivariate) GC and dynamic Bayesian networks on inference problem for both synthesized and experimental data, including GE microarray data. They concluded, that for a small sample size, the inference of DBN is better than of the GC approach, otherwise the GC performs better (in the sense of common precision measures). The drawback of dynamic Bayesian networks was is their computational intensity for large graphs.

In this paper we consider the method from Äijö and Lähdesmäki [1] applying DBN. For experimental comparison to other methods, we considered the publicly available Matlab implementation of this method by the authors. The method

from [1] is based on ODE method and uses non-parametric modeling of molecular kinetics and Bayesian analysis. The method can use both steady-state and time-series data. The experimental results of this methods demonstrated in [1] that this approach provides more accurate network structure predictions than other commonly used ODE and Bayesian methods. Therefore we prefer to use this method for comparison instead of considering ODE and DBN separately.

The model of Äijö and Lähdesmäki, which we call here ODE-DBN, is based on the commonly used first-order ODE model where the ODE describing the unknown function f responsible for the gene regulation is replaced by a first-order approximation of the rates of gene expression as

$$\frac{dx_i(t_k)}{dt} \approx \Delta x_i(t_k) = \frac{x_i(t_{k+1}) - x_i(t_k)}{t_{k+1} - t_k} \tag{3}$$

for a given set of measurement time points. The scalar $x_i(t)$ denotes the expression of gene i at time t and the vector $\mathbf{x}_i(t)$ denotes the expressions of genes that regulate gene i. The method uses Gaussian processes to learn the unknown regulation function from the data. The values of the unknown function f are modeled by a Gaussian process

$$f(x) \approx GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{4}$$

where GP denotes a Gaussian process, $m(x)$ is a mean function and $k(\mathbf{x}, \mathbf{x}')$ is a covariance function. It is further assumed that the mean function is identically zero. It is further assumed that there is a normal i.i.d. additive noise on the measurements and the predictions of the GP are computed analytically from the marginal likelihood, given the covariance matrices (more details in [1]). The method ODE-DBN has two goals: estimation of the non-parametric kinetic models and inference of the network structure. For a given model structure, the regulatory function can be estimated by means of a Gaussian process with the given covariance matrix. Bayesian model structure selection, where the goal is to choose explanatory variables $\mathbf{x_i}$ for each gene i, can be obtained via the marginal likelihood. The posterior probability of a given model can be obtained by applying Bayes theorem. The actual inference procedure is done separately for each gene in the network. That is, for each gene, the model of the ODE is being fit with different combinations of explanatory variables $\mathbf{x}$ and the posterior probabilities are computed. The posterior probabilities of network models are summarized using a square connection matrix, where the (i, j) element represents the posterior probability that gene j is regulated by gene i. Each element of the connection matrix can be computed by summing posterior probabilities of all networks that contain a directed connection from x_i to x_j. The method has an exponential computational complexity of order $O(n2^n)$ where n is the number of genes. The authors tested the method on small networks with five genes on yeast data with time series measurement with length 20 or 15. Another experiment was done with the network of 100 genes and was computed by means of distributed computing. The method was compared to both single ODE method (TSNI method from Bansal et al., [7]) and to single Bayesian networks (BANJO

method, [23]) as well as to the Bayesian networks from Zou and Conzen [25] with respect to common precision recall statistics. Based on the experiments in [1], the ODE-DBN method outperforms the inference of the TSNI method as well as of the method from Zou and Conzen on the time-series data and dynamic and static Bayesian networks on time-series and steady-state data.

4 Graphical Granger Lasso Models

Microarrays of gene expression data are represented by high-dimensional vectors and have short time series of the observations. The related parameter estimation problems are therefore ill-posed, so the straightforward application of the GC method is unfeasible [18]. As a remedy, the Granger method with a penalization method is applied.

Fujita et al. in [11] in 2007 applied a (multivariate) sparse vector autoregressive model SVAR with lasso regression, called a graphical GC.

Consider a graphical model with n variables (can be the number of genes), observed over T time points, and let d be the order of the VAR model or the effective number of time lags ($d = T - 1$). Let X^t denote the design matrix corresponding to t-th time point, and X_i^t be its i-th column.

The Lasso estimate of the graphical Granger model is found by solving the following estimation problem for $i = 1, \ldots, n$:

$$\arg \min_{\theta^t \in R^n} \|X_i^T - \sum_{t=1}^{d} X^{T-t}\theta^t\|_2^2 + \lambda \sum_{t=1}^{d}\sum_{j=1}^{n} |\theta_j^t| w_j^t. \tag{5}$$

The formula (5) has been studied in many variations, which mainly concern the form of the penalty function. Shojaie and Michalidis combined GC with the so called truncating lasso penalty or with so called adaptive lasso penalty [20] and proved their consistency in [21]. Lozano et al. studied the graphical Granger models with group Lasso penalization function in [17]. The complexity of the Granger Lasso depends on the complexity of the selected optimization method (non-convex optimization). The complexity of the optimization method applied in [20] is quadratical with respect to n, the number of genes. In this paper we will consider the VAR with Lasso regression from [3] and extend it to the multivariate case. The implementation of Lasso Granger from [3] has the computational quadratical complexity with respect to n.

5 Copula Granger method

Bahadori und Liu in [4] proved that Granger causality (i.e without any regularization) cannot be consistent in a high-dimensional regime, where insufficient number of observations is given. Utilizing the high dimensional advantages of Lasso regularization, they introduced the semi-parametric approach Copula Granger and showed its consistency in high dimensions as well as its ability to efficiently capture nonlinearity in the data.

The G-NPN model is defined as follows. One says a set of time series $X = (X_1, \ldots, X_n)$ has G-NPN distribution $G - NPN(X, B, F)$ if there exist functions $\{F_j\}_{j=1}^n$ such that $F_j(X_j)$ for $j = 1, \ldots, n$ are jointly Gaussian and can be factorized according to the VAR model with coefficients $B = \{\beta_{i,j}\}$. More specifically, the joint distribution for the transformed random variables $Z_j \hat{=} F_j(X_j)$ can be factorized as following

$$p_Z(z) = \mathcal{N}(z(1, \ldots, L)) \times \prod_{j=1}^{n} \prod_{t=L+1}^{T} p_{\mathcal{N}}(z_j(t); \sum_{t=1}^{n} \beta_{i,j}^T z_i^{t,Lagged}, \sigma_j)$$

where $p_{\mathcal{N}}(z; \mu, \sigma)$ is the Gaussian density function with mean μ and variance σ^2 and $z_i^{t,Lagged} = [z_i(t-L), \ldots, z_i(t-1)]$ is the history of z_i up to time t, L is the maximal time lag, and $\beta_{\mathbf{i,j}} = [\beta_{\mathbf{i,j}}(\mathbf{1}), \ldots, \beta_{\mathbf{i,j}}(\mathbf{L})]$ is the vector of coefficients modeling the effect of time series z_j on the target time series.

The causality is defined as follows: the time series z_j Granger causes z_i if at least one value in the coefficient vector $\beta_{\mathbf{j}}$ is nonzero by statistical significant sense.

Based on the copula method from [16], the G-NPN model aims to separate the marginal properties of the data from its dependency structure. The marginal distribution of the data can be efficiently estimated using the non-parametric techniques with exponential convergence rate [4].

Learning G-NPN models consists of three steps: (i) Find the empirical marginal distribution for each time series $\hat{F}_i$. (ii) Map the observations into the copula space as $\hat{f}_i(X_i^t) = \hat{\mu}_i + \hat{\sigma}_i \Phi^{-1}(\hat{F}_i(X_i^t))$. (iii) Find the GC among $\hat{f}_i(X_i^t)$. In practice the Winsorized estimator of the distribution function is used, to avoid the large numbers $\Phi^{-1}(0^+)$ and $\Phi^{-1}(1^-)$, [4].

Bahadori and Liu have proved that the convergence rate for Copula-Granger is the same as the one for Lasso. This suggests efficient Granger graph learning in high dimensions via Copula-Granger.

The Copula Granger Lasso method was tested with respect to the Granger method and Granger Lasso method on synthetic and experimetal data (Twitter application) with the best precision for Copula Granger Lasso method [4]. To our knowledge, any comparison of the Copula Granger Lasso method to Granger Lasso together with DBN has not been published yet.

6 Application of Granger Lasso Methods to Gene Regulatory Networks: Experimental Results

We investigated the three above discussed methods on genetic regulatory networks. Our selected data set is from the gene database of the genome-wide expression of cell cycle genes in human cancer cell lines (HeLa) analyzed by Whitfield et al. [22]. We used the preselected 19 genes, whose gene regulatory network was reconstructed based on the biological experiments o Li et al. [15]. This causal network we used as a benchmark structure for comparison of the discussed methods. The 19 genes, which play a substantial role at the human cancer

cell lines, have the following names: PCNA, NPAT, E2F1, CCNE1, CDC25A, CDKN1A, BRCA1, CCNF, CCNA2, CDC20, STK15, BUB1B, CKS2, CDC25C, PLK1, CCNB1, CDC25B, TYMS, DHFR. The causal structure for these genes identified by Li et al. was adopted from the figure from [17] and is in Figure 1.

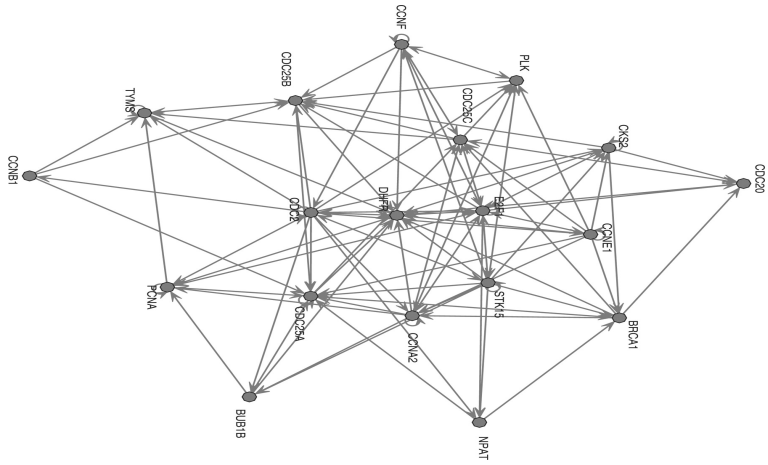


Fig. 1. Causal structure from biological experiments for 19 selected genes (adopted from [17])

The gene expressions in the database from [22] for the 19 genes were given for 48 time observations. In our experiments we used the following Matlab codes: for the inference measured by the combination ODE and Bayesian method (ODE-DBN) we used the code GP4GRN from Tarmo Äijö [1] which we extended with graphical outputs using publicly available Matlab graphical software Graphviz4Matlab Version 2.24. For experiments with Granger Lasso method we used the code from Bahadori [5] and extended it to the multivariate case (i.e from bivariate causal relationships to the multivariate ones). This we extended with the same graphical outputs using Graphviz4Matlab. Similarly we extended the code for Copula Granger Lasso method from the same source.

By TP, TN, FP, FN, respectively we denote the number of true positive outcomes, of true negative outcomes, of false positive outcomes and of false negative outcomes, respectively. We compared these methods for the HeLa data with respect to these classification performance measures: balanced classification rate $BCR = \frac{1}{2}(TP/(TP+FN) + TN/(TN+FP))$, precision $PREC = TP/(TP+FP)$, true positive rate $TPR = TP/(TP+FN)$ and classification accuracy $CA = (TP+TN)/(TP+TN+FP+FN)$. For each of these criteria we ran optimization processes for each of the method.

All the measures had significantly comparable results for the GP4GRN method and Granger Lasso method. It is not surprising, since the number of genes in the

network was small and the strength of the Granger Lasso causality is for large graphs.

The computational time of GP4GRN was very demanding, for each gene ca 2 min of real time at a PC workstation with 64-bit processor, which in our concrete case was 38 minutes. Granger Lasso require only a few seconds run for our simulation. The Copula Granger Lasso, which was running also a few seconds, had the best precision with respect to the above mentioned measures with significant differences to the other methods, see Table 1.

For example, the BCR measure for GP4GRN was 0.5774, for Granger Lasso 0.05789 and for Copula Granger Lasso 0.8006.

Concerning the validation of the methods, it can be done in many ways. The main goal of this paper was to compare the three methods on the well-known HeLa data from [22] which has been already applied for testing other methods [17], [20]. We tested the robustness of the methods with respect to noise with normal distribution having values of order $10^{-6}, 10^{-5}, \ldots, 10^{-1}$ which was added to the original data, so our statements about robustness concern only this type and orders of perturbation. Since the exact distribution of the error function of GE data is unknown, one can do similar tests for other types of noise, for example with Laplace distribution [19].

The precision of the inference results was the best for Copula Granger Lasso method, while the GP4GRN method issued into the overfitting and spurious results (the output GRN was overfitted with causal connections) already with low levels of noise on data. Concretely, the levels of random normal noise were of order $10^{-6}, 10^{-5}, \ldots, 10^{-1}$ on the time series with 48 time measurements of the data from [22]. The performance of the methods measured for BCR and CA measures of the non-perturbed data are summarized in the Table 1.

Table 1. Comparison of the precision of the three methods for various criteria

Method	GP4GRN	Granger Lasso	Copula Granger Lasso
BCR	0.5774	0.5789	0.8006
CA	0.7507	0.5789	0.8006
TP	95 overfitting	38	58

The performance of the methods measured by values for BCR, CA and TP of the perturbed data with levels of normal noise were of order $10^{-6}, 10^{-5}, \ldots, 10^{-1}$ is summarized in the Table 2.

Table 2. Comparison of the precision of the three methods with data perturbed by random noise of order (P. order) 10^{-6} and of 10^{-5} for various criteria; the values in the brackets are the deviations in per cent of the precision error from the precision error for unperturbed method.

P. order 10^{-6}	GP4GRN	Granger Lasso	Copula Granger Lasso
BCR	0.5850 (+1.31%)	0.6011(+3, 8%)	0.7895(−2, 6%)
CA	0.7507 (0%)	0.6011(+3, 8%)	0.7895(−2, 6%)
TP	95 overfitting	38	51
P. order 10^{-5}	GP4GRN	Granger Lasso	Copula Granger Lasso
BCR	0.5850(+1.31%)	0.5850(+0.7%)	0.7313(−8, 6%)
CA	0.7507(0%)	0.7507 (+29, 6%)	0.7313(−8, 6%)
TP	65 overfitting	96 overfitting	49

For the data perturbed at the order of $10^{-4}, \ldots, 10^{-1}$ of noise, both the GP4GRN and Granger Lasso methods exhibited a strong overfitting in the values of TP in the output GRNs, so the results of these methods were spurious. In case of Copula Granger Lasso, this method was robust to the perturbations of the data up to the order 10^{-1} and the precision results remained similar for the unperturbed data, see the Table 3.

Table 3. Comparison of change of the precision of Copula Granger Lasso with respect to BCR, CA and TP depending on the order of random noise on the data.

P. order	10^{-4}	10^{-3}	10^{-2}	10^{-1}
BCR	0.7867(−1.7%)	0.7837(−2.1%)	0.7701(−3.8%)	0.759(−6.4%)
CA	0.7867(−1.7%)	0.7837(−2.1%)	0.7701(−3.8%)	0.759(−6.4%)
TP	51	51	47	40

Figures 2 to 5 show the results of the codes for GP4GRN, Granger Lasso and Copula Granger Lasso respectively, in the gridded form; Figure 5 is the gridded graph corresponding to the graph in Figure 1.

7 Conclusion

We have tested the causality detection of three methods, a DBN-ODE based method, Granger Lasso and Copula Granger Lasso on gene regulatory networks with a genetic data set (HeLa) given by microarrays of gene expression data. The best method with respect to the precision and computational costs was Copula Granger Lasso, which as a non-linear method was able to detect the most causal relationships. Both Copula Granger Lasso and Granger Lasso methods are scalable methods and can be easily expanded to higher dimensions. Because of the low precision of GP4GRN and high computationally costs in our experiments, we conclude that GP4GRN is not feasible for large gene regulatory networks. This method seems to be more appropriate for modeling of local dynamics within a small genetic regulatory network, rather than for detection of general inference relationships in a large genetic regulatory network. We believe that the assumption of Gaussian processes, on which are dynamic Bayesian networks based, is in genetic regulatory networks violated and this violation is more transparent with increasing the size of the network.

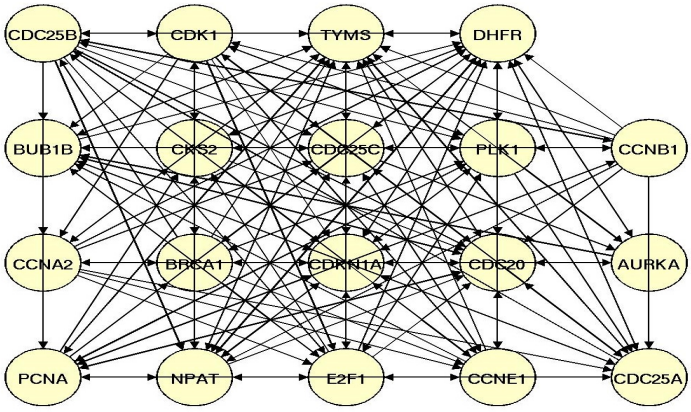


Fig. 2. Causal structure for selected 19 selected genes achieved by GP4GRN: the output graph shows overfitting (too many causal connections)

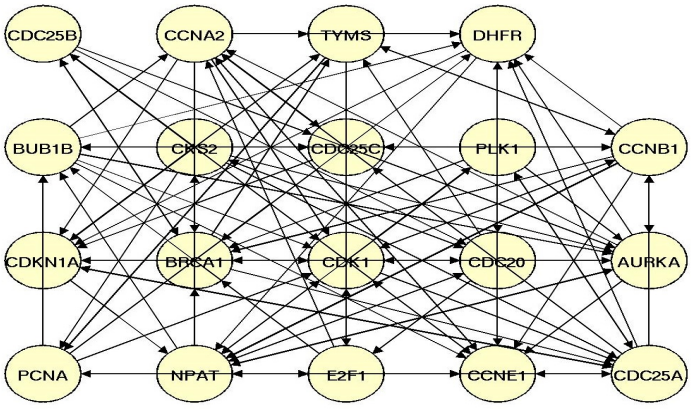


Fig. 3. Causal structure for selected 19 selected genes achieved by Granger Lasso method

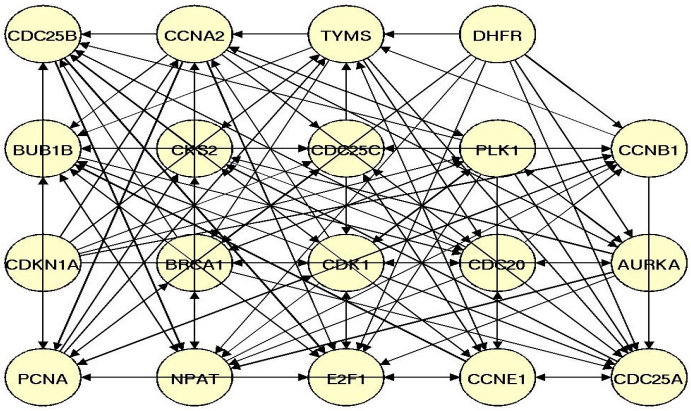


Fig. 4. Causal structure for selected 19 selected genes achieved by Copula Granger Lasso method

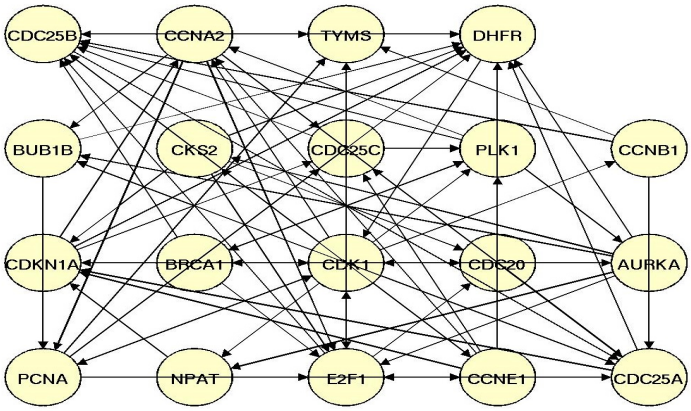


Fig. 5. Benchmark: the gridded graph of the causal structure for selected 19 selected genes from Figure 1

References

1. Äijö, T., Lahdesmäki, H.: Learning Gene Regulatory Networks from Gene Expression Measurements Using Non-Parametric Molecular Kinetics. Bioinformatics, no 22, Vol. 25 (2009)
2. Ancona, N., Marinazzo, D., Stramaglia, S.: Radial Basis Function Approach to Nonlinear Granger Causality of Time Series, Physical Review E 70 (2004)
3. Arnold, A., Liu, Y., Abe, N.:T emporal Causal Modeling with Graphical Granger Methods. In Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD-07) 2007
4. Bahadori, T., Y. Liu, Y.: An Examination of Large-Scale Granger Causality Inference. SIAM Conference on Data Mining (SDM'13) 2013
5. Website, `http// www-scf.usc.edu/∼mohammab/codes/codes.htm`
6. Bansal M., Belcastro, V., Ambesi-Impiombato, A, di Bernardo, D.: How to Infer Gene Networks from Expression Profiles. Mol. Syst. Biol., 3, 78 (2007)
7. Bansal, M., Della Gatta, G., di Bernardo, D. : Inference of Gene Regulatory Networks and Compound Mode of Action from Time Course Gene Expression Profiles. Bioinformatics, 22, 815822 (2006)
8. Broomhead, D.S., Lowe, D.: Multivariate Functional Interpolation and Adaptive Networks, Complex Systems 2, 321 - 355 (1988)
9. Cao, J. and Zhao, H.: Estimating dynamic models for gene regulation networks. Bioinformatics, 24, 16191624, 2008.
10. Chen, Y., Rangarajan, G., Feng, J., Ding, M.: Analyzing Mulitiple Non- Linear Time Series with Extended Granger Causality. Phys. Lett. A 324, 26-35 (2004)
11. Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Yamaguchi, R., Miyano, S., Ferreira, C.E. Modeling Gene Expression Regulatory Networks with the Sparse Vector Autoregressive Model, BMC Systems Biology, 1:37 (2007)
12. Granger, C.W.J.: Investigating Causal Relations by Econometric and Cross- Spectral Methods, Econometrica 37, 424-438 (1969)
13. Kaminski, M., Ding, M.: Truccolo, W.A., Bressler, S.L.: Evaluating Causal Relations in Neural Systems: Granger Causality, Directed Transfer Function and Statistical Assessment of Significance. Biol Cybern, 85, 14557 (2001)
14. Jensen, F.V.: An Introduction to Bayesian Networks, London, UCL Press (1996)
15. Li, X., Rao, S., Jiang, W., Li, C., Xiao, Y., Guo, Z., Zhang, Q., Wang, L., Du, L., Li, J., Li, L., Zhang, T., Wang, Q.K.: Discovery of Time-Delayed Gene Regulatory Networks Based on Temporal Gene Expression Profiling. BMC Bioinformatics, 7, 26 (2006)
16. Liu, H., Lafferty, J.D., Wasserman, T.: The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. Journal of Machine Learning Research, 10, 22952328 (2009)
17. Lozano, A.C., Abe, N., Liu, Y., Rosset, S.: Grouped Graphical Granger Modeling for Gene Expression Regulatory Networks Discovery. Vol. 25 ISMB, pp. i110-i118 (2009)
18. Marinazzo, D., Pellicoro, M., Stramaglia, S.: Kernel-Granger Causality and the Analysis of Dynamic Networks. Physical Review E, 77:056215 (2008)
19. Purdom, E., Holmes, S.P. Error Distribution for Gene Expression Data, Statistical Applications in Genetics and Molecular Biology, Vol. 4, 1, 16 (2005)
20. Shojaie, A. Michalidis, G.: Discovering Graphical Granger Causality Using the Truncating Lasso Penalty, Bioinformatics 26, 18: i517-i523 (2010)

21. Shojaie, A., Basu, S. Michalidis, G.: Adaptive Thresholding for Reconstructing Regulatory Networks from Time Course Gene Expression Data, Manuscript at www.biostat.washington.edu (2011)
22. Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown. P.O., Botstein, D.: Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. Moll.Biol.Cell., 13(6):1977-2000 (2002)
23. Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., Jarvis, E.D.: Advances to Bayesian Network Inference for Generating Causal Networks from Observational Biological Data. Bioinformatics, 20, 35943603 (2004)
24. Zou, C., Feng, J.: Granger Causality vs Dynamic Bayesian Network Inference: A Comparative Study. BMC Bioinformatics, 10:122 (2009)
25. Zou, M., Conzen, S.D.: A New Dynamic Bayesian Network (DBN) Approach for Identifying Gene Regulatory Networks from Time Course Microarray Data. Bioinformatics, 21, 7179 (2005)

On Approximate Fully Probabilistic Design of Decision Making Strategies

Miroslav Kárný

Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic. Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic,
school@utia.cas.cz

Abstract. An efficient support of a single decision maker is vital in constructing scalable systems addressing complex decision-making (DM) tasks. Fully probabilistic design (FPD) of DM strategies, an extension of dynamic Bayesian DM, provides a firm basis for such a support. The limited cognitive and evaluation resources of the supported decision maker cause that theoretically optimal solutions are realised only approximately. Thus, the truly efficient support has to include reliable means for constructing approximate solutions of DM subtasks. The current paper deals with the design of the approximately optimal DM strategy for a known environment model and adequately described DM preferences. The design relies on: **a)** the explicit minimiser found within FPD; **b)** randomised nature of the strategy provided by FPD.

Keywords decision making; Bayesian learning; minimum cross-entropy principle; fully probabilistic design of DM strategies; linear-quadratic DM

1 Introduction

The paper addresses a particular problem within a research aiming at creation of a systematic support of DM. The support has to respect that any real decision maker devotes a limited cognitive and evaluation resources to single DM problem and mostly has to use an approximation of theoretically optimal DM strategy. A design of such strategy is made here for a specific but widely applicable DM.

1.1 Basic Notions

This subsection fixes basic notions, which strongly vary over different DM-inspecting domains (statistics, economy, control theory, machine learning, etc.).

The decision maker designs and uses the DM *strategy* $\mathsf{s} = (\mathsf{s}_t)_{t\in\boldsymbol{t}} \in \mathbf{s}$, $\boldsymbol{t} = \{1, 2, \ldots, T\}$[1]. The DM *rules* s_t, forming the strategy s, are indexed by the discrete time t and map non-decreasing available *knowledge* $(k_t \in \boldsymbol{k_t})_{t\in\boldsymbol{t}}$, $\boldsymbol{k}_{t-1} \subseteq \boldsymbol{k}_t$, on *actions* $(a_t \in \boldsymbol{a_t})_{t\in\boldsymbol{t}}$, $\mathsf{s}_t : \boldsymbol{k}_{t-1} \to \boldsymbol{a}_t \neq \emptyset$.

[1] Throughout, $\boldsymbol{z}$ denotes a set of possible instances of z.

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

The considered knowledge increments are *data records* $d_t \in \boldsymbol{d_t} = \boldsymbol{k}_t \setminus \boldsymbol{k}_{t-1}$ ($\setminus$ denotes substraction of sets). The data record consists of the observed *environment response* r_t and of the applied action a_t. Thus, $d_t = (r_t, a_t)$ and $k_{t-1} = (d_{t-1}, \ldots, d_1, k_0)$, where k_0 denotes *prior knowledge*.

The DM strategy is designed with the aim to satisfy decision maker's DM preferences in the best possible way. They are expressed here as preferences with respect to possible closed-loop *behaviours* $b \in \boldsymbol{b}$

$$b = (g_t,\, a_t,\, k_{t-1}). \tag{1}$$

The part g_t collects variables up to the DM *horizon* T, which are considered by the decision maker but unavailable for choosing the action a_t.

1.2 FPD Formulation of DM Under Uncertainty

The addressed *DM under uncertainty* arises whenever the available knowledge k_{t-1} and the chosen action a_t do not allow the decision maker to determine uniquely the value of g_t, at least for some $t \in \boldsymbol{t}$. The classical axiomatisation [16, 1] of DM under uncertainty leads to Bayesian DM, which selects the optimal DM strategy s^{L} as a minimiser of an *expected loss*

$$\mathsf{s}^{\mathsf{L}} \in \mathrm{Arg}\min_{\mathsf{s}\in\mathbf{s}} \mathsf{E}_{\mathsf{s}}[\mathsf{L}|k_0] = \mathrm{Arg}\min_{(\mathsf{s}_t:\boldsymbol{k}_{t-1}\to\boldsymbol{a}_t)_{t\in\boldsymbol{t}}} \int_{\boldsymbol{b}} \mathsf{L}(b)\mathsf{f}_{\mathsf{s}}(b|k_0)\ \mathrm{d}b. \tag{2}$$

Bayesian DM requires specification of a *loss* $\mathsf{L} : \boldsymbol{b} \to (-\infty, \infty]$, *quantifying decision-maker's preferences*, and of the probability distribution of the possible behaviours $b \in \boldsymbol{b}$. It serves for evaluation of the conditional *expectations* $\mathsf{E}_{\mathsf{s}}[\cdot|k_0]$ for strategies $\mathsf{s} \in \mathbf{s}$ and it is given by the *probability density* (pd, $\mathsf{f}_{\mathsf{s}}(b|k_0)$) of behaviours b conditioned on a prior knowledge k_0 with respect to a measure $\mathrm{d}b$.

The exploited *fully probabilistic design (FPD)* of DM strategies [8, 20] quantifies DM preferences via an *ideal pd* $\mathsf{f}_I(b|k_0)$, which expresses desirability of possible behaviours $b \in \boldsymbol{b}$. FPD selects the strategy-dependent loss $\mathsf{L}_{\mathsf{s}} = \ln(\mathsf{f}_{\mathsf{s}}/\mathsf{f}_I)$. With this loss, the optimal DM strategy s^o becomes the minimiser of the Kullback-Leibler divergence $\mathcal{D}(\mathsf{f}_{\mathsf{s}}||\mathsf{f}_I)$ (KLD, [12])

$$\begin{aligned}\mathsf{s}^o \in \mathrm{Arg}\min_{\mathsf{s}\in\mathbf{s}} \mathsf{E}[\ln(\mathsf{f}_{\mathsf{s}}/\mathsf{f}_I)|k_0] &= \mathrm{Arg}\min_{\mathsf{s}\in\mathbf{s}} \int_{\boldsymbol{b}} \mathsf{f}_{\mathsf{s}}(b|k_0) \ln\left(\frac{\mathsf{f}_{\mathsf{s}}(b|k_0)}{\mathsf{f}_I(b|k_0)}\right)\ \mathrm{d}b \\ &= \mathrm{Arg}\min_{(\mathsf{s}_t:\boldsymbol{k}_{t-1}\to\boldsymbol{a}_t)_{t\in\boldsymbol{t}}} \mathcal{D}(\mathsf{f}_{\mathsf{s}}||\mathsf{f}_I).\end{aligned} \tag{3}$$

It is always possible to construct explicitly a FPD problem formulation, which is arbitrarily close to the given Bayesian DM task [10] and there are practically significant FPD tasks having no Bayesian counterpart [9].

1.3 The Addressed Problem and Solution Idea

The design of the optimal DM strategy (2) reduces to *dynamic programming* [2]. It gives *deterministic* strategy s^{L} generating actions, which are minimising

arguments in the functional equation evolving *value function* $\zeta^{\mathsf{L}}(k_t)$ against time

$$\zeta^{\mathsf{L}}(k_{t-1}) = \min_{a_t \in \boldsymbol{a_t}} \mathsf{E}[\zeta^{\mathsf{L}}(k_t)|a_t, k_{t-1}],\ \zeta^{\mathsf{L}}(k_T) = \mathsf{L}(b). \tag{4}$$

The existence of its analytical solution is an exception and a version of approximate dynamic programming [18] is inevitable.

The design of the optimal DM strategy s^o (3) is similar to (4) and also calls for an approximation in generic case. Its design is addressed in this paper. The proposed approximation exploits that: **i)** FPD has an explicit minimiser [8], **ii)** the *value function* $\zeta(k_t)$ *in FPD* solves a nonlinear integral equation, which determines the unique *randomised optimal DM strategy*.

The proposed approximation exploits the fact that the integral equation for the value function $\zeta(k_t)$ has to hold for any knowledge k_t even if it resulted from an application of non-optimal actions. Thus, it suffices to find a function, which solves the discussed equation on a sufficiently rich subset of $\boldsymbol{k}_t$ and then we surely get an approximation of the value function.

Technically, the integral equation is converted into a probabilistic model of a parametric approximation of the value function. Then, parameter estimates are updated via the Bayes rule on realised (non-optimal) past. The application of the corresponding randomised DM strategy makes the acquired knowledge sufficiently rich. The inevitable approximation errors can be and should be taken into account by employing stabilised forgetting [11]. This measure is advisable to any approximate sequential learning [6].

1.4 Layout

Section 2 specifies the assumptions delimiting the supported DM tasks and recalls the exploited information about FPD. Section 3 forming the core of the paper proposes the approximation of the optimal FPD strategy. Section 4 applies the general result to a widely used linear-quadratic dynamic DM (control, [13]). Section 5 provides a numerical illustration. Section 6 concludes the text.

2 Stationary FPD Caring about Observable Behaviour

In this preparatory section, the DM task leading to a stationary version of FPD is formulated and solved. For the sake of presentation simplicity, it deals with preferences specified for observable behaviours only. Thus, the part g_t of the behaviour b in (1) consists of yet unobserved environment responses $(r_\tau)_{\tau \geq t}$ and non-applied actions $(a_\tau)_{\tau > t}$.

The pd $\mathsf{f}_{\mathsf{s}}(b|k_0)$, describing behaviours $b \in \boldsymbol{b}$ under a DM strategy $\mathsf{s} \in \mathbf{s}$, can be factorised via the chain rule [15]

$$\mathsf{f}_{\mathsf{s}}(b|k_0) = \prod_{t \in \boldsymbol{t}} \underbrace{\mathsf{f}_{\mathsf{s}}(r_t|a_t, k_{t-1})}_{\text{environment model}} \times \underbrace{\mathsf{f}_{\mathsf{s}}(a_t|k_{t-1})}_{\text{DM-rule model}} \tag{5}$$

$k_t = ((r_t, a_t), k_{t-1}) = (d_t, k_{t-1})$ is the knowledge available for choosing a_{t+1}.

2.1 Considered Class of DM Tasks

The supported DM tasks are delimited by the following conditions.

- The environment model is a time-invariant, strategy-independent, state-space model $\mathsf{m}(x_t|a_t, x_{t-1})$ with the finite-dimensional real state $x_t \in \boldsymbol{x_t}$ and action $a_t \in \boldsymbol{a_t}$. The state x_t is a known image of its previous value x_{t-1} and of the observed data record $d_t = (r_t, a_t)$. Thus, for all $t \in \boldsymbol{t}$,

$$\mathsf{f}_{\mathsf{s}}(r_t|a_t, k_{t-1}) = \mathsf{m}(x_t|a_t, x_{t-1}).$$

- The initial state x_0 is assumed to be a part of the prior knowledge k_0.
- The DM rules s_t having the same model (5) are operationally equivalent and they are formally identified with their model. Thus, for all $t \in \boldsymbol{t}$,

$$\mathsf{s}_t(a_t|k_{t-1}) = \mathsf{f}_{\mathsf{s}}(a_t|k_{t-1}).$$

- The ideal pd $\mathsf{f}_I(b|k_0)$ only cares about preferences on the observed states and actions and thus it can be factorised as follows

$$\mathsf{f}_I(b|k_0) = \prod_{t\in\boldsymbol{t}} \mathsf{m}_I(x_t|a_t, x_{t-1})\mathsf{s}_I(a_t|x_{t-1}), \tag{6}$$

  where the given pds m_I, s_I in (6) are assumed to be time-invariant.
- The design is performed for the DM horizon $T \to \infty$.

2.2 Optimal DM Strategy To Be Approximated

Proposition 1 (Solution of Stationary FPD) *Let a stabilising DM strategy $\mathsf{s}^s \in \boldsymbol{s}$ exist, which means that*

$$c_{\mathsf{s}^s} = \lim_{T\to\infty} \frac{1}{T}\mathcal{D}(\mathsf{f}_{\mathsf{s}^s}||\mathsf{f}_I) < \infty.$$

Then, the optimal DM strategy, minimising the KLD (3), is stabilising, stationary $\mathsf{s}^o(b|k_0) = \prod_{t\in\boldsymbol{t}} \mathsf{s}^o(a_t|x_{t-1})$ and determined by the time-invariant DM rule

$$\mathsf{s}^o(a_t|x_{t-1}) = \frac{\mathsf{s}_I(a_t|x_{t-1})\exp\left[-\mathsf{D}(a_t, x_{t-1}) - \mathsf{H}(a_t, x_{t-1})\right]}{\underbrace{\int_{\boldsymbol{a_t}} \mathsf{s}_I(a_t|x_{t-1})\exp\left[-\mathsf{D}(a_t, x_{t-1}) - \mathsf{H}(a_t, x_{t-1})\right]\ \mathrm{d}a_t}_{\exp[-\mathsf{h}(x_{t-1})]}} \tag{7}$$

$$\mathsf{D}(a_t, x_{t-1}) = \int_{\boldsymbol{x_t}} \mathsf{m}(x_t|a_t, x_{t-1}) \ln\left(\frac{\mathsf{m}(x_t|a_t, x_{t-1})}{\mathsf{m}_I(x_t|a_t, x_{t-1})}\right)\ \mathrm{d}x_t \geq 0 \tag{8}$$

$$\mathsf{H}(a_t, x_{t-1}) = \int_{\boldsymbol{x_t}} \mathsf{m}(x_t|a_t, x_{t-1})\mathsf{h}(x_t)\ \mathrm{d}x_t \geq -c \tag{9}$$

$$c = \lim_{T\to\infty} \frac{1}{T}\mathcal{D}(\mathsf{f}_{\mathsf{s}^o}||\mathsf{f}_I) \in [0, c_{\mathsf{s}^s}] \ \Rightarrow\ \zeta(x_t) = c + \mathsf{h}(x_t) \geq 0. \tag{10}$$

Proof It is omitted as it follows steps used in proving standard dynamic programming [2]. It only exploits the fact that the KLD reaches its minimum zero value for coinciding arguments. You can consult [3] containing the proof concerning general case with preferences specified also for unobserved states. □

Remarks

- The functional equations (7) – (9) rarely have an analytical solution. Their approximate solution is proposed in Section 3.
- The function $\exp[-\mathsf{h}(x)]$ is proportional to the stationary pd of the state when the optimal strategy is used. It is seen from (7) and the conditioning rule $\mathsf{pd}(a|x) = \mathsf{pd}(a,x)/\mathsf{pd}(x)$. This interpretation should be respected when selecting the set of functions in which its approximation is searched for.
- The decisive function $\mathsf{h}(x_t)$ is the shifted value of the non-negative value function $\zeta(x_t)$ (10). Its non-negativity implies $\mathsf{H}(a_t, x_{t-1}) \geq -c$ (9).
- The function $\mathsf{D}(a_t, x_{t-1})$ (8) is non-negative as it is the conditional KLD of the environment model m from its ideal counterpart m_I.
- All involved functions are assumed to be time invariant. The time-invariance of the environment model $\mathsf{m}(x_t|a_t, x_{t-1})$ is asymptotically guaranteed if it is obtained as the predictive pd resulting from Bayesian learning, [15]. Thus, the presented treatment is extendable to this case.

3 Approximation of the Optimal Strategy

Here, the approximation of the optimal DM strategy is searched for. It consists of approximations of the functions D, H defining the pd s^o (7), cf. Proposition 1.

The conditional KLD $\mathsf{D}(a_t, x_{t-1}) = \int_{\boldsymbol{x_t}} \mathsf{m}(x_t|a_t, x_{t-1}) \ln\left(\frac{\mathsf{m}(x_t|a_t,x_{t-1})}{\mathsf{m}_I(x_t|a_t,x_{t-1})}\right) \mathrm{d}x_t$ in (7) is time-invariant and can be, at least approximately, evaluated off-line. Thus, the approximation concerns primarily the shifted value function $\mathsf{h}(x_t)$ and its expectation $\mathsf{H}(a_t, x_{t-1})$ with respect to the environment model

$$\mathsf{H}(a_t, x_{t-1}) = \mathsf{E}[\mathsf{h}|a_t, x_{t-1}] = \int_{\boldsymbol{x_t}} \mathsf{m}(x_t|a_t, x_{t-1})\mathsf{h}(x_t)\ \mathrm{d}x_t.$$

3.1 Technical Elaboration

The proposed approximation uses:

- parametric approximation of $\mathsf{h}(x) \approx \mathsf{h}(x, \Theta)$ inducing the approximation
$$\mathsf{H}(a, x_{t-1}) = \mathsf{E}[\mathsf{h}|a, x_{t-1}] \approx \mathsf{H}(a, x_{t-1}, \Theta) = \mathsf{E}[\mathsf{h}(\cdot, \Theta)|a, x_{t-1}, \Theta];$$
- mean-value theorem applied to the integral over $\boldsymbol{a}_t$, Proposition 1 & (11);
- decomposition of expectation $\mathsf{H}(a, x_{t-1}, \Theta) = \mathsf{E}[\mathsf{h}(\cdot, \Theta)|a, x_{t-1}, \Theta]$ in $\mathsf{h}(x_t, \Theta)$ and *innovation* $\varepsilon(a, x_{t-1}, \Theta)$: $\mathsf{H}(a, x_{t-1}, \Theta) = \mathsf{h}(x_t, \Theta) + \varepsilon(x_t, a, x_{t-1}, \Theta)$ [15];
- minimum KLD (cross-entropy) principle [17], which extends a partial information about a pd into the complete pd.

Proposition 1 implies that the function $\mathsf{h}(x, \Theta)$ should solve the equation

$$\exp[-\mathsf{h}(x_{t-1}, \Theta)] = \int_{\boldsymbol{a_t}} \mathsf{s}_I(a_t|x_{t-1}) \exp\left[-\mathsf{D}(a_t, x_{t-1}) - \mathsf{H}(a_t, x_{t-1}, \Theta)\right] \, \mathrm{d}a_t, \quad (11)$$

which *has to hold for any state* $x_{t-1} \in \boldsymbol{x_{t-1}}$ *even if it resulted from use of non-optimal past actions*. An application of mean-value theorem to this equation, introduction of innovations and logarithmic transformation provide

$$\begin{aligned} -\mathsf{h}(x_{t-1}, \Theta) &= \underbrace{\ln \left[\int_{\boldsymbol{a_t}} \mathsf{s}_I(a_t|x_{t-1}) \exp[-\mathsf{D}(a_t, x_{t-1})] \, \mathrm{d}a_t \right]}_{\phi(x_{t-1})<0} \\ &\quad + \ln \left(\exp \left[- \int_{\boldsymbol{x}_t} \mathsf{m}(x_t|\underline{a}(x_{t-1}, \Theta), x_{t-1}) \mathsf{h}(x_t, \Theta) \, \mathrm{d}x_t \right] \right) \\ &= \phi(x_{t-1}) - \mathsf{h}(x_t, \Theta) + \varepsilon_t(x_t, \underline{a}(x_{t-1}, \Theta), x_{t-1}, \Theta), \end{aligned} \quad (12)$$

where $\underline{a}(x_{t-1}, \Theta)$ denotes the action resulting from the mean-value theorem.

The time and Θ invariant function $\phi(x_{t-1})$ is *negative*. It can be prepared off-line and thus it is fully determined by the knowledge k_{t-1}. The innovations

$$\varepsilon_t(x_t, \underline{a}(x_{t-1}, \Theta), x_{t-1}, \Theta) = \mathsf{h}(x_t, \Theta) - \int_{\boldsymbol{x}_t} \mathsf{m}(x_t|\underline{a}(x_{t-1}, \Theta), x_{t-1}) \mathsf{h}(x_t, \Theta) \, \mathrm{d}x_t$$

are, by construction, zero mean and uncorrelated with their past values, [15],

$$\int_{\boldsymbol{x_t}} \varepsilon_t(x_t, \underline{a}(x_{t-1}, \Theta), x_{t-1}, \Theta) \mathsf{m}(x_t|\underline{a}(x_{t-1}, \Theta), x_{t-1}) \, \mathrm{d}x_t = 0.$$

This property and (12) imply that the positive random value of the value function $\zeta(x_t, c, \Theta) = c + \mathsf{h}(x_t, \Theta)$ has conditional expectation $\zeta(x_{t-1}, c, \Theta) + \phi(x_{t-1}) \in (0, \infty)$, i.e.

$$\mathsf{E}[\zeta(x_t, c, \Theta) - \zeta(x_{t-1}, c, \Theta)|\zeta(x_{t-1}, \Theta), k_{t-1}, \Theta] = \phi(x_{t-1}) < 0. \quad (13)$$

The minimum KLD principle [17] completes this information about the conditional expectation into the exponential distribution

$$\mathsf{f}(\zeta(x_t, c, \Theta)|\zeta(x_{t-1}, c, \Theta), k_{t-1}, c, \Theta) = \frac{\exp\left[-\frac{\zeta(x_t, c, \Theta)}{\zeta(x_{t-1}, c, \Theta) + \phi(x_{t-1})}\right]}{\zeta(x_{t-1}, c, \Theta) + \phi(x_{t-1})}. \quad (14)$$

In order to avoid discussion of non-linear Bayesian learning, which is out our scope, we also adopt the approximation

$$\zeta(x_{t-1}, c, \Theta) = c + \mathsf{h}(x_{t-1}, \Theta) \approx (\hat{c}_{t-1} + \mathsf{h}(x_{t-1}, \hat{\Theta}_{t-1}))\chi(c + \mathsf{h}(x_{t-1}, \hat{\Theta}_{t-1}) \geq 0), \quad (15)$$

where $\hat{c}_{t-1}, \hat{\Theta}_{t-1}$ are point estimates of c, Θ based on k_{t-1} and $\chi(\cdot)$ is an indicator function of the set in its argument. In this way, the parametric model relating $\zeta(x_t, c, \Theta)$ to the knowledge k_{t-1} and unknown c, Θ is obtained

$$\begin{aligned}
\mathsf{f}(\zeta(x_t, c, \Theta)|k_{t-1}, c, \Theta) &= \alpha_{t-1}^{-1} \exp\left[-(c + \mathsf{h}(x_t, \Theta))\alpha_{t-1}\right] \chi(c + \mathsf{h}(x_{t-1}, \hat{\Theta}_{t-1}) \geq 0) \\
\alpha_{t-1}^{-1} &= \hat{c}_{t-1} + \mathsf{h}(x_{t-1}, \hat{\Theta}_{t-1}) + \phi_{t-1} \qquad (16) \\
&= \hat{c}_{t-1} + \mathsf{h}(x_{t-1}, \hat{\Theta}_{t-1}) + \ln\left(\int_{\boldsymbol{a_t}} \mathsf{s}_I(a_t|x_{t-1}) \exp[-\mathsf{D}(a_t, x_{t-1})] \, \mathrm{d}a_t\right).
\end{aligned}$$

The gained parametric model is used in Bayesian learning, which evolves the posterior pd $\mathsf{f}(c, \Theta|k_t)$ on the unknown c, Θ. The evolution has the form, cf. (14), (15) and (16)

$$\mathsf{f}(c, \Theta|k_t) \propto \mathsf{f}(c, \Theta|k_{t-1}) \exp\left[-(c + \mathsf{h}(x_t, \Theta))\alpha_{t-1}\right] \qquad (17)$$

Non-negativity of $\zeta(x_t, c, \Theta)$ and its conditional expectation is the key information about c, which subtracts in (13). It implies that $c \geq \max_{\tau \leq t-1}(-\mathsf{h}_\tau)$.

3.2 Discussion

Here, explanatory comments are added to the above technical manipulations. In summary, the proposed learning of DM strategy relies on the heuristic steps:

- *The parametric expression $c + \mathsf{h}(x, \Theta)$ of the value function $\zeta(x)$ is supposed to approximate well the value function for some $\Theta \in \boldsymbol{\Theta}$.*
  This assumption can be met by an appropriate choice of the function $\mathsf{h}(x, \Theta)$, by exploiting a "universal approximation property" [4]. In this respect, it is important that, by construction, the approximated function $\exp[-\mathsf{h}(x)]$ is proportional to the stationary distribution of x_t for the optimal strategy.
- *The pd of the approximate value function $\zeta(x_t, c, \Theta) = c + \mathsf{h}(x_t, \Theta)$ has been derived via maximum entropy while neglecting a direct information about the environment model $\mathsf{m}(x_t|a_t, x_{t-1})$.*
  This assumption is unrestrictive as the scalar $\mathsf{h}(x_t, \Theta)$ depends on multivariate x_t and the adopted DM strategy in a complex unknown way. It means that a negligible amount of useful and truly available knowledge is neglected. The dependence on the environment model and the ideal counterpart of the applied DM strategy projects into the weight α_{t-1} (16).
- *The crude approximation (15) is adopted.*
  This assumption is generally unnecessary. It has helped us in suppressing the need to discuss non-linear Bayesian learning, which is out of our scope.

The following points are also worth discussing.

- The posterior pd on Θ serves for approximating the optimal DM strategy, i.e. for estimation of $\mathsf{H}(a_{t+1}, x_t) = \mathsf{E}[\mathsf{h}(\cdot)|a_{t+1}, x_t]$. It hints to take

$$\begin{aligned}
\mathsf{H}(a_{t+1}, x_t) &\approx \int_{\boldsymbol{x_{t+1}}} \mathsf{m}(x_{t+1}|a_{t+1}, x_t) \int_{\boldsymbol{\Theta}} \mathsf{h}(x_{t+1}, \Theta)\mathsf{f}(\Theta|k_t) \, \mathrm{d}\Theta \, \mathrm{d}x_{t+1} \\
&\approx \int_{\boldsymbol{x_{t+1}}} \mathsf{m}(x_{t+1}|a_{t+1}, x_t)\mathsf{h}(x_{t+1}, \hat{\Theta}_t) \, \mathrm{d}x_{t+1} = \hat{\mathsf{H}}(a_{t+1}, x_t).
\end{aligned}$$

The last approximate equality delimits the needed point estimate $\hat{\Theta}_t$ of Θ.

- The function $-\mathsf{D}(a_{t+1}, x_t) - \acute{\mathsf{H}}(a_{t+1}, x_t|k)$ is immediately used for generating a_{t+1}, see (7). After applying a_{t+1} and recording x_{t+1}, the learning process can continue. The randomised nature of the optimal DM strategy in FPD makes the used DM strategy explorative. The higher is uncertainty about the parameter Θ the flatter is the pd used for generating a_{t+1}. It indicates qualitatively plausible variations of the exploration extent.
- The actions $\underline{a}(x_{t-1}, \Theta)$ considered in approximate evaluations (12) origin from the ideal counterpart of the DM strategy, neither from the optimal nor the used DM strategy. The basic idea of the construction indicates that the learning running on non-optimal states, caused by the applied non-optimal actions, is counteracted by the weight α_{t-1} (16) determined by the combination environment model – ideal strategy.
- It is possible to introduce additional weighting suitable whenever learning contains some approximation error [6]. We shall not employ it in order to check whether the proposed learning copes with the "incorrect data".

4 Application to Linear-Gaussian DM

The influence and extent of the applicability of adopted approximations as well as of heuristic assumptions, see Section 3, are yet unclear. Thus, it makes sense to check the proposed procedure on a case with a known solution. Linear-Gaussian DM treated here serves to this purpose. It is given by the following assumptions.

- The environment model is linear Gaussian

$$\mathsf{m}(x_t|a_t, x_{t-1}) = \mathcal{N}_{x_t}(Ax_{t-1} + Ba_t, R)$$
$$\mathcal{N}_x(\mu, R) = |2\pi R|^{-0.5} \exp\left[-0.5(x-\mu)' R^{-1}(x-\mu)\right],$$

  where A, B, determining its conditional expectation, as well as positive definite covariance $R > 0$ are known matrices of dimensions compatible with the vectorial state x_t and action a_t. $'$ denotes transposition.
- The ideal counterpart of the environment model is chosen also Gaussian

$$\mathsf{m}_I(x_t|a_t, x_{t-1}) = \mathcal{N}_{x_t}(0, R).$$

  It reflects the wish to push the state to zero (so called regulation problem, [13]). The equality of covariances of the environment model and its ideal counterpart respects the fact that R represents the lowest reachable covariance. The ideal counterpart of the DM strategy is chosen also Gaussian

$$\mathsf{s}_I(a_t|x_{t-1}) = \mathcal{N}_{a_t}(0, q).$$

  This ideal pd represents the wish to spare acting energy $0.5a_t' q^{-1} a_t$.

In this case, the exact solution of FPD is known, [5]. It holds

$\exp[-\mathsf{h}(x)] = \mathcal{N}_x(0, S)$ the covariance $S > 0$ is known as Riccati matrix
$\mathsf{s}^o(a_t|x_{t-1}) = \mathcal{N}_{a_t}(-L'x_{t-1}, Q)$,

where the matrices $S > 0$, $Q > 0$ as well the matrix L (control law) are determined by parameters of the environment model and those of ideal pds.

The proposed procedure specialises to this case as follows.

The function $\mathsf{D}(a_t, x_{t-1})$ (8), the conditional KLD of the pd m from the pd m_I, can be computed analytically, e.g. [7],

$$\mathsf{D}(a_t, x_{t-1}) = 0.5\,(Ax_{t-1} + Ba_t)'\,R^{-1}\,(Ax_{t-1} + Ba_t)\,.$$

The function $\phi(x_{t-1})$ (12) is also given analytically

$$\begin{aligned}\phi(x_{t-1}) &= \ln\left[\int_{\boldsymbol{a_t}} \mathsf{s}_I(a_t|x_{t-1})\exp[-\mathsf{D}(a_t, x_{t-1})]\;\mathrm{d}a_t\right] \\ &= 0.5\left[\ln\left(\left|I + qB'R^{-1}B\right|\right) - x'_{t-1}A'(R + BqB')^{-1}Ax_{t-1}\right],\end{aligned} \tag{18}$$

where I denotes unit matrix. The neat final form is obtained by employing so called Woodbury formula.

The next approximation $\mathsf{h}(x, \Theta)$ of $\mathsf{h}(x)$ admits the needed comparisons

$$\exp[-\mathsf{h}(x,\Theta)] = \mathcal{N}_x(0,\Theta),\ \Theta > 0\ \Rightarrow\ \mathsf{h}(x,\Theta) = 0.5(\ln(|\Theta|) + x'\Theta^{-1}x) + \textit{constant}. \tag{19}$$

The forms of $\phi(x_{t-1})$ (18) and of $\mathsf{h}(x)$ (19) specialise the learning (17) to

$$\begin{aligned}&\mathsf{f}(\Theta|k_t) \propto \mathsf{f}(\Theta|k_{t-1})\exp\left[-(c + \mathsf{h}(x_t,\Theta))\alpha_{t-1}\right]\chi(c + \mathsf{h}(x_t,\hat{\Theta}_t) \geq 0) \\ &\propto |\Theta|^{-0.5\nu_t}\exp\left[-(c\nu_t + 0.5\mathrm{tr}\left(\Theta^{-1}V_t\right))\right]\chi(c \geq \bar{c}_t) \\ &\bar{c}_t = \max\left[\bar{c}_{t-1}, -0.5\left(\ln\left(\left|\hat{\Theta}_t\right|\right) + x'_t\hat{\Theta}_t x_t\right)\right] \\ &V_t = V_{t-1} + \alpha_{t-1}x_t x'_t,\quad \nu_t = \nu_{t-1} + \alpha_{t-1},\quad \alpha_{t-1} = \\ &\frac{2}{2\hat{c}_{t-1} + \ln\left(|I + qB'R^{-1}B|\,|\hat{\Theta}_{t-1}|\right) + x'_{t-1}(\hat{\Theta}^{-1}_{t-1} - A'(R + BqB')^{-1}A)x_{t-1}} \\ &V_0 > 0, \nu_0, \bar{c}_0 > 0 \text{ determine the prior pd in the self-reproducing form} \\ &\mathsf{f}(c,\Theta|k_0) \propto |\Theta|^{-0.5\nu_0}\exp\left[-\left(c\nu_0 + 0.5\mathrm{tr}\left(\Theta^{-1}V_0\right)\right)\right]\chi(c \geq \bar{c}_0).\end{aligned} \tag{20}$$

The final formula (20) is intuitively plausible as:

- Θ should estimate the Riccati matrix, which is covariance matrix of the state in the closed loop with the optimal DM strategy. The proposed learning provides such estimate in the form of the *weighted* covariance. The adopted maximum-likelihood estimates $\hat{c}_t$, $\hat{\Theta}_t$ of c, Θ for the knowledge k_t are

$$\hat{c}_t = \bar{c}_t,\ \hat{\Theta}_t = \frac{V_t}{\nu_t}. \tag{21}$$

- The weight of the dyad increment $x_t x'_t$ is the higher the closer is x_{t-1} to 0.
- The relative closeness of x_{t-1} to zero is determined by relations between properties of the controlled environment (A, B, R) and the cost q of actions.

The approximately optimal DM strategy corresponding to (7) and (21) is

$$\begin{aligned}
&\hat{\mathsf{s}}_{t+1}(a_{t+1}|x_t) \propto \mathcal{N}_{a_t}(0,q)\exp[-\mathsf{D}(a_{t+1},x_t)-\hat{\mathsf{H}}(a_{t+1},x_t)]\\
&\propto \exp\left[-0.5\left(a'_{t+1}q^{-1}a_{t+1}+(Ax_t+Ba_{t+1})'\left(\hat{\Theta}_t^{-1}+R^{-1}\right)(Ax_t+Ba_{t+1})\right)\right]\\
&=\mathcal{N}_{a_{t+1}}(-\hat{L}_t x_t,\hat{Q}_t)\\
&\hat{Q}_t=(q^{-1}+B'(\hat{\Theta}_t^{-1}+R^{-1})B)^{-1},\quad \hat{L}_t=\hat{Q}_t B'(\hat{\Theta}_t^{-1}+R^{-1})A.
\end{aligned}$$

Structurally, it corresponds with the optimal DM strategy. Limited experimental experience indicates that the procedure often approaches the optimal DM strategy. The approximation quality can be improved by employing the stabilised forgetting counteracting the accumulation of approximation errors [6].

5 Numerical Example

This section illustrates numerically behaviour of the algorithm in linear-Gaussian case described in the previous section. The environment model is specified by

$$A=\begin{bmatrix}0.70 & -0.30 & 0.80\\ 0.70 & 0.95 & 0.20\\ 0.20 & 0.00 & 0.90\end{bmatrix},\quad B=\begin{bmatrix}1.00\\ 0.50\\ 0.00\end{bmatrix},\quad R=\begin{bmatrix}1.00 & -0.20 & 0.20\\ -0.20 & 0.29 & 0.11\\ 0.20 & 0.11 & 0.17\end{bmatrix},$$

where the covariance is positive definite as it was generated as product of its Choleski factors. In the inspected regulation problem and for the scalar action, preferences are specified just via the ideal action variance $q = 10$. The results are shown for $T = 100$ allowing to display time trajectories. Non-presented runs up to $T = 50000$ confirmed stability of the solution and of the closed DM loop.

The optimal stationary strategy is given by the Gaussian pd

$$\mathcal{N}_{a_t}(-[0.817,\ 0.788,\ -0.409]x_{t-1},0.069),$$

while the proposed procedure provides

$$\mathcal{N}_{a_t}(-[0.788,\ 0.781,\ -0.531]x_{t-1},0.068).$$

Closeness of sample moments of states and actions with optimal and approximate strategy indicates that the found strategy approximates well the optimal one. Importantly, the essentially same approximate strategy has been obtained

$$\mathcal{N}_{a_t}(-[0.794,\ 0.787,\ -0.512]x_{t-1},0.064)$$

when the learning run with the optimal controller. The learning was also run with enforced zero action. It lead to the controller

$$\mathcal{N}_{a_t}(-[0.746,\ 0.722,\ -0.622]x_{t-1},0.071),$$

with poorer, but still quite-reasonable, closed-loop behaviour. The mild deterioration of quality can be attributed to the lack of exploration.

The possibility to learn reasonable strategy from non-optimal closed-loop behaviour is the focal feature of the example as it indicates that the adopted concept is sound. Numerically, it manifests on time course of the weights α_t (16). They become (relatively) large if the closed-loop behaviour is locally (even by chance) close to the optimal one. Fig. 1 illustrates this statement by showing time-courses of this weight in all described configurations of experiments.

For completeness, Figure 2 shows state evolutions in all configurations.

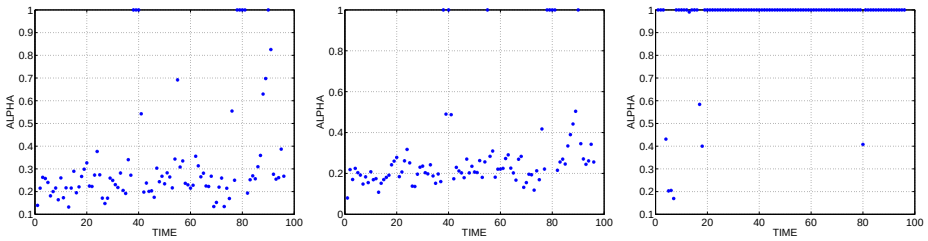


Fig. 1. Time courses of α_t (16). The left one corresponds to the closed-loop with the proposed strategy. The middle one reflects learning with the optimally closed loop. The right one concerns learning while action is fixed at zero value: the relatively high values of α_t are caused by the lack of informative data.

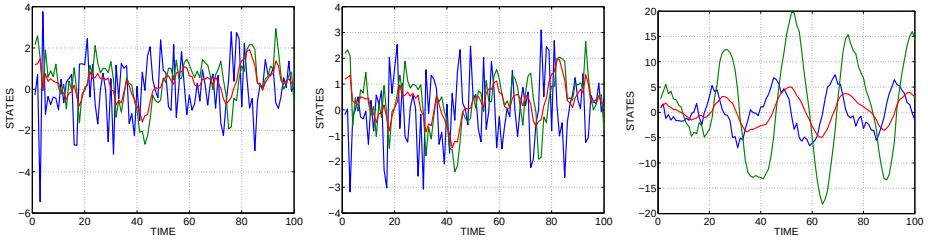


Fig. 2. Time courses of state x_t. The left one corresponds to the closed-loop with the proposed strategy. The middle one concerns the optimally closed loop. The right one concerns the loop with the action fixed at zero value. *Scales reflect regulation quality.*

6 Conclusions

The paper tries tailor approximate dynamic programming to fully probabilistic design of DM strategies. The presented preliminary results indicate that the addressed problem is solvable in the outlined way but otherwise the paper is an open-ended story. The logical necessity of respective development steps is the weakest conceptual point. Technically, the future work should focus on:

- analysing the proposed solution (at least via simulations);
- guiding in parameterisations of the function $\exp[-\mathsf{h}(x_t)]$ (universal approximation [4], probably by dynamic mixtures [14, 19]);
- combining with Bayesian learning of the environment model, [15];
- addressing the DM problem with indirectly observed state, [8];
- applying forgetting as a universal counter-measure against accumulation of approximation errors [6].

Acknowledgements This research has been supported by GAČR 13-13502S. The text has been substantially influenced by discussions with Dr. T.V. Guy.

References

1. Berger, J.: Statistical Decision Theory and Bayesian Analysis. Springer, New York (1985)
2. Bertsekas, D.: Dynamic Programming and Optimal Control. Athena Scientific, Nashua, US (2001), 2nd edition
3. Guy, T.V., Kárný, M.: Stationary fully probabilistic control design. In: Filipe, J., Cetto, J.A., Ferrier, J.L. (eds.) Proc. of the Second Int. Conference on Informatics in Control, Automation and Robotics. pp. 109–112. INSTICC, Barcelona (2005)
4. Haykin, S.: Neural Networks: A Comprehensive Foundation. Macmillan, New York (1994)
5. Kárný, M.: Towards fully probabilistic control design. Automatica 32(12), 1719–1722 (1996)
6. Kárný, M.: Approximate Bayesian recursive estimation. Inf. Sci. (2013), submitted
7. Kárný, M., Böhm, J., Guy, T.V., Jirsa, L., Nagy, I., Nedoma, P., Tesař, L.: Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer (2006)
8. Kárný, M., Guy, T.V.: Fully probabilistic control design. Systems & Control Letters 55(4), 259–265 (2006)
9. Kárný, M., Guy, T.: On support of imperfect Bayesian participants. In: Guy, T., Kárný, M., Wolpert, D. (eds.) Decision Making with Imperfect Decision Makers, vol. 28. Springer, Berlin (2012), Intelligent Systems Reference Library
10. Kárný, M., Kroupa, T.: Axiomatisation of fully probabilistic design. Information Sciences 186(1), 105–113 (2012)
11. Kulhavý, R., Zarrop, M.B.: On a general concept of forgetting. Int. J. of Control 58(4), 905–924 (1993)
12. Kullback, S., Leibler, R.: On information and sufficiency. Annals of Mathematical Statistics 22, 79–87 (1951)
13. Meditch, J.: Stochastic Optimal Linear Estimation and Control. Mc. Graw Hill (1969)
14. Nagy, I., Suzdaleva, E., Kárný, M., Mlynářová, T.: Bayesian estimation of dynamic finite mixtures. Int. Journal of Adaptive Control and Signal Processing 25(9), 765–787 (2011)
15. Peterka, V.: Bayesian system identification. In: Eykhoff, P. (ed.) Trends and Progress in System Identification, pp. 239–304. Pergamon Press, Oxford (1981)
16. Savage, L.: Foundations of Statistics. Wiley, New York (1954)
17. Shore, J., Johnson, R.: Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy. IEEE Tran. on Inf. Th. 26(1), 26–37 (1980)
18. Si, J., Barto, A., Powell, W., Wunsch, D. (eds.): Handbook of Learning and Approximate Dynamic Programming. Wiley-IEEE Press, Danvers (May 2004)
19. Titterington, D., Smith, A., Makov, U.: Statistical Analysis of Finite Mixtures. John Wiley, New York (1985)
20. Todorov, E.: Linearly-solvable Markov decision problems. In: Schölkopf, B., et al (eds.) Advances in Neural Inf. Processing, pp. 1369 – 1376. MIT Press, NY (2006)

Preliminaries of probabilistic hierarchical fault detection

Ladislav Jirsa, Lenka Pavelková, and Kamil Dedecius

Department of Adaptive Systems,
Institute of Information Theory and Automation,
Academy of Sciences of the Czech Republic,
Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic
{jirsa,pavelkov,dedecius}@utia.cas.cz
http://www.utia.cas.cz

Abstract. The paper proposes a novel probabilistic fault detection and isolation (FDI) system that enables to evaluate dynamically the industrial system condition (health) at any level of its functional hierarchy. The investigated industrial system is considered as a set of interconnected individual components. Each component acts in its noisy environment as an imperfect participant, more or less dependent on neighbouring components and, in turn, influencing some others. The nature of the problem prevents us from expressing sufficiently hard propositions about the health of the system as a whole at once but we can observe and construct propositions at lower system hierarchies. These propositions (opinions) are combined at higher levels using the rules of probabilistic logic, retaining the ignorance and finally yielding a single opinion on the health of the whole monitored system.

Keywords: Fault detection; FDI; probabilistic logic; system health.

1 Introduction

A fault is defined as an unpermitted deviation of at least one characteristic property of a variable from an acceptable behaviour. Therefore, the fault is a state that may lead to a malfunction or failure of the system [1, 2].

With increasing demands for safety and efficiency of complex processes, fault detection and isolation (FDI) becomes part of control systems in chemical, nuclear and aerospace engineering, automotive systems, power plant stations [3], software development etc. [4]. Together with FDI, controller capable to prevent failure or system reconfiguration that ensures the reliable and safe operation in the presence of component failures [5] might be implemented as well. FDI itself consists in binary opinion whether the system is in faulty state and indication of location and nature of the fault.

There exist three main classes of FDI methods: (i) *knowledge-based* FDI, exploiting human factor expertise, (ii) *signal-based* FDI considering properties of single or multiple signals and exploiting bounds checking, change- point detection, correlation and regression analyses etc., and (iii) *process-model-based* FDI,

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

reflecting a high-level model-based view on the whole manufacturing process. Quantitative methods use explicit process model in combination with statistics and decision theory, qualitative methods are based on artificial intelligence approach (pattern recognition, fuzzy theory, neural networks, spectral analysis etc.), see e.g. [6–9].

There are many process-model-based methods to evaluate faulty state, e.g. full-state observer-based methods or unknown input observer methods (using state-space system models), parity relations methods (using linear transformation of predicted and observed output), optimisation-based methods (minimising sensitivity to noise and maximising sensitivity to faults), methods based on Kalman filter, stochastic approach (description of a system by probabilistic distributions), system identification (tracking of model parameters), artificial intelligence techniques and others [4]. To deal with unobservable state variables, candidate methods are used to estimate the system evolution. A set of possible states (candidates) is constructed and used for comparison of output and the model to predict the expected future state of the system given each candidate [10], [11], [12]. The system structure can be represented either directly by the model or by using temporal logic describing possible sequence of events in case of fault occurrence within particular components [10],[12].

If we focus on an industrial plant, we may distinguish many possible fault sources. For instance, *sensors* are typically sources of gross errors, e.g. due to a fixed failure, a constant bias (positive or negative) or an out-of range failure. Some of the sensed variables are used for subsequent process control, which, under failure, may lead to significant degradation of production quality, unless this state is quickly detected and an appropriate action is taken. Another possible fault sources are *actuators*. While total breakdown of an actuator can be easily detected in most cases, a slow deterioration of actuator's performance is a more challenging problem. Its detection can be easier if the actuator provides suitable feedback signal(s). *Hardware faults* stretch from trivial irreversible malfunction of a hardware component to hardly discoverable degradation of function caused e.g. by insufficient cooling of computer chip sets. *Software faults* (permanent or transient) may be caused by improper software configuration and incompatibilities, timing problems and even faults following from the lack of testing and bad programming habits. Other system faults can be initiated, e.g., by overloading of the operating system or communication lines due to exceptional situation, unacceptable signal-to-noise ratio, echoes etc.

The heterogeneous sources of faults inevitably place considerable demands on related FDI. The situation is yet more complicated due to different possible time developments of faults. In this respect, three basic fault types can be distinguished [1] and should be detectable by the proposed system: (i) *abrupt fault* causing undesirable stepwise change of a signal at once. This fault type is usually easily detected with only minimal delay. However, it may lead to immediate deviation of production quality beyond acceptable limits; (ii) *incipient fault* typical with its continuous drift from desirable value. Its recognition is closely tied with the character of the drift, mainly its time and "shape" properties; (iii) *intermit-*

tent fault occuring in intervals, usually irregular. These faults are generally very problematic due to their difficult detectability and isolation.

To summarize, a monitoring and processing of the system as a whole results generally in a solution (i) tailored for a particular system, i.e single-purpose (ii) combining different probability distributions of particular quantities of interest, either discrete of continuous, (iii) having a high dimensionality.

In this paper, we focus on a novel proposal of a dynamic FDI system based on probabilistic approach to fault detection. In the presented approach, the system of interest is decomposed into blocks, representing individual physical or logical system units (e.g. sensors, actuators, communication lines etc.). To each particular block, an observer is assigned that provides an opinion of the respective block health and related uncertainty. These observers can be considered as imperfect participants communicating via their connections within a structure. We aim to combine the information provided by involved participants to obtain a resulting value of system health.

The individual information pieces are fused together using the probabilistic logic framework in order to evaluate the health of the overall system. Probabilistic logic combines the capacity of probability theory to handle uncertainty with the capacity of deductive logic to exploit structure. We focus on a special type of probabilistic logic called subjective logic (SL) that explicitly takes uncertainty into account. It allows probability values to be expressed with degrees of uncertainty. In general, SL is suitable for modelling and analysing situations characterised by uncertainty and incomplete knowledge [13–15]. Note that the evaluation of opinions of the particular block health is not part of the current paper.

The paper is organised as follows. Section 2 provides basics of SL theory needed for its anticipated application to the problem of the system health monitoring. Section 3 gives a simple simulated example of industrial system and an evaluation of its health using rules of SL.

2 Representation and fusion of FDI-relevant knowledge

This section briefly deals with basics of SL framework as defined in [13–15]. We focus on such features that are important to the solving our problem that lies in (i) a representation of the knowledge about the health of individual industrial system blocks and (ii) combining these particular information pieces to obtain opinion of the overall health of the examined industrial system.

2.1 Basic notion of belief theory

In SL, the representation of uncertain probabilities is based on a belief model similar to the one used in [16]. The first step in applying this model is to define an exhaustive set of of mutually exclusive elementary states of a given system, called the frame of discernment or state space and denoted by Θ. The powerset of Θ, denoted by 2^{Θ}, contains all possible subsets of Θ including Θ itself.

Elementary state in a frame of discernment Θ will be called atomic sets because they do not contain subsets. It is assumed that only one atomic set can be true at any one time. If a set is assumed to be true, then all supersets are considered true as well. An observer (subject, participant) who believes that one or several sets in the powerset of Θ might be true can assign belief masses to these sets. Belief mass on an atomic set $x \in 2^\Theta$ is interpreted as the belief that the set in question is true. Belief mass on a non-atomic set $x \in 2^\Theta$ is interpreted as the belief that one of the atomic sets it contains is true, but that the observer is uncertain about which of them is true.

A belief mass assignment (BMA) m_Θ distributes a total belief mass of 1 amongst the subsets of Θ such that the belief mass for each subset is positive or zero. Function $m_\Theta : 2^\Theta \to [0,1]$ fulfills:

$$m_\Theta(x) \geq 0,\ m_\Theta(\emptyset) = 0,\ \sum_{x \in 2^\Theta} m_\Theta(x) = 1, \tag{1}$$

For each subset $x \in 2^\Theta$, the number $m_\Theta(x)$ is called the belief mass of x.

A belief mass $m_\Theta(x)$ expresses the belief assigned to the set x and does not express any belief in subsets of x in particular. A BMA is called dogmatic if $m_\Theta(\Theta) = 0$ because the total amount of belief mass has been committed. In contrast to belief mass, the belief in a set must be interpreted as an observers total belief that a particular set is true. A belief in x not only depends on belief mass assigned to x but also on belief mass assigned to subsets of x. Each subset $x \subseteq \Theta$ such that $m_\Theta(x) > 0$ is called a focal element of Θ. Note that in case all focal elements are elementary states then we speak about Bayesian BMA. A total belief that a particular state is true is expressed by the belief function $b : 2^\Theta \to [0,1]$ defined by

$$b(x) = \sum_{\emptyset \neq y \subseteq x} m_\Theta(y),\ x, y \in 2^\Theta.$$

Similarly to belief, a disbelief is interpreted as the total belief that a state is not true. Disbelief function corresponding with m_Θ is the function $d : 2^\Theta \to [0,1]$ defined by

$$d(x) = \sum_{y \cap x = \emptyset} m_\Theta(y),\ x, y \in 2^\Theta.$$

The uncertainty function corresponding with m_Θ is the function $u : 2^\Theta \to [0,1]$ defined by

$$u(x) = \sum_{\substack{y \cap x \neq \emptyset \\ y \not\subseteq x}} m_\Theta(y),\ x, y \in 2^\Theta.$$

Due to (1), the sum of the belief, disbelief and uncertainty functions is equal one, i.e.

$$b(x) + d(x) + u(x) = 1,\ x \in 2^\Theta,\ x \neq \emptyset. \tag{2}$$

In subjective logic, subjective opinions express specific types of beliefs, and represent the input and output arguments of the subjective logic operators.

Opinions expressed over binary state spaces are called binomial. Opinions defined over state spaces larger than binary are called multinomial. In this paper, we focus on the binomial opinion only as they suit FDI concept where the state space consist of two states, i.e. functionality/nonfunctionality of specific block of given system.

A MBA where the possible focal elements are Θ and/or elementary states (singletons) of Θ, is called a Dirichlet BMA function. The same mapping in the case of binary state spaces is called Beta belief mass distribution.

Base rate function $a : \Theta \rightarrow [0, 1]$ represents a priori probability expectation before any evidence has been received and fulfill

$$a(\emptyset) = 0 \text{ and } \sum_{x \in \Theta} a(x) = 1 \tag{3}$$

The combination of a Dirichlet MBA (or Beta MBA) and a base rate function can be comprised in a composite function called an opinion. Subjective opinions represent a special type of general belief functions. The subjective opinion model extend the traditional belief function model in the sense that opinions take base rates (it correspond to a prior information) into account whereas belief functions ignore base rates.

The probability transformation [17] projects a MBA onto a probability expectation value denoted by $p(x)$ as follows

$$p(x) = \sum_{y \subseteq \Theta} m_{\Theta}(y) \frac{|x \cap y|}{|y|}, \; x, y \in 2^{\Theta} \tag{4}$$

2.2 Elements of binomial subjective opinions

A subjective opinion expresses a subjective belief of a particular subject (participant) about the truth of propositions including a degree of uncertainty. The propositions are represented by elementary states as defined in section 2.1. An opinion is denoted as ω_x^A where A is the subject who provides this opinion, and x is the proposition (state) to which the opinion applies. The proposition x is assumed to belong to a state space Θ which is usually not included in the opinion notation. The subject, the proposition and its frame are attributes of an opinion. Indication of subjective belief ownership is normally omitted whenever irrelevant, e.g. when only one subject is considered. A general multinomial opinion applies to a collection of propositions. A binomial opinion applies to a single proposition. Hereafter, we focus on binomial opinions only.

The binomial opinion is defined as follows: Let $\Theta = \{x, \bar{x}\}$ be a binary frame. A binomial opinion about the truth of state x is the ordered quadruple

$$\omega_x = (b, d, u, a) \tag{5}$$

where:
belief b is the belief mass in support of x being true,

disbelief d is the belief mass in support of x being false,
uncertainty u is the amount of uncommitted belief mass,
base rate a is the a priori probability in the absence of committed belief mass.

These components satisfy (2) and it holds b, d, u, $a \in [0, 1]$.

The probability expectation $p(x)$ (4) is defined by

$$p(x) = E_x = b + au \tag{6}$$

2.3 Binomial Beta opinion

Binomial opinion class has an equivalence mapping to Beta probability density function (pdf) under specific conditions. This mapping then gives subjective opinions a basis in notions from classical probability and statistics theory.

A general uncertain binomial opinion (i.e. with $u > 0$) corresponds to a Beta pdf denoted as $B(p|\alpha, \beta)$ where α and β are its two evidence parameters, $p = p(x)$ is defined by (6). Beta pdfs are expressed as

$$B(p|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, 0 \leq p \leq 1,\ \alpha > 0,\ \beta > 0, \tag{7}$$

with the restriction that $p \neq 0$ if $\alpha < 1$ and $p \neq 0$ if $\beta < 1$.

Let r denote the number of observations of x, and let s denote the number of observations of $\bar{x}$. Then α and β parameters can be expressed as a function of the observations (r, s) in addition to the base rate a.

$$\begin{aligned} \alpha &= r + Wa \\ \beta &= s + W(1-a) \end{aligned} \tag{8}$$

The default non-informative prior weight $W = 2$ produces a uniform Beta pdf in case of default base rate $a = 1/2$ and $r = s = 0$.

The probability expectation value of the Beta pdf is defined as follows

$$E(B(p|\alpha, \beta)) = \frac{\alpha}{\alpha + \beta} = \frac{r + Wa}{r + s + W}. \tag{9}$$

The mapping from the parameters of a binomial opinion $\omega_X = (b, d, u, a)$ to the parameters of $B(p|\alpha, \beta)$ is defined as follows.

Let $\omega_X = (b, d, u, a)$ be a binomial opinion, and let $B(p|\alpha, \beta)$ with α, β defined by (8) be a Beta pdf, both over the same proposition x, i.e over the binary state space $\{x, \bar{x}\}$.

The opinions ω_X and $B(p|\alpha, \beta)$ are equivalent through the following mapping:

$$\left.\begin{aligned} b &= \tfrac{r}{W+r+s} = \tfrac{\alpha - Wa}{\alpha+\beta} \\ d &= \tfrac{s}{W+r+s} = \tfrac{\beta - W(1-a)}{\alpha+\beta} \\ u &= \tfrac{W}{W+r+s} = \tfrac{W}{\alpha+\beta} \end{aligned}\right\} \Leftrightarrow \left\{\begin{aligned} &\underline{\text{for } u \neq 0:} && \underline{\text{for } u = 0:} \\ &\alpha = \tfrac{W(b+au)}{u} && \\ &\beta = \tfrac{W(d+(1-a)u)}{u} && E[p] = b \\ &1 = b + d + u && 1 = b + d \end{aligned}\right. \tag{10}$$

The equivalence between binomial opinions and Beta pds is very powerful because subjective logic operators then can be applied to density functions and vice versa, and also because binomial opinions can be determined through statistical observations. For more details see [15].

2.4 Operators of subjective logic

Subjective logic provides a set of operators where input and output arguments are in the form of binomial opinions defined over binary frames. By using these operators, an efficient computation of mathematically complex models is enabled. Most of the operators correspond to well-known operators from binary logic and probability calculus. Additional operators exist for modelling special situations, such as when fusing opinions of multiple observers.

Below, some selected operators are described in detail:

Let $\Theta_1 = \{x, \overline{x}\}$ and $\Theta_2 = \{y, \overline{y}\}$ be two separate frames with independent opinions $\omega_X = (b_x, d_x, u_x, a_x)$ and $\omega_Y = (b_y, d_y, u_y, a_y)$, respectively.

Binomial multiplication corresponds to the logical AND and probability product. Notation: $\omega_{X\wedge Y} = \omega_X \cdot \omega_Y$

$$\begin{aligned} b_{x\wedge y} &= b_x b_y + \frac{(1-a_x)a_y b_x u_y + (1-a_y)a_x b_y u_x}{1-a_x a_y} \\ d_{x\wedge y} &= d_x + d_y - d_x d_y \\ u_{x\wedge y} &= u_x u_y + \frac{(1-a_y)b_x u_y + (1-a_x)b_y u_x}{1-a_x a_y} \\ a_{x\wedge y} &= a_x a_y \end{aligned} \tag{11}$$

Binomial comultiplication corresponds to the logical OR and probability coproduct. Notation: $\omega_{X\vee Y} = \omega_X \sqcup \omega_Y$

$$\begin{aligned} b_{x\vee y} &= b_x + b_y - b_x b_y \\ d_{x\vee y} &= d_x d_y + \frac{(1-a_y)a_x d_x u_y + (1-a_x)a_y d_y u_x}{a_x + a_y - a_x a_y} \\ u_{x\vee y} &= u_x u_y + \frac{a_y d_x u_y + a_x d_y u_x}{a_x + a_y - a_x a_y} \\ a_{x\vee y} &= a_x + a_y - a_x a_y \end{aligned} \tag{12}$$

Let two subjects A and B observe the same $X = \{x, \overline{x}\}$ in the same time instant and evaluate their opinions ω_X^A and ω_X^B.

Averaging Fusion $\omega_X^{A\underline{\diamond}B} = \omega_X^A \underline{\oplus} \omega_X^B$

$$\begin{aligned} b^{A\underline{\diamond}B} &= \frac{b^A u^B + b^B u^A}{u^A + u^B} \\ u^{A\underline{\diamond}B} &= \frac{2u^A u^B}{u^A + u^B} \end{aligned} \tag{13}$$

3 Example: monitoring of the system condition

In this section, an application of SL to the problem of the health monitoring of a technological process is presented.

We suppose that the investigated system is composed of a set of basic blocks. We assume that the blocks are monitored by a device-specific subjects that provide binary opinions on the functionality of the relevant blocks. This information is transformed into opinion on the block functionality using (10). We demonstrate the principle of combining involved opinion only. Also, the influence of changes in one block on the whole system is examined. For these purposes, we use a simulated system as defined below.

Let us consider a simple system of position adjustment to be monitored (see Fig. 1). The system consists of three basic blocks X (position measurement), Y (velocity measurement) and Z (actuator) that are organised in two units. The first unit contains blocks X and Y. They are interchangeable, i.e. information obtained by Y can be used to substitute information by X and vice versa (redundancy). The functionality of the sensor X is monitored by two subjects A (analysing noise and giving opinion ω_X^A) and B (analysing response and giving opinion ω_X^B), the functionality of the sensor Y is monitored by subject C (giving opinion ω_Y^C). The functionality of the actuator Z in the second unit is monitored by subject D (giving opinion ω_Z^D). Note that Fig. 1 does not describe physical composition of the system but units in a hierarchical structure showing how information of one unit affects the others. Each unit can be, on different levels of abstraction, created by sub-units etc.

Subjects A and B observe the same X simultaneously. Their opinions are therefore composed together by averaging fusion (13). The sensors X and Y are interchangeable, i.e. functionality of at least one of them is sufficient for a correct performance of the system. Therefore, opinions on their functionality is represented by comultiplication (12). Finally, the two major units must both work at the same time, therefore opinion on their mutual operation state is obtained as a multiplication (11) of opinions on each unit.

We denote subjects' opinions as ω_X^A, ω_X^B, ω_Y^C, ω_Z^D and opinion on the overall system functionality as ω. Then, using notation from Table **??**, we get

$$\omega \equiv (b, d, u, a) = \left[\left(\omega_X^A \underline{\oplus} \omega_X^B\right) \sqcup \omega_Y^C\right] \cdot \omega_Z^D. \tag{14}$$

The opinions are given as follows

$$\begin{aligned} \omega_X^A &= (0.9, 0.0, 0.1, 0.8) \\ \omega_X^B &= (0.8, 0.1, 0.1, 0.5) \\ \omega_Y^C &= (0.8, 0.2, 0.0, 0.5) \\ \omega_Z^D &= (0.9, 0.0, 0.1, 0.3) \end{aligned} \tag{15}$$

which indicate a high belief in the blocks' functionality with low (or absent) uncertainty and prior doubts about the block D ($a = 0.3$). Then, according to (14), $\omega = (0.89, 0.01, 0.1, 0.25)$ indicating high functionality of the system.

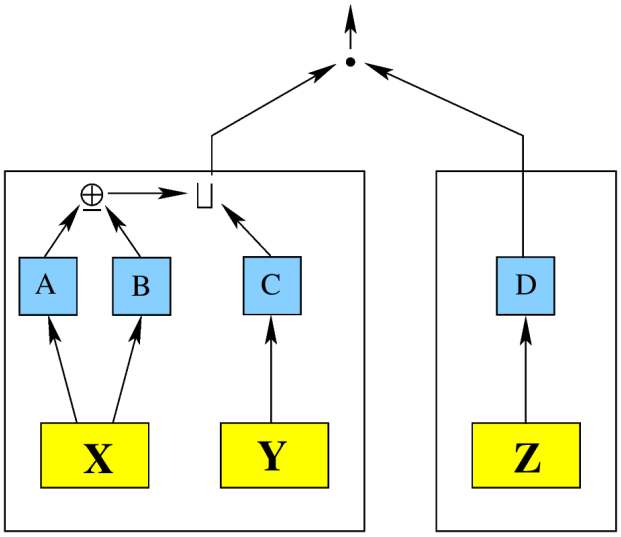


Fig. 1. Scheme of a monitored system. X, Y: monitored sensors, Z: another device, A, B, C, D: monitoring subjects

If we decrease belief in ω_X^A to $\omega_X^A = (0.2, 0.7, 0.1, 0.8)$, we get $\omega = (0.83, 0.08, 0.09, 0.25)$ which still keeps high performance because of the redundancy of X and Y.

If we use (15) but decrease belief in ω_Z^D to $\omega_Z^D = (0.2, 0.7, 0.1, 0.3)$, we get $\omega = (0.20, 0.70, 0.09, 0.75)$ showing poor overall performance and strong influence of the isolated block Z.

A set of experiment follows that examines how changes in the opinion of one block influence the behaviour of the whole system.

3.1 Influence of belief and disbelief

Let us keep values in (15) except of ω_Z^D. We consider $\omega_Z^D = (b_Z^D, d_Z^D, 0.1, 0.3) = (b_Z^D, 0.9 - b_Z^D, 0.1, 0.3)$ where b_Z^D lies within possible ranges given by (2), i.e. $b_Z^D \in [0, 0.9]$. The dependence of individual entries of ω on b_Z^D is shown in Fig. 2. The b and d are influenced very strongly because D enters the top level directly.

Now, we consider varying belief/disbelief in ω_X^A whereas other opinions fulfill (15). Similarly to the above mentioned case, $\omega_X^A = (b_X^A, 0.9 - b_X^A, 0.1, 0.8)$, where $b_X^A \in [0, 0.9]$. Results are shown in Fig. 3. It is obvious that an influence of the subject A involved in a more complex unit is less significant than in the previous case.

3.2 Influence of uncertainty

Now, we use (15) but change uncertainty in ω_Z^D, $u_Z^D \in [0,1]$. Then, according to (2), $b_Z^D = 1 - u_Z^D$ and $\omega_Z^D = (1 - u_Z^D, 0.0, u_Z^D, 0.3)$. The course of ω is shown in Fig. 4.

The same experiment for $\omega_X^A = (1 - u_X^A, 0, u_X^A, 0.8)$ is shown in Fig. 5, $u_X^A \in [0,1]$.

Again, we can see a strong influence of ω_Z^D. In this experiment, influence of A's uncertainty is practically negligible because the opinion on measurement is backed-up both by another subject B and also by another sensor Y.

3.3 Influence of base rates change

Let (15) be used and base rate a_Z^D is to be varying: $\omega_Z^D = (0.9, 0.0, 0.1, a_Z^D)$, $a_Z^D \in [0,1]$. The influence of a_Z^D on overall ω is shown in Fig. 6. Increasing prior judgement a_Z^D on isolated block Z increases belief and decreases uncertainty of overall opinion, whereas disbelief remains practically unchanged. The most significant effect is linear increase of base rate, see (11).

Finally, we use (15) and vary $a_X^A \in [0,1]$ in $\omega_X^A = (0.9, 0.0, 0.1, a_X^A)$. The influence on ω is shown in Fig. 7. Overall base rate is, again, affected most significantly. On the other hand, increasing value of a_X^A, increases uncertainty and slightly decreases disbelief.

4 Conclusion

In this paper, we proposed a novel type of probabilistic logic-based fault detection system with a highly modular and scalable structure. The decomposed system is represented by a collection of interconnected blocks, that can be interpreted as individual participants, whose opinions on particular block health is evaluated via Bayesian modelling. The methodology to obtain these opinions is specific according to nature of a particular unit and it is not addressed in the present paper. In order to obtain an information about the health of the whole monitored system, these particular opinions are fused together using the rules of probabilistic (more precisely subjective) logic. The resulting FDI system provides the human operator with information about the system functionality as a whole and at the same time enables to recognise health of particular blocks.

The proposed methodology (i) has capability of modular description and scalability, (ii) enables individual application of suitable probabilistic mechanism for each block, and (iii) avoids the dimensionality problem by using combination of low-dimensional units.

The future work comprises (i) evaluation of opinions on system health at each block and (ii) analysis of feasibility of the proposed system.

Acknowledgement

The project is supported by the grant MŠMT 7D12004 (E!7262 ProDiSMon) and by the Czech Science Foundation, project no. 13–13502S.

References

1. Isermann, R.: Model-based fault-detection and diagnosis status and applications. Annual Reviews in Control **29**(1) (2005) 71 – 85
2. Isermann, R.: Fault Diagnosis Applications: Model Based Condition Monitoring, Actuators, Drives, Machinery, Plants, Sensors, and Fault-tolerant Systems. Springer Verlag (2011)
3. Huang, X., Qi, H., Liu, X.: Implementation of fault detection and diagnosis system for control systems in thermal power plants. In: Proceedings of the 6th World Congress on Intelligent Control and Automation, Dalian, China, Institute of Electrical and Electronics Engineers (IEEE) (June 21–23 2006) 5777–5781
4. Hwang, I., Kim, S., Kim, Y., Seah, C.E.: A survey of fault detection, isolation, and reconfiguration methods. IEEE Transactions on Control Systems Technology **18**(3) (May 2010)
5. Zhang, Y., Jiang, J.: Bibliographical review on reconfigurable fault-tolerant control systems. Annual Reviews in Control **32**(2) (2008) 229 – 252
6. Ding, S.: Model-based fault diagnosis techniques: design schemes, algorithms, and tools. Springer (2008)
7. Gustafsson, F.: Statistical signal processing approaches to fault detection. Annual Reviews in Control **31**(1) (2007) 41–54
8. Venkatasubramanian, V., Rengaswamy, R., Yin, K., Kavuri, S.N.: A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. Computers & Chemical Engineering **27**(3) (March 2003) 293–311
9. Venkatasubramanian, V., Rengaswamy, R., Kavuri, S.N., Yin, K.: A review of process fault detection and diagnosis: Part III: Process history based methods. Computers & Chemical Engineering **27**(3) (March 2003) 327–346
10. Williams, B.C., Nayak, P.P.: A model-based approach to reactive self-configuring systems. In: In Proceedings of AAAI-96. (1996) 971–978
11. Balaban, E., Cannon, H.N., Narasimhan, S., Brownston, L.S.: Model-based fault detection and diagnosis system for NASA Mars subsurface drill prototype. In: 2007 IEEE Aerospace Conference, Big Sky, Montana, Institute of Electrical and Electronics Engineers (IEEE) (March 2007) 13
12. Magni, L., Scattolini, R., Rossi, C.: A fault detection and isolation method for complex industrial systems. IEEE Transactions on Systems, Man and Cybernetics — Part A: Systems and Humans **30**(6) (November 2000) 860–864
13. Jøsang, A.: A logic for uncertain probabilities. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **9**(03) (2001) 279–311
14. Jøsang, A.: Probabilistic logic under uncertainty. In: Proceedings of the thirteenth Australasian symposium on Theory of computing - Volume 65. CATS '07, Darlinghurst, Australia, Australia, Australian Computer Society, Inc. (2007) 101–110
15. Jøsang, A.: Subjective logic. Draft book Available at: http://persons. unik. no/josang/papers/subjective_logic.pdf, visited **26** (2010)
16. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
17. Dubois, D., Prade, H.: On several representations of an uncertain body of evidence. Fuzzy Information and Decision Processes (1982) 167–181

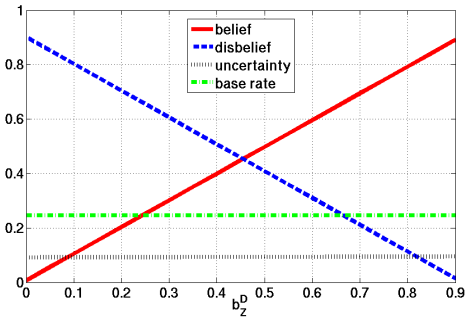


Fig. 2. Dependence of overall opinion $\omega = (b, d, u, a)$ on belief b_Z^D of subject D. Solid line: belief b, dashed line: disbelief d, dotted line: uncertainty u, dash-dot line: base rate a

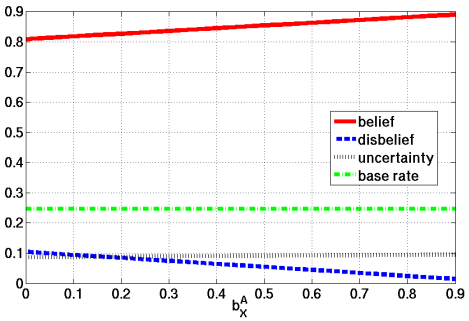


Fig. 3. Dependence of overall opinion $\omega = (b, d, u, a)$ on belief b_X^A of subject A. Solid line: belief b, dashed line: disbelief d, dotted line: uncertainty u, dash-dot line: base rate a

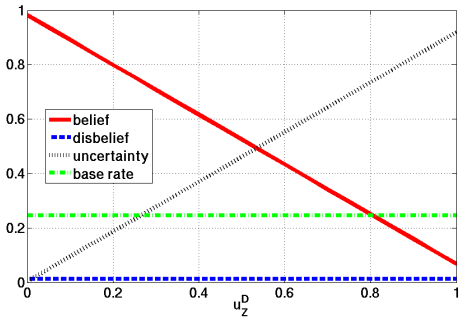


Fig. 4. Dependence of overall opinion $\omega = (b, d, u, a)$ on uncertainty u_Z^D of subject D. Solid line: belief b, dashed line: disbelief d, dotted line: uncertainty u, dash-dot line: base rate a

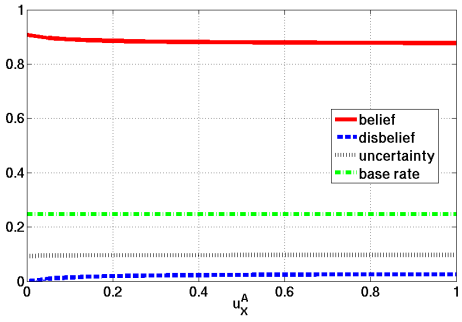


Fig. 5. Dependence of overall opinion $\omega = (b, d, u, a)$ on uncertainty u_X^A of subject A. Solid line: belief b, dashed line: disbelief d, dotted line: uncertainty u, dash-dot line: base rate a

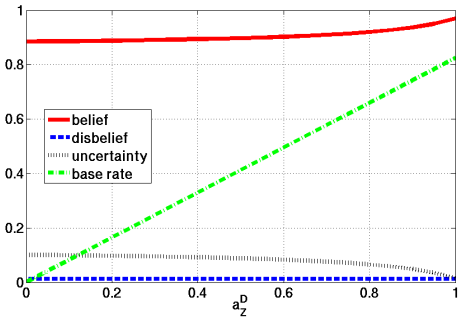


Fig. 6. Dependence of overall opinion $\omega = (b, d, u, a)$ on base rate a_Z^D of subject D. Solid line: belief b, dashed line: disbelief d, dotted line: uncertainty u, dash-dot line: base rate a

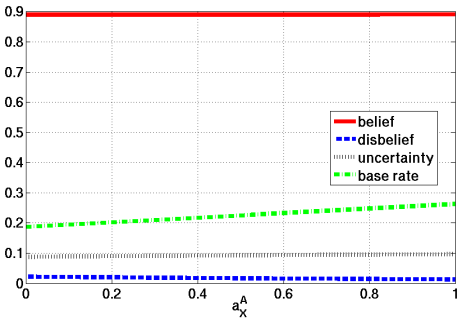


Fig. 7. Dependence of overall opinion $\omega = (b, d, u, a)$ on base rate a_X^A of subject A. Solid line: belief b, dashed line: disbelief d, dotted line: uncertainty u, dash-dot line: base rate a

A NOTE ON WEIGHTED COMBINATION METHODS FOR PROBABILITY ESTIMATION

Vladimíra Sečkárová*

Institute of Information Theory and Automation of the ASCR,
Pod Vodárenskou věží 4, CZ-182 08 Prague 8, Czech Republic
seckarov@utia.cas.cz

Department of Probability and Mathematical Statistics,
Charles University in Prague, Czech Republic

Abstract. To successfully learn from the information provided by available information sources, the choice of automatic method combining them into one aggregate result plays an important role. To respect the reliability in the source's performance each of them is assigned a weight, often subjectively influenced. To overcome this issue, we briefly describe the method based on Bayesian decision theory and elements of information theory. In particular we consider discrete-type information, represented by probability mass functions (pmfs) and obtain an aggregate result, which has also form of pmf. This result of decision making process is found to be a weighted linear combination of available information. Besides the brief description of the novel method, the paper focuses on its comparison with other combination methods. Since we consider the available information and unknown aggregate as pmfs, we mainly focus on the case when the parameter of binomial distribution is of interest and the sources provide appropriate pmfs.

Keywords: weighting methods, parameter estimation, Kerridge inaccuracy, maximum entropy principle, binomial distribution

1 Introduction

Exploiting available information plays an important role in many parts of mathematics such as parameter estimation, quality control, etc. and their applications. Usually, the processed information originates in many sources. The sources of information can range from experts in particular field to sensors measuring physical variables. To obtain the reliable result of interest based on these data we need to assign each source a weight. This weight should express reliability of a particular source and is usually assigned by an extra expert, thus is subjectively

* This research has been partially supported by GACR 13-13502S and SVV-2013 - 267 315.

influenced. It is worthwhile, especially in complex situations, to prevent the subjectivity. This paper focuses on the objective choice of weights under several commonly acceptable assumptions.

Throughout the paper we assume the sources provide the information about a common random vector having finite amount of realizations. The probability distribution over this random vector depends on an unknown, generally multidimensional, parameter, representing the ideal aggregated information. Our aim, the parameter estimate based on available data, will then be a combination of these data. Useful survey on combination methods from the mathematical and behavioural point of view can be found in [4]. To obtain the aggregate we express the parameter estimation task as a task of decision making and exploit the basic steps of Bayesian decision theory (see e.g. [6], Section VIIID) to compose an optimal decision. The optimality criterion is based on the minimization of an expected loss. The specific loss function we adopt is based on the elements of information theory. The parameter of interest is assumed to be a probability mass function (pmf), i.e. a column vector whose non-negative elements sum to unity, and the data provided by information sources are in the form of pmfs, too. A nice survey on the combination methods using elements of information theory can be found in [1], describing approaches with weights more or less subjectively influenced. To eliminate the subjective influence, we work with the method based on the Kerridge inaccuracy [8] and maximum entropy principle [14], which leads to a final weighted combination with weights determined during the construction of this combination [13]. The weights in the final combination are based on the information included in provided data and thus no subjective influence is added to them.

The main goal of this paper is to compare the proposed method with other methods in the considered field. We focus on the estimation of the parameter in binomial distribution. The methods serving to comparison belong to the group of empirical Bayes methods [11], where the prior distribution is computed from the previous observations.

The paper is organized as follows: the next section provides a brief description of our approach based on Bayesian decision theory and elements of information theory, the third section provides an overview of methods used for comparison, the fourth section gives the resulting estimates obtained by the considered methods based on the same data sample. The fifth section contains conclusion and topics for the future work. The previously published derivations connected with our approach (see [13]) can be found in the Appendix.

2 Proposed Method

In this section we briefly introduce a method combining the available information based on elements of information theory. The motivation for this particular choice is based on the aim of elimination of the subjectivity in the weights. For the weights' assignments we focus on the natural part included in provided

data, i.e. we exploit the amount of information included in them. This of course requires specific setup described in the following paragraphs.

Let us start with a parameter estimation task, where the parameter h has the form of pmf, belonging to the probabilistic simplex. To obtain its estimate $\hat{h}$ we exploit Bayesian decision theory and look for the estimate minimizing the expected loss function. We select the loss function as a function computing the inaccuracy between a pair of pmfs - the Kerridge inaccuracy K(.,.) (see [3]). The estimate $\hat{h}$ then coincides with the conditional expectation E[.|.] with respect to the posterior probability density function (pdf) $\pi(h|D)$ of the unknown parameter h (an optimal aggregate) conditioned on available data $D = (g_1, \ldots, g_s)^T$ formed by pmfs g_j given by s sources, $s < \infty$:

$$\begin{aligned}\hat{h} &= \arg\min_{\tilde{h}\in\widetilde{H}} \mathrm{E}_{\pi(h|D)}[\mathrm{K}(h,\tilde{h})|D] \\ &= \arg\min_{\tilde{h}\in\widetilde{H}} \mathrm{K}(\mathrm{E}_{\pi(h|D)}[h|D],\tilde{h}) = \mathrm{E}_{\pi(h|D)}[h|D], \end{aligned} \quad (1)$$

where $\widetilde{H} = \left\{(\tilde{h}(x_1),\ldots,\tilde{h}(x_n)) : \sum_{i=1}^n \tilde{h}(x_i) = 1, \tilde{h}(x_i) > 0,\ i = 1\ldots,n\right\}$ and $g_j \in \widetilde{H}$, $j = 1,\ldots,s$. Sources describe a common random vector X having possible outcomes $\{x_i\}_{i=1}^n$, $n < \infty$, i.e. provide the probabilities $g_j(x_i) = P_j(X = x_i)$, $j = 1,\ldots,s$.

To compute the estimate (1) we need to determine the posterior pdf $\pi(h|D)$, which is yet unknown. To determine its form, we exploit the maximum entropy principle [14]. It leads to the convex optimization:

$$\hat{\pi}(h|D) = \arg\min_{\widetilde{\pi}(h|D)\in\mathrm{M}} \left[\int_H \widetilde{\pi}(h|D)\log\widetilde{\pi}(h|D)\mathrm{d}h\right], \quad (2)$$

where $\mathrm{M} = \{\widetilde{\pi}(h|D) : \mathrm{E}_{\widetilde{\pi}(h|D)}(\mathrm{K}(g_j,h)|D) \le \beta_j(D),\ j = 1,\ldots,s,$
$\int_H \widetilde{\pi}(h|D)\mathrm{d}h = 1\}$.

The constraints in M express the assumption the j^{th} source will accept h as a compromise (optimal aggregate) if it serves as a good approximation of j^{th} pmf. According to [3] we expect that from the Bayesian point of view the Kerridge inaccuracy employed in the set M should reach low values for good approximations. The optional scalar $\beta_j(D)$ reflects "tolerance" of j^{th} source to accept h as an approximation of its opinion g_j.

The optimization results in pdf of Dirichlet distribution (see Subsection 6.1) and the final point estimate $\hat{h}$ of h is the expected value of this distribution and has the form of a weighted combination of given pmfs g_j (see Subsection 6.3):

$$\mathrm{E}_{\hat{\pi}(h|D)}(h(x_i)|D) = \hat{h}(x_i) = \lambda_0^*(D) + \sum_{j=1}^s \lambda_j^*(D) g_j(x_i), \quad i = 1,\ldots,n, \quad (3)$$

where

$$\lambda_0^*(D) = \frac{1}{n + \sum_{j=1}^s \lambda_j(D)}, \quad \lambda_j^*(D) = \frac{\lambda_j(D)}{n + \sum_{j=1}^s \lambda_j(D)}.$$

The last step of the described method involves the computation of the weights, which heavily depends on evaluation of Kuhn-Tucker multipliers $\lambda_j(D) \geq 0$ arising in the optimization task (2). The straightforward derivation of the multipliers can be found in Subsection 6.2. The problem, which has not been solved yet, is that the multipliers still depend on upper bounds $\beta_j(D)$ for expected Kerridge inaccuracies in (2). Here, we leave each $\beta_j(D)$ free and inspect the behaviour of the estimator (3) as a function of corresponding $\lambda_j(D)$, $j = 1, \ldots, s$. A promising objective solution is being elaborated in Section 4 and suggests the upper bounds (linearly shifted - see (9)) as the mean Kerridge inaccuracy in the following form:

$$\beta_k^*(D) = \frac{\sum_{j=1}^{s} \mathrm{K}(g_k, g_j)}{s} = \mathrm{K}(g_k, h_{\mathrm{data}}), \quad k = 1, \ldots, s, \tag{4}$$

where $h_{\mathrm{data}}(x_i) = \sqrt[s]{\prod_{j=1}^{s} g_j(x_i)}, \quad i = 1, \ldots, n$, i.e. geometric pool of opinions is taken as an aggregate acceptable by all information sources. Thus we can see, that the weights exploit the information captured in the provided data. Since they depend on the choice of the upper bounds $\beta_j(D)$, we study this connection in Section 4.

3 Empirical Bayes Methods

In this section we briefly introduce the methods we use for comparison in Section 4. As mentioned earlier, we try to avoid a subjective influence on the weights used in the aggregation process. The first idea, when there is no extra expert assigning the weights to available sources, is to use the equal weights. This approach can be found e.g. in [1].

Since equal weights do not reflect the reliability of the sources at all, we focus on different type of methods, namely, methods exploiting previously obtained data. The next section will bring the comparison of our method with group of empirical Bayes methods. These methods exploit the Bayes' formula in order to get the estimate of an unknown parameter. Their advantage lies in using the prior distribution based on the data available from the previous time instants, rather than choosing a specific prior distribution. When the prior information enters the Bayes' formula the final estimate is a weighted combination of available data.

The empirical Bayes methods are well-applicable in case when the estimation of a multidimensional parameter being a pmf is of interest. We consider four different approaches in this field, i.e. Griffin-Krutchkoff's estimator [7], Copas' second estimator [5], Lemon's estimator [10] and smooth incomplete beta estimator [12]. Formulas belonging to the mentioned estimators, using a common notation, can be found in [12]. Griffin-Krutchkoff's estimator provide a linear optimal estimator, where the optimality origins in minimizing the risk based on squared error loss. Similar situation considered Copas proposing an estimator, which is again assumed to be linear, it minimizes mean squared loss and guarantees a minimax estimate. Lemon's estimator uses a mean value of specifically chosen functions reflecting the current and previous data as posterior pmf. In

particular the estimator focuses on a conditional probability of the modelled variable conditioned by the unknown parameter while plugging in some estimate of this parameter. All three methods can be easily applied to estimation of the parameter of binomial distribution. Finally, we inspect a smooth incomplete beta estimator, derived particularly for the case of binomial distribution. Here, it is suggested to use a function based on incomplete beta function. The parameter of interest and the final estimate have both form of pmf. In all cases the resulting pmf is a weighted combination of available data, thus these methods are the perfect choice for comparison with our method in (3).

The difference between the above group of methods and method in (3) is in the approach to the available data. To obtain the final estimate the former use the empirical prior distribution based on the previous observations. The latter does not use any prior information and combines data pieces at once.

4 Comparison

In this section a comparison of the proposed method and empirical Bayes methods is given. Assume we are interested in estimation of the probability $p \in (0, 1)$ of success and the probability of failure $1 - p$ in N independent trials modeled by binomial distribution.

4.1 Illustrative Example

Thus in the case of the empirical Bayes methods we are looking for the estimate $\hat{p}$ (at the same time for the estimate $1-\hat{p}$) of an unknown parameter p of random variable Y distributed according to $Bi(N, p)$. The probability of k successes in N independent trials is then $P(Y = k) = \binom{N}{k} p^k (1 - p)^{N-k}$. Let us assume we observe N trials at s time instants. Each time we obtain the number of successes y_j and failures $N - y_j$. To obtain the aggregate $\hat{p}$ in empirical Bayes methods we use the binomial fractions y_j/N, which can be viewed as empirical estimates of p. We can then also get the estimate of probability of failures simply by computing $1 - \hat{p}$.

To apply the method proposed in (3) we take a look at the considered situation from a different perspective. The unknown probability of success and failure form a 2-dimensional unknown parameter h. Let X denote a random variable having two realizations ($n = 2$), namely, $\{x_1, x_2\}$ ={success,failure} and thus $h = (h(x_1), h(x_2))^T$. Also assume we have s sources providing pmfs g_j, $j = 1, \ldots, s$. We realize that according to the notation in the previous paragraph we have

$$\begin{aligned} h =& (h(x_1), h(x_2))^T = (p, 1 - p)^T \\ \hat{h} =& (\hat{h}(x_1), \hat{h}(x_2))^T = (\hat{p}, 1 - \hat{p})^T \\ g_j =& (g_j(x_1), g_j(x_2))^T = (y_j/N, (N - y_j)/N)^T, \quad j = 1, \ldots, s. \end{aligned}$$

Now we can focus on how the previously mentioned methods work. We generate four random values, number of successes, from $Bi(10, 1/3)$. That is, we have $s = 4$ and for each $j = 1, \dots, 4$ we can compute the binomial fractions $y_j/10$ and their counterparts $(10 - y_j)/10$ to get pmfs g_j. The data are the following:

$$D = \begin{pmatrix} \left(\frac{y_1}{10}, \frac{10-y_1}{10}\right) \\ \cdots \\ \cdots \\ \left(\frac{y_4}{10}, \frac{10-y_4}{10}\right) \end{pmatrix} = \begin{pmatrix} (g_1(x_1), g_1(x_2)) \\ \cdots \\ \cdots \\ (g_4(x_1), g_4(x_2)) \end{pmatrix} = \begin{pmatrix} (0.3, 0.7) \\ (0.4, 0.6) \\ (0.2, 0.8) \\ (0.1, 0.9) \end{pmatrix}$$

The upper picture in Fig. 1 shows the behaviour of the Kuhn-Tucker multipliers $\lambda_j(D)$ depending on the values of $\beta_j^*(D)$, $j = 1, \dots, 4$, see (4). We decrease linearly shifted bounds $\beta_j^*(D)$ (see (9)) as follows:

$$\beta_{j,l}^*(D) = \beta_j^*(D) \times (0.85 - l \times 0.0084) \text{ for instants } l = 1, \dots, 100, \quad (5)$$

$$\beta_{j,1}^*(D) = \mathrm{K}(g_j, h_{\mathrm{data}}). \quad (6)$$

In case of dynamic setup, where with each time point we obtain a new data, the empirical Bayes methods update the estimate by new data. In case of our method, the data $g_{j,t}$ in (3) can be viewed as the estimate given by j^{th} source based on its data up to time point t. In the next time step, a new estimate $g_{j,t+1}$ is obtained and again, formula (3) is used to combine all available $g_{j,t+1}$, $j = 1, \dots, s$.

The bottom picture in Fig. 1 brings the final estimate (final aggregate) $\hat{h}$ computed by method in (3) with changing value of $\lambda_j(D)$. Also the source with the highest and the lowest entropy are drawn.

The resulting estimates $\hat{h}$ are given in the Fig. 2. They were obtained using equal weights (EW) and the following methods: Griffin-Krutchkoff's estimator (GK), Copas' second estimator (Co), Lemon's estimator (Le) and a smooth incomplete beta estimator (BE), all briefly introduced in the previous section. We can see that even under a small number of available data our method and Copas' second estimator performed quite well regarding the information, that the data were drawn from binomial distribution with probability of success equal to 1/3 (probability of failure is thus 2/3). In particular, Copas' estimator coincides with the estimator based on equal weights, the mean value of drawn data. The results obtained from Lemon's, Griffin-Krutchkoff's estimators and smooth incomplete beta estimator differ from the true value of $h = (h(x_1), h(x_2))^T$, but are closer to what we would naturally expect from obtained data, which can be misleading for small sample cases. A case with larger sample is studied in Subsection 4.2.

At the end we note that while the empirical Bayes methods exploited the fact that we focus on binomial distribution, our method (3) do not need the information about the original distribution to obtain $\hat{h}$. This predestinates our method to be a normative method for estimation of pmfs.

4.2 Monte Carlo Simulations

In this subsection we study the behaviour of considered methods in Monte Carlo study. We assume the same setup as introduced in Subsection 4.1, thus we gen-

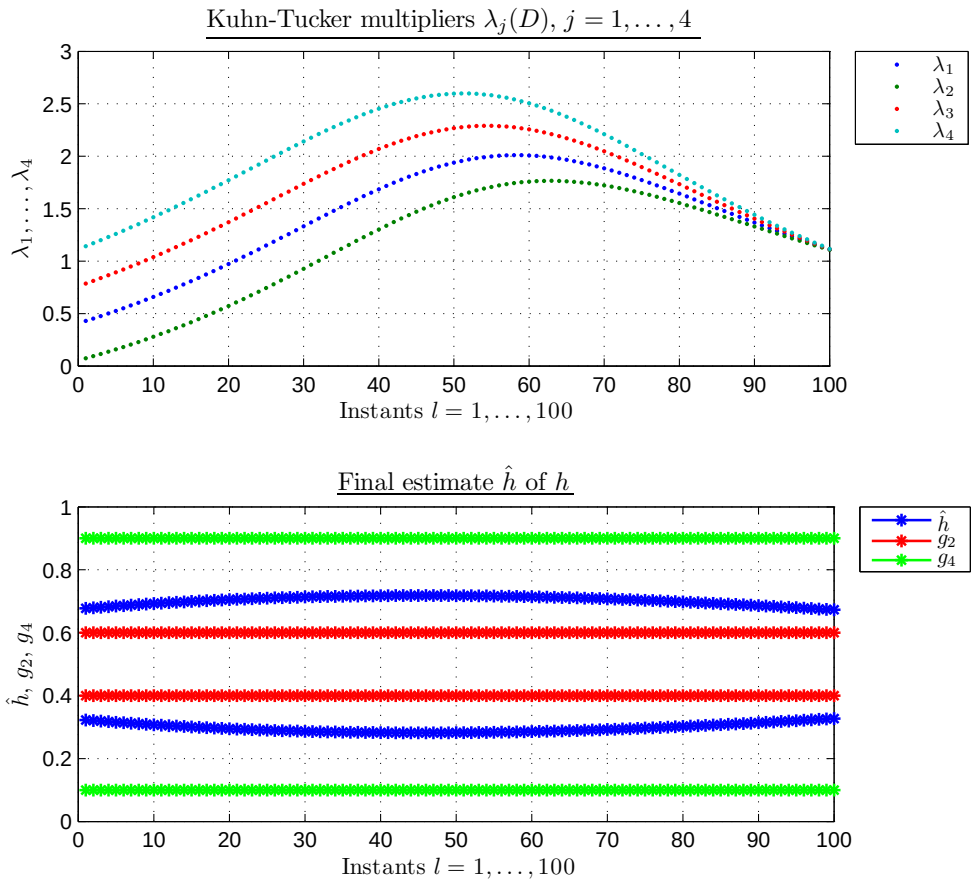


Fig. 1. The behaviour of the $\lambda_j(D)$, $j = 1, \ldots, 4$ based on 100 different decreasing values of $\beta_j^*(D)$ using (5) and the final weighted combination $\hat{h}(x_i)$, $i = 1, 2$ based on computed $\lambda_j(D)$, $j = 1, \ldots, 4$.

erate 10 and 1000 4-tuples from binomial distribution $Bi(10, 1/3)$ and with each new set of data we compute the estimates as in Subsection 4.1. In both cases the upper bounds $\beta_j(D)$ used in our method were with each new set of random values set to $\beta_j^*(D) = \beta_{j,1}^*(D) \times 0.40$, where $\beta_{j,1}^*(D)$ is defined in (6) (see also (4)).

To compare the estimators we are interested in common values as sample mean, computed from all 40 (or 4000) values, mean value and variance of $\hat{h}(x_1)$ obtained from 10 (or 1000) estimates given by considered estimators. We also compute the square error, exploiting the squared distance of the values of par-

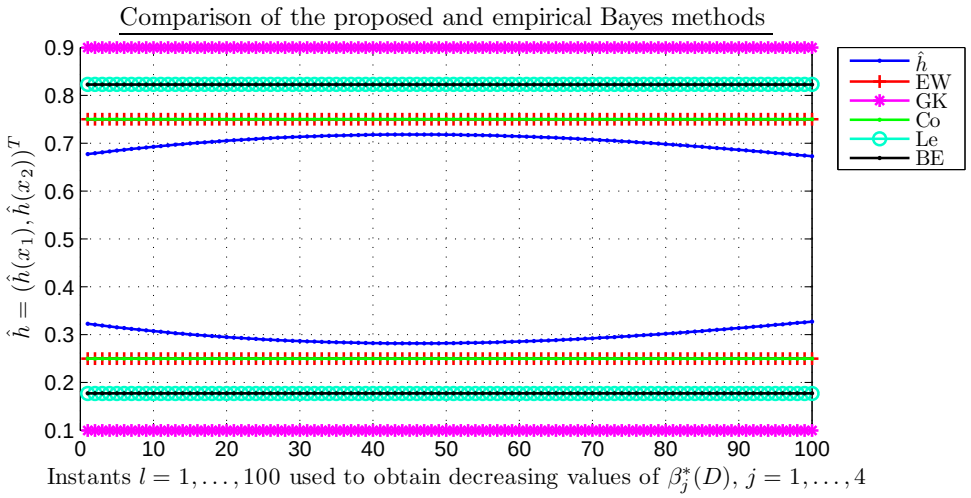


Fig. 2. Empirical Bayes methods in comparison with method based on the Kerridge inaccuracy and maximum entropy principle (3) in case when the estimate $\hat{h}(x_1)$ of the parameter $p = h(x_1)$ of binomial distribution is of interest (also its counterpart $\hat{h}(x_2) = 1 - \hat{h}(x_1)$ is drawn). For the proposed method, the estimate was computed for decreasing values of $\beta_j^*(D)$, $j = 1, \ldots, 4$, see (5).

ticular estimator from the sample mean:

$$\text{sq.error} = \sum_{m=1}^{M} (\hat{h}_m(x_1) - \text{ sample mean})^2, \quad M = 10, 1000$$

In case when only 10 4-tuples were generated, the sample mean equals 0.3850 and the results are the following:

	Proposed method	EW	GK	Co	Le	BE
mean($\hat{h}(x_1)$)	0.3415	0.3850	0.4100	0.3850	0.3939	0.3939
var($\hat{h}(x_1)$)	0.0037	0.0068	0.0121	0.0068	0.0061	0.0061
sq.error	0.0526	0.0615	0.1153	0.0615	0.0557	0.0557

We can see that according to the data in the sample, the estimator based on equal weights and Copas' second estimator work well, Lemon's and smooth incomplete Beta estimator are slightly different from the sample mean. If we take the information about the true distribution of generated data, $Bi(10, 1/3)$, our method gives really good estimate of the unknown parameter $h(x_1)$ even for such small sample.

In case of generating 1000 4-tuples (first 10 4-tuples coincide with those used previously) the sample mean is 0.3350. The results for considered methods are the following:

	Proposed method	EW	GK	Co	Le	BE
mean($\hat{h}(x_1)$)	0.2994	0.3350	0.3307	0.3344	0.3335	0.3335
var($\hat{h}(x_1)$)	0.0028	0.0051	0.0142	0.0062	0.0117	0.0117
sq.error	4.0902	5.0514	14.2110	6.1772	11.7356	11.7222

Here we see that all of the considered Bayes estimators perform really well, the estimate $\hat{h}(x_1)$ based on our method is slightly different from the sample mean and the true value $h(x_i)$. On the other hand the variance of $\hat{h}(x_1)$ and the squared error is the lowest among all considered estimators, which after fixing the values of the upper bounds $\beta_j(D)$ can lead estimates based on our estimator being closer to the sample mean and true value of the unknown parameter $h(x_1)$.

5 Conclusion and Future Work

In this paper we briefly described the method for combining data based on estimation of an unknown parameter. Both, data and parameter, are being pmfs. This method is based on the Kerridge inaccuracy and maximum entropy principle. The final estimate is a weighted combination of data, where the weights are obtained without any subjective influence, yet are non-trivial. They heavily depend on the Kuhn-Tucker multipliers arising during the computation. The aim of this paper consists in comparison with empirical Bayes methods while considering the binomial distribution and estimation of its parameter – probability of success. The results are satisfactory, even on a very small sample, the proposed method worked really well compared to the empirical Bayes methods. Thus after fixing the value of Kuhn-Tucker multipliers, which is the aim of our future work, the method has a great potential in small sample theory and many other fields of statistics.

6 Appendix

6.1 Determination of the estimate $\hat{\pi}(h|D)$ of the posterior pdf $\pi(h|D)$

To determine the estimate of the posterior pdf $\pi(h|D)$ we focus on the Kuhn-Tucker function of the optimization task (2) and arrange it as follows:

$$\begin{aligned} \mathrm{L}(\widetilde{\pi}(h|D);\boldsymbol{\lambda}(D)) = &\int_H \widetilde{\pi}(h|D)\log\left(\frac{\widetilde{\pi}(h|D)}{\frac{\prod_{i=1}^{s} h(x_i)^{(\sum_{j=1}^{s}\lambda_j(D)g_j(x_i)+1)-1}}{Z(\lambda_1(D),\ldots,\lambda_s(D))}}\right)\mathrm{d}h \\ &- \log Z(\lambda_1(D),\ldots,\lambda_s(D))\underbrace{\int_H \widetilde{\pi}(h|D)\mathrm{d}h}_{=1} - \sum_{j=1}^{n}\lambda_j(D)\beta_j(D) \\ &- \lambda_{s+1}(D)\left(\int_H \widetilde{\pi}(h|D)dh - 1\right), \end{aligned} \tag{7}$$

where $Z(\lambda_1(D),\dots,\lambda_s(D))$ is a normalizing constant, $\lambda_j(D) \geq 0$ are Kuhn-Tucker multipliers, $j = 1,\dots,s+1$ and $\boldsymbol{\lambda}(D) = (\lambda_1(D),\dots,\lambda_s(D))$. According to the properties of the Kullback-Leibler divergence KLD(.,.) [9], the first term is minimal for $\widetilde{\pi}(h|D)$ being the pdf of the Dirichlet distribution with parameters $\sum_{j=1}^{s} \lambda_j(D) g_j(x_i) + 1$, $i = 1,\dots,n$. The last term of (7) is equal to zero, the rest does not depend on $\widetilde{\pi}(h|D)$ and does not influence the minimization. Thus the estimate $\hat{\pi}(h|D)$ of the posterior pdf $\pi(h|D)$ in (2) is a pdf of Dirichlet distribution with parameters mentioned above.

6.2 Determination of the Kuhn-Tucker multipliers

In this subsection we derive the formula for Kuhn-Tucker multipliers $\lambda_j(D)$ arising in the optimization task (2) and playing the key role in the combination (3). Thus we compute the first derivatives of the Kuhn-Tucker function (7) with respect to $\lambda_j(D)$, $j = 1,\dots,s$ and set each derivative equal to zero in order to find a minimum of this Kuhn-Tucker function. We omit the first and the last term of considered Kuhn-Tucker function from differentiation. The first term is already minimized - $\hat{\pi}(h|D)$ is a pdf of Dirichlet distribution $Dir(1 + \sum_{j=1}^{s} \lambda_j(D) g_j(x_i), i = 1,\dots,n)$ and according to the properties of the Kullback-Leibler divergence we have $\mathrm{KLD}(\hat{\pi}(h|D)||\hat{\pi}(h|D)) = 0$. Since $\hat{\pi}(h|D)$ is a pdf, the last term is equal to zero.

The first derivative of (7) with respect to $\lambda_k(D)$ looks then as follows:

$$
\begin{aligned}
&\frac{\partial}{\partial \lambda_k}\left(-\log Z(\lambda_1(D),\dots,\lambda_s(D)) - \sum_j \lambda_j(D)\beta_j(D)\right)\\
&= \frac{\partial}{\partial \lambda_k}\left(-\log \frac{\prod \Gamma(1+\sum_j \lambda_j(D) g_k(x_i))}{\Gamma(n+\sum_j \lambda_j(D))}\right) - \beta_k(D)\\
&= -\sum_i \psi\left(1+\sum_j \lambda_j(D) g_k(x_i)\right) g_k(x_i) + \psi\left(n+\sum_j \lambda_j(D)\right) 1 - \beta_k(D)\\
&= -\sum_i \psi_i g_k(x_i) + \psi_0 - \beta_k(D) \quad \forall \lambda_j, j = 1,\dots,s, \qquad (8)
\end{aligned}
$$

where ψ is the digamma function, see [2].

By using one-sided inverse - left inverse - we obtain the following system of nonlinear equations:

$$
\begin{aligned}
-D_{(s\times n)}\boldsymbol{\psi}_{(n\times 1)} + \boldsymbol{\psi}_{0,(s\times 1)} &= \boldsymbol{\beta}_{(s\times 1)}\\
-D_{(s\times n)}\boldsymbol{\psi}_{(n\times 1)} &= \boldsymbol{\beta}_{(s\times 1)} - \boldsymbol{\psi}_{0,(s\times 1)}\\
I_n \boldsymbol{\psi}_{(n\times 1)} &= -D^{-1}_{\mathrm{left},(n\times s)}\left(\boldsymbol{\beta}_{(s\times 1)} - \boldsymbol{\psi}_{0,(s\times 1)}\right)\\
\boldsymbol{\psi}_{(n\times 1)} &= -D^{-1}_{\mathrm{left},(n\times s)}\boldsymbol{\beta}^{*}_{(s\times 1)}, \qquad (9)
\end{aligned}
$$

where $D = (g_1, \ldots, g_s)^T$ (see Section 2). Thus:

$$\begin{array}{ccc} \psi(1 + \sum_j \lambda_j(D) g_j(x_1)) = \sum_j -D^{-1}_{\text{left},1j} \beta^*_j(D) \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ \psi(1 + \sum_j \lambda_j(D) g_j(x_n)) = \sum_j -D^{-1}_{\text{left},nj} \beta^*_j(D) \end{array}$$

and to obtain the multipliers we use the inverse digamma function:

$$\begin{array}{ccc} \sum_j \lambda_j(D) g_j(x_1) = \psi^{-1}(\sum_j -D^{-1}_{\text{left},1j} \beta^*_j(D)) - 1 \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ \sum_j \lambda_j(D) g_j(x_n) = \psi^{-1}(\sum_j -D^{-1}_{\text{left},nj} \beta^*_j(D)) - 1. \end{array}$$

The results using matrix notation are the following:

$$\begin{aligned} (D^T)_{(n\times s)} \boldsymbol{\lambda}_{s\times 1} &= (\psi^{-1}(-D^{-1}_{\text{left},(n\times s)} \boldsymbol{\beta}^*_{(s\times 1)}))_{(n\times 1)} - \mathbf{1}_{(n\times 1)} \\ \boldsymbol{\lambda}_{(s\times 1)} &= (D^T)^{-1}_{\text{left},(s\times n)} \left((\psi^{-1}(-D^{-1}_{\text{left},(n\times s)} \boldsymbol{\beta}^*_{(s\times 1)}))_{(n\times 1)} - \mathbf{1}_{(n\times 1)} \right). \end{aligned} \tag{10}$$

6.3 Determination of the final combination

We exploit the formula for the expected value of random vector having Dirichlet distribution $Dir(1 + \sum_{j=1}^{s} \lambda_j(D) g_j(x_i), i = 1, \ldots, n)$ (see Subsection 6.1) and conclude the following:

$$\begin{aligned} \mathrm{E}_{\hat{\pi}(h|D)}(h(x_i)|D) = \hat{h}(x_i) &= \frac{1 + \sum_{j=1}^{n} \lambda_j(D) g_j(x_i)}{\sum_{i=1}^{n} \left(1 + \sum_{j=1}^{s} \lambda_j(D) g_j(x_i)\right)} \\ &= \frac{1 + \sum_{j=1}^{n} \lambda_j(D) g_j(x_i)}{n + \sum_{j=1}^{s} \lambda_j(D)} \end{aligned}$$

for $i = 1, \ldots, n$.

References

1. Abbas, A.E. : A Kullback-Leibler View of Linear and Log-Linear Pools. Decision Analysis, vol. 6/1, pp. 25–37 (2009)
2. Abramowitz, M. and Stegun, I. A. (Eds.): Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover Publications, New York (1972)
3. Bernardo, J.M.: Expected Information as Expected Utility. Ann. Stat., vol. 7, pp. 686–690 (1979)
4. Clemen, R.T. and Winkler, R.L.: Combining Probability Distributions From Experts in Risk Analysis. Risk Analysis, vol. 19/2, pp. 187–203 (1999)
5. Copas, J. B.: Empirical Bayes Methods and the Repeated Use of a Standard. Biometrika, vol. 59, pp. 349–360 (1972)

6. Fine, T.L.: Theories of Probability: An Examination of Foundations. Academic Press, London (1973)
7. Griffin, B.S. and Krutchkoff, R.G.: Optimal Linear Estimators: An Empirical Bayes Version with Application to the Binomial Distribution. Biometrika, vol. 58, pp. 195–201 (1971)
8. Kerridge, D.F.: Inaccuracy and Inference. J. R. Stat. Soc., Ser. B, vol. 23, pp. 184–194 (1961)
9. Kullback, S. and Leibler, R.A.: On information and sufficiency. Ann. Math. Stat., vol. 22, pp. 79-86 (1951)
10. Lemon, G.H. and Krutchkoff, R.G.: An Empirical Bayes Smoothing Technique. Biometrika, vol. 56, pp. 361–365 (1969)
11. Maritz, J.S. : Empirical Bayes Methods. Methuen's Monographs on Applied Probability and Statistics. Methuen and Co Ltd., London (1970)
12. Martz, H.F. and Lian, M.G. : Empirical Bayes Estimation of the Binomial Parameter. Biometrika, vol. 61/3, pp. 517–523 (1974)
13. Sečkárová, V. : On Supra-Bayesian Weighted Combination of Available Data Determined by Kerridge Inaccuracy and Entropy. Pliska Stud. Math. Bulgar., vol. 22, pp. 159-168 (2013)
14. Shore, J.E. and Johnson, R.W.: Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-entropy. IEEE Trans. Inf. Theory, vol. 26, pp. 26–37 (1980)

Estimating Efficiency Offset between Two Groups of Decision-Making Units

Karel Macek[1]

Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic,
Pod Vodárenskou věží 4, Prague 8, 180 00, Czech Republic, karel.macek@utia.cas.cz

Keywords: Data Envelopment Analysis; Local Regression; Efficiency Comparison; Interval Estimation

Abstract. The comparison of two groups of decision-making units (DMUs) has been already subject of scientific reflection. So far, some statistical tests have been developed. This article addresses estimating the difference between expected outputs of two groups of DMUs. In contrast to other efficiency evaluation methods, this publication focuses on quantitative assessment of this difference, not on the hypothesis testing. The article focuses on single output DMUs and the designed statistical tests are examined on various simulated data sets as well as on one real-world example. Some of them stem from the data envelopment analysis, others are related to the local regression.

1 Introduction

Efficiency evaluation of decision making units (DMUs) has attracted the attention since 1957 [1]. Later on, a comparison of two or more groups of decision making units extended the basic framework of individual efficiencies.

One of the most important efforts was [2] where the statistical foundation of the data envelopment analysis (DEA) is introduced as well as a statistical test dedicated to comparison of efficiency offgroups in two groups of DMUs. This work was extended to 5 alternatives in [3].

This work offers a way how to examine the expected offset between two groups of DMUs. Some of them are based on simple tests from [3] while others apply local regression as a benchmark [4, 5]. The two groups of DMUs could correspond to farms in with green vs. classic approach or branches of two banks. Finally, those two groups could correspond to operation of two systems (e.g. factories) or one system with two different configurations (e.g. one factory using two different production programs). These examples are summarized in Table 1. In this table, the three first rows focus on the qualitative assessment which was mainly based on the hypothesis testing. This can be used for determination if there are some differences in the efficiency between the two groups of the DMUs. The last row in the table is our contribution, where it is also important to quantify the efficiency in an explicit way. If a new technology, a decision support system, or an outsourced service are paid and bought, it is important to quantify the value added by a new approach. In case of so called performance contracts [1], this quantification

[1] See e.g. http://www1.eere.energy.gov/femp/financing/espcs.html

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

determines the payments to the provider of the new approach and improves the competition between providers.

Table 1. Examples of offset measurement of two groups of DMUs

DMUs	Inputs	Output	Groups differ by	Purpose of offset evaluation
Family farmers [6]	cultivated area, working days	net income	traditional vs. green farming	is green farming sufficient for the families
Coffee retailers [7]	costs of goods sold; sales, general, and administrative expenses; depreciation/amortization	revenue	fair-trade vs. others	competitiveness impact of socially responsible sourcing
Universities [8]	staff; non-personnel expenditures	students, publications, third party funds	German vs. Swiss	evaluation of EU initiated reforms
Operation days of a building [5]	daily average of the ambient temperature	power consumption	original vs. new controller of the HVAC system	evaluation of savings achieved by the new technology

At more general level, the efficiency evaluation of DMUs is of high importance in large-scale distributed systems where the quality of different decision making approaches has to be evaluated. This can lead to propagation of positive experience within a DMU network.
The text is organized as follows: Sect. 2 introduces used notation. The notation is used in Sect. 3 for the problem formulation, i.e. the estimation of output offset between two groups of DMUs. Consequently, some estimates of the offset are provided in Sect. 4. Those estimates are examined on both simulated and real data in Sect. 5. The text is concluded in Sect. 6 by a short summary.

2 Notation

Before we will define specific notation for the addressed domain, we introduce some general notation. We will use $\mathbb{N}$ for natural numbers, $\mathbb{R}$ for the set of real numbers and $\mathbb{R}^N$ for N dimensional real vectors. The equality by definition is

denoted by $\equiv$. The conditioned probability density function are denoted as $f(\cdot|\cdot)$ and are distinguished by their arguments. The conditioned expected value is defined as $\mathcal{E}[a|b] \equiv \int af(a|b)\mathrm{d}a$.
Let us consider two groups of DMUs. Each DMU transforms the inputs to output and each group can use different mechanisms for this transformation. Mathematically, each DMU has a single output $y \in \mathbb{R}$ and a vector input $\mathbf{x} \in \mathbb{R}^m$, having dimension $m \in \mathbb{N}$. For the comparison, we have data in form $D = (\mathbf{x}^{(i)}, y_i, k_i)_{i=1}^n$ where i are indices of data , $\mathbf{x}^{(i)}$ denotes ith vector with components $\mathbf{x}_j^{(i)}$, and $k_i \in \{1,2\}$ is an index of the group. We assume that the DMUs within each group are homogeneous, i.e. the input-output transformation is described by the probability density function $f(y|\mathbf{x},k)$. This dependency can be modeled using a reference r output and noise terms, i.e.:

$$y = r(\mathbf{x}) + u_k \tag{1}$$

where $r : \mathbb{R}^m \to \mathbb{R}$ and u_k is a general noise term, not necessarily zero-mean. This model assumes that the noise u_k is dependent on the group, but not on the inputs. The tests introduced in [3] do this assumption which is from our point of view not very realistic. In the real situations the output variance might depend on the inputs. Typically, the more input, the higher variance. Therefore, we introduce also an alternative model instead of (1)

$$y = r(\mathbf{x}) + v_k(\mathbf{x}). \tag{2}$$

where v_k is a noise, depending on $\mathbf{x}$.
Furthermore, we assume that each DMU operates under different conditions and using different inputs. Thus each group has its typical inputs and conditions. Therefore we assume the inputs to have pdf $f(\mathbf{x}|k)$. Finally, we introduce the probability that a randomly selected DMU will belong to the first group $\rho \equiv \mathbf{P}(k=1)$. Then the marginal pdf of $\mathbf{x}$ is

$$f(\mathbf{x}) = \rho f(\mathbf{x}|k=1) + (1-\rho) f(\mathbf{x}|k=2). \tag{3}$$

3 Problem Formulation

Consider we have a given input $\tilde{\mathbf{x}} \in \mathbb{R}^m$. We let one DMU from both groups transform this input. First, the expected difference of outputs for given $\tilde{\mathbf{x}}$ is:

$$\delta(\mathbf{x}) \equiv \mathcal{E}[y|\tilde{\mathbf{x}}, k=2] - \mathcal{E}[y|\tilde{\mathbf{x}}, k=1]. \tag{4}$$

Next, the expected average difference equals:

$$\Delta \equiv \mathcal{E}[\delta(\mathbf{x})] = \int_{\mathbb{R}^m} \delta(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}. \tag{5}$$

Note that in case of independence of noise on the input (1), it holds

$$\Delta = \delta(\mathbf{x}) \qquad \forall \mathbf{x} \in \mathbb{R}^m. \tag{6}$$

This proposition can be proved by application of the definition in (1), definitions of δ and Δ. The application of the additivity to (4) leads to elimination of the x dependent parts. A formal proof is beyond the scope of this text. We use the

simpler model (1 because we need to know the analytical value of Δ which are carried out in Sect. 5. In case of (2), we it seems to be necessary to approach the real value of Δ using intensive Monte-Carlo simulations.

The problem addressed in this text consists in estimating $\delta(\mathbf{x})$ and Δ from available data. This estimation of the offset between two groups of DMUs can be of three forms: (i) a point estimate $\hat{\Delta}$, (ii) an interval estimate $(\hat{\Delta}_{\min}, \hat{\Delta}_{\max})$, or (iii) a posterior pdf $f(\Delta|D)$ where D are the available data.

4 Considered Estimates

In this Sect. we provide information on the considered estimates of Δ. First, we will introduce the benchmarking models that are data-driven and have minimal assumptions about the structure of (1) and (2). Note we work with a common benchmark for data from both groups[2]. Consequently, we will describe the algorithm for the estimation of Δ where the benchmarking models are used at the first step.

DEA Benchmarking. As proposed in [3], the reference r which has been introduced in (1-2) can be estimated as follows, corresponding to the BCC[3] model [9]:

$$\hat{r}_{\text{bcc}}(\tilde{\mathbf{x}}) = \max\{\phi| \tag{7}$$

$$\sum_{i=1}^{n} \lambda_i y_i = \phi; \tag{8}$$

$$\sum_{i=1}^{n} \lambda_i \mathbf{x}_j^{(i)} \leq \tilde{\mathbf{x}}_j, \forall j = 1, \dots m; \tag{9}$$

$$\sum_{i=1}^{n} \lambda_i = 1; \tag{10}$$

$$\lambda_i \geq 0, \quad \forall i = 1 \dots n\} \tag{11}$$

Let us interpret this reference briefly. For a given $\tilde{\mathbf{x}}$, we are looking for a maximal combined output (8) while the combined inputs are limited by the given one (9). The allowed combinations are convex as stated in conditions on $\lambda_1 \dots \lambda_n$ in (10), (11).

This estimate can be calculated by solving a linear programming problem. We mention also usual modification of the basic DEA apprach. First, we consider the

[2] It is possible also the creation two benchmarks, but this is not addressed in this article. Elsewhere [5], we clarify the motivation for a common benchmark for both groups carefully.

[3] BCC stands for Barker, Charnes, and Cooper who were authors of this model.

FDH[4] model [10]

$$\hat{r}_{\text{fdh}}(\tilde{\mathbf{x}}) = \max\{\phi| \tag{12}$$

$$\sum_{i=1}^{n} \lambda_i y_i = \phi; \tag{13}$$

$$\sum_{i=1}^{n} \lambda_i \mathbf{x}_j^{(i)} \leq \tilde{\mathbf{x}}_j, \forall j = 1, \ldots m; \tag{14}$$

$$\sum_{i=1}^{n} \lambda_i = 1; \tag{15}$$

$$\lambda_i \in \{0,1\}\} \tag{16}$$

Then also the basic DEA model denoted as CCR[5] [11]

$$\hat{r}_{\text{ccr}}(\tilde{\mathbf{x}}) = \max\{\phi| \tag{17}$$

$$\sum_{i=1}^{n} \lambda_i y_i = \phi; \tag{18}$$

$$\sum_{i=1}^{n} \lambda_i \mathbf{x}_j^{(i)} \leq \tilde{\mathbf{x}}_j, \forall j = 1, \ldots m; \tag{19}$$

$$\lambda_i \geq 0\} \tag{20}$$

The value of the benchmark can be calculated for each $x \in \mathbb{R}^{N_x}$ using a linear programming procedure where ϕ stands for the objective function and the equalities and inequalities for constrains. The DEA approaches differ each other with respect to the conditions on λ_i, as it can be seen in (10), (11), (15), (16), and (20). One can see that conditions for FDH are the most strict while for the CCR are the less strict. Thus, the FDH will rate more units efficient than BCC or CCR.

Local Regression Benchmarking. Another way how to construct a benchmark is a local polynomial regression [4] - abbreviated as LPR - of the degree p. We will use only the single input version, i.e. $m = 1$:

$$\hat{r}_{\text{lpr}}(\tilde{\mathbf{x}}) = \sum_{i=1}^{n} \lambda_i(\tilde{\mathbf{x}}) y_i \tag{21}$$

where the vector $\lambda(\mathbf{x})^T = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ is calculated as

$$\lambda(\tilde{\mathbf{x}})^T = \mathbf{e}^T (X_{\tilde{\mathbf{x}}}^T W_{\tilde{\mathbf{x}}} X_{\tilde{\mathbf{x}}})^{-1} X_{\tilde{\mathbf{x}}}^T W_{\tilde{\mathbf{x}}} \tag{22}$$

where $\mathbf{e} = (1, 0, \ldots, 0) \in \mathbb{R}^{p+1}$

$$X_{\tilde{\mathbf{x}}} = \begin{pmatrix} 1 & \mathbf{x}_1^{(1)} - \tilde{\mathbf{x}} & \ldots & \frac{(\mathbf{x}_1^{(1)} - \tilde{\mathbf{x}})^p}{p!} \\ 1 & \mathbf{x}_1^{(2)} - \tilde{\mathbf{x}} & \ldots & \frac{(\mathbf{x}_1^{(2)} - \tilde{\mathbf{x}})^p}{p!} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbf{x}_1^{(n)} - \tilde{\mathbf{x}} & \ldots & \frac{(\mathbf{x}_1^{(n)} - \tilde{\mathbf{x}})^p}{p!} \end{pmatrix} \tag{23}$$

[4] FDH stands for free disposable hull.

[5] CCR stands for Charnes, Cooper, and Rhodes who were authors of this model.

and $W_{\tilde{\mathbf{x}}}$ is a diagonal matrix where $w_i(\tilde{\mathbf{x}}) = K((\mathbf{x}_1^{(i)} - \tilde{\mathbf{x}})/h)$ is the weight. Parameter $h > 0$ is a bandwidth parameter and K is a kernel function. We adopted the Gaussian kernel:

$$K(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \tag{24}$$

Other applicable kernels, like Epanechnikov kernel, as well as general properties of kernel functions are described e.g. in [4].
In contrast to DEA approaches, the local polynomial regression has two parameters, namely the bandwidth h and p and the quality of fit depends on them. The optimization of p can be done using systematic search since the set finite and small. For each p, the bandwidth h can be optimized using leave-one-approach, details are provided in [4].

Estimating Offset between Two Groups of DMUs. Now, we are about to describe the method itself. We will do it an a step-by-step way:

1. **Benchmark on data** - we use the data D to calculate the $\hat{r}(\mathbf{x}^{(i)})$. We can use one of the benchmarking models provided in the previous paragraphs.
2. **Calculation the residuals** - we introduce residuals as

   $$e_i = y_i - \hat{r}(\mathbf{x}^{(i)}) \qquad \forall i = 1, \ldots, n(.)$$

3. **Fitting the residuals** - we calculate the regression model. We adopted the ordinary least square approach with an indicator (dummy) variable as used in [3] that is applicable in the case of model (1).

   $$e_i = \beta_0 + \beta_1 \mathbf{1}(k_i = 2) + \xi$$

   where $\mathbf{1}(k_i = 2)$ is indicator that the ith DMU belongs to group 2 and ξ is a zero mean noise.
4. **Point estimate of the offset** - we interpret a regression parameter as a point estimate, namely the regression parameter β_1 can be interpreted as an estimate of Δ because it expresses the average difference between both groups of DMUs.
5. **Estimating the variance of the offset** - we use usual statistical inference on linear regression parameters for variance of the offset, since linear regression [12] estimates β_1 and consequently its variance, too. The estimate will be denoted as b_1 and the variance $SE(b_1)$. It is calculated as follows:

   $$SE(b_1) = s \sqrt{\frac{n}{n\sum_{i=1}^{n}(k_i^2) - \left(\sum_{i=1}^{n} k_i\right)^2}}$$

   where

   $$s = \sqrt{\frac{\sum_{i=1}^{n}(e_i - \hat{e}_i)^2}{n-2}}.$$

   with $\hat{e}_i$ being output from the regression model in (25)
6. **Construction of interval estimate of the offset** - using this pdf, we can provide the interval estimate as

   $$\begin{aligned} \Delta_{\min} &= b_1 + t_{n-2}(\alpha/2) SE(b_1) \\ \Delta_{\max} &= b_1 + t_{n-2}(1-\alpha/2) SE(b_1) \end{aligned}$$

   for given level of significance $\alpha \in [0,1]$ where t_{n-2} denotes the cdf of Student distribution with n degrees of freedom.

7. **Density estimation of the offset** - finally, the pdf of Δ is as follows:

$$f(\Delta|D) = f_{n-2}\left(\frac{\Delta - b_1}{SE(b_1)}\right) \tag{25}$$

where f_{n-2} is pdf of Student distribution with $n-2$ degrees of freedom and $SE(b_1)$ is defined in (25).

5 Comparison on Simulated Data

In this Sect. we evaluate the proposed methods on the simulated data so we can evaluate their quality. The data are simulated from models like (2). The evaluation methods use the simulated data only without any knowledge of the models used. We examine general methods for the offset estimation without any prior knowledge and we test if the methods are able to fit the unknown model sufficiently. The use of the simulated data allows us to calculate the offsets analytically from the models and compare them with the data-driven estimates.

5.1 Simulated Data

The following examples are dedicated to the numerical tests and our primary focus is not their real interpretation[6].

Example 1 offers a monotonous, concave function, as expected for the DEA estimation [2] and the noise u_k is left-half-normally distributed.

$$\begin{aligned} r(x) &= -x^2 + 2x + 15 \\ u_1 &= -|d_1| \qquad d_1 \sim \mathcal{N}(0, \sigma_1) \\ u_2 &= -|d_2| \qquad d_2 \sim \mathcal{N}(0, \sigma_2) \\ \rho &= 1/2 \\ f(\mathbf{x}) &= 1 \qquad \forall \mathbf{x} \in [0,1] \end{aligned}$$

In this case $\Delta = \delta(\mathbf{x}) = \mathcal{E}[U_2] - \mathcal{E}[U_1] = (\sigma_2 - \sigma_1)\sqrt{2/\pi}$. For the experiments we used $\sigma_1 = 1$ and $\sigma_2 = 2$. The number of instances is $n = 200$.

Example 2 modifies the noise of the previous one so

$$\begin{aligned} u_1 &\sim \mathcal{N}(\mu_1, \sigma_1) \\ u_2 &\sim \mathcal{N}(\mu_2, \sigma_2) \end{aligned}$$

Then $\Delta = \mu_2 - \mu_1$. We use $\mu_1 = 1$ and $\mu_2 = 2$ and $\sigma_1 = \sigma_2 = 1$.

Example 3 has same structure as Example 1, but $r(x) = 4x$ for $k = 1$ and $r(x) = 5x$ otherwise. Furthermore, $\sigma_1 = \sigma_2 = 0.2$. From definition (6), it can easily inspected $\Delta = 0.5$ for this case.

[6] Possible interpretation of those models can relate to companies in a segment. The input x can be interpreted as the market share of a company. The output can be the operational costs of the company.

Evaluation Approach We have formulated a set of problems, where the exact values of Δ are analytically known. To compare quality of the proposed estimates, we run the identification procedure $n_s = 100$ times. Thus, we obtain n_s estimates. We want to calculate how they match the exact values. We measured:

- The quality of the point estimate using MSE $= \frac{1}{n_s}\sum_{s=1}^{n_s}(\hat{\Delta}_s - \Delta)^2$ which should be as small as possible,
- The quality of interval estimate using scores whether the exact value is within the interval $[\Delta_{s,\min}, \Delta_{s,\max}]$

$$I = \frac{1}{n_s}\sum_{s=1}^{n_s}\mathbf{1}(\hat{\Delta}_{s,\min} \leq \Delta \leq \hat{\Delta}_{s,\max})$$

  and it holds that $I \geq 1-\alpha$ is a good result and $I = 1-\alpha$ is a very good result. We used the level of significance $\alpha = 0.05$.
- The probability distribution function as the logarithm of the joint probability $L = \sum_{s=1}^{n_s}\log f(\Delta|D)$ which should be as big as possible.

Results Tables 1–3 show the results for given tests on the formulated examples. Table 1 shows quite good results in the *MSE* and high *L* for all methods with the exception of *CCR* which fails in all examples[7]. The interval estimates seem not to be very satisfactory since the value should be 0.95 or greater. Only FDH with the value of 0.88 approaches this value. Table 2 shows results for normally distributed noise which are not so good as in the previous case. Table 3 is the only one where CCR was successful. It could be assumed since CCR is dedicated to proportional dependencies between inputs and outputs.

Table 2. Results for Example 1

Approach	*MSE*	*I*	*L*
BCC	0.052	0.640	-5.180
CCR	1.2e8	0.980	-1.2e10
FDH	0.036	0.880	-.3.640
LPR	0.135	0.510	-13.464

5.2 Real Data

We used the same data as in [5] where we assessed the savings achieved by improved control of a heating, air-conditioning and ventilation system. The data set consists of 200 records from an HVAC control system containing (i) daily gas consumption y, (ii) average daily ambient weather x, and index of strategy used during given day k. Since the lower temperature, the higher heating, we used negative values of the ambient temperature. From Fig. 1 it can be seen the CCR leads

[7] The success in the interval estimates for CCR is given by a very wide variance.

Table 3. Results for Example 2

Approach	*MSE*	*I*	*L*
BCC	3.982	0.000	-398.246
CCR	9.0e5	0.960	-9.05e7
FDH	4.142	0.000	-414.182
LPR	3.383	0.000	-338.336

Table 4. Results for Example 3

Approach	*MSE*	*I*	*L*
BCC	1.01	0	-175.74
CCR	0.99	0	-173.93
FDH	1.24	0.32	-4.39
LPR	0.92	0	-199.14

to very flat pdf. Other methods demonstrate the savings, but their estimates of $f(\Delta)$ differ.
From the practical applicability of this evaluation framework, following conclusion can be drawn: if the average achieved savings Δ would be a part of a contract (e.g. the customer pays a ratio of the savings back to the provider), the used evaluation method have to be specified.

6 Conclusions and Future Work

In this text we introduced offset between two groups of DMUs and discussed ways how to estimate it. These estimates have been tested on three examples. We have shown that DEA based estimates are more appropriate for cases where the usual DEA assumptions (one sided noise) are satisfied. The local regression approaches are applicable where those assumptions are not satisfied. Next research shall focus on the estimates of $\delta(\mathbf{x})$ that are of practical importance for evaluation of changes in particular DMUs. Theoretical aspects of the proposed tests shall be subject of deeper investigation since the estimates are not very satisfactory in two of the three examples. Finally, the generalization for multiple-output models shall be addressed.

7 Acknowledgement

This research has been supported by GAČR 13-13502S. The author wishes to express his sincere thanks to Tatiana Valentine Guy and Miroslav Kárný for selfless help with editing and their support in general.

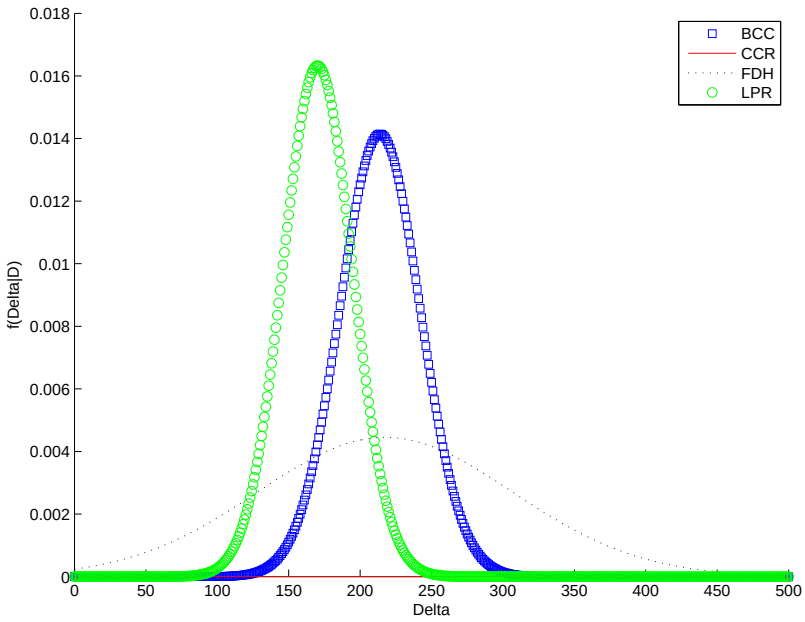


Fig. 1. Pdfs of Δ for particular approaches. BCC and LPR seem to be informative while CCR is practically flat. The achieved savings are very likely between 100 and 300.

References

1. Farrell, M.J.: The Measurement of Productive Efficiency. Journal of the Royal Statistic Society **120**(III) (1957) 253–281
2. Banker, R.: Maximum likelihood, consistency and data envelopment analysis: A statistical foundation. Management Science **39** (1993) 1265–1273
3. Banker, R., Zheng, Z., Natarajan, R.: Dea-based hypothesis tests for comparing two groups of decision making units. European Journal of Operational Research **206**(1) (2010) 231–238
4. Wasserman, L.: All of Nonparametric Statistics. Springer, New York (2006)
5. Macek, K., Mařík, K.: A methodology for quantitative comparison of control solutions and its application to HVAC (heating, ventilation and air conditioning) systems. Energy **44**(1) (2012) 117–125
6. Gomes, E.G., de Mello, J.C.C.B.S., de Freitas, A.C.R.: Efficiency measures for a non-homogeneous group of family farmers. Pesuisa Operacional **32**(3) (2012) 561–574
7. Joo, S.J., Min, H., Kwon, I.W.G., Kwon, H.: Comparative efficiencies of specialty coffee retailers from the perspectives of socially responsible global sourcing. The International Journal of Logistics Management **21** (2010) 490–509
8. Olivares, M., Schenker-Wicki, A.: The dynamics of productivity in the swiss and german university sector: A non-parametric analysis that accounts for heterogeneous production. Technical report, University of Zurich (2012)
9. Banker, R.D., Charnes, A., Cooper, W.W.: Some Models for Estimating Technical and Scale Inefficiencies inData Envelopment Analysis. Management Science **30** (1984) 1078–1092
10. Deprins, D., Simar, L., H.Tulkens: Measuring labor efficiency in post offices. In Marchand, M., Pestieau, P., Tulkens, H., eds.: The Performance of Public Enterprises: Concepts and Measurement. North Holland (1984) 243–257
11. Charnes, A., Cooper, W., Rhodes, E.: Measuring the efficiency of decision making units. European Journal of Operational Research **2**(6) (1978) 429 – 444
12. Kutner, M.H., Nachtsheim, C.J., Neter, J.: Applied Linear Regression Models. Fourth international edn. McGraw-Hill/Irwin (September 2004)

Economic Prediction using Heterogeneous Data Streams from the World Wide Web

Abby Levenberg[1], Edwin Simpson[2], Stephen Roberts[1,2], and Georg Gottlob[1,3]

[1] Oxford-Man Institute of Quantitative Finance,
University of Oxford,
abby.levenberg@oxford-man.ox.ac.uk
[2] Machine Learning Research Group,
Department of Engineering Science,
University of Oxford,
sjrob@robots.ox.ac.uk, edwin@robots.ox.ac.uk
[3] Department of Computer Science,
University of Oxford,
georg.gottlob@cs.ox.ac.uk

Abstract. Learning to predict financial and economic variables of interest is a hard problem with a large body of literature devoted to it. Of late there has been a significant amount of work on using sources of *text* from the Web (such as Twitter or Google Trends) to predict financial and economic variables. Much of this work has relied on some form or other of superficial sentiment analysis to represent the text. In this work we present a novel approach to predicting economic variables using multiple heterogenous streams of Web data. We can incorporate different data types into our model – such as time series and text – by first treating each data stream as a separate source with its own features and predictive distribution. For the text data streams we use a novel approach to prediction using a sentiment composition model to generate features. We then use a Bayesian classifier combination model to combine the independent "weak" predictions into a single prediction of the Nonfarm Payroll index, a primary economic indicator. Our results show that using a classifier combination model over multiple streams can achieve very high predictive accuracy.

Keywords: heterogeneous data streams, economic prediction, classifier combination, text sentiment

1 Introduction

There is a vast amount of data available on the Internet from a huge number of distinct online sources and the rate of its output is increasing daily. Currently there is significant interest in both industrial and academic research that aims to utilize such *big data* provided by the WWW to make predictions and gain insights into various aspects of daily life. Of late there has been a lot of work using textual WWW data to make predictions of a financial nature attempting to find correlations between the data and various lead economic and financial indicators such as the stock market or employment rates. Structured extraction of and learning from these online sources of data is a useful and challenging problem that spans the machine learning, information extraction, and quantitive finance research communities.

In this work we forecast the trend of the United States *Nonfarm Payrolls* (NFP), a monthly economic index that measures employment growth (decay) and is considered an important indicator of the welfare of the U.S. economy.[4] The NFP index is part of the Current Employment Statistics Survey, a comprehensive report released by the United States Department of Labor, Bureau of Labor Statistics, on the state of the national labor market. Released on the first Friday of each month, the index is given as the *change* in the number of (nonfarm) employment compared to the prior month. Besides indicating the state of the economy, the NFP is an index that "moves the market" upon its release [17] with the market reacting positively to a increase in the index and negatively to a decline. It is of interest to anyone with an stake in the market, such as banks, hedge funds, prop traders, etc., to try

[4] http://research.stlouisfed.org/fred2/series/PAYNSA?cid=32305

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

and make an accurate and timely prediction of its direction. As such, as the NFP release data nears there is a significant amount of speculation in the media from economists attempting to forecast its direction and value.

We show that such a prediction is possible using freely available data from the WWW. We present a novel extraction and machine learning framework to access and combine features from heterogeneous *data streams* from disparate online sources. We make use of both text and real-valued streams. For the text streams we present a novel approach to prediction using features generated from a state-of-the-art sentiment composition algorithm. We combine these text streams with relevant timeseries data mined from the WWW. We use these streams to learn accurate predictions of the future trend of our economic variable of interest. Since each stream provides its own predictive distribution we show how to fully exploit the information from separate streams of various data types by using an Independent Bayesian Classifier Combination (IBCC) model to obtain high accuracy in our predictive task. Using features from multiple WWW streams is a contribution to the current literature and presents a number of challenges that we address in the following sections.

In the next section we review the relevant literature in the area of using WWW data to make economic predictions. In Section 3 we present a stream-based framework for online extraction for multiple unrelated heterogeneous data sources. We also describe the IBCC model which enables us to aggregate any number of stream specific base classifiers into a single prediction. In Section 4 we describe the data streams in further detail. In Section 5 we report on correlating our data streams with economic trends and using the complete streaming framework with the IBCC model to predict the NFP. We show results that are state-of-the-art.

2 Prior Work

Much previous work has concentrated on the combination of information sources of the same type. In this paper, however, we combine heterogeneous data streams, time series and textdata, to achieve robust prediction models. Our review is therefore divided into two categories of prior work: time series and text data.

2.1 Time Series Prediction

Arguably the whole field of quantitative financial analysis revolves around the ability to detect signals within and between time series data. As such there is a large body of literature on using time series data to predict financial variables of interest. Techniques range from simple heuristics based on intuition and market knowledge to state-of-the-art machine learning algorithms such as genetic algorithms and deep networks. Over the last decades many textbooks have been and continue to be published that describe a huge number of techniques for finding correlations between various financial and economic time series, for example [3], [8], and [11]. Numerous journals and conferences are devoted to disseminating the latest approaches for financial timeseries prediction and regression for analysts, traders, quants and academics. For example, the *Journal of Time Series Econometrics* and the *Journal of Time Series Analysis* are journals devoted entirely to publishing the latest findings in this area.

2.2 Text Prediction

Utilizing the information implicit in market news and opinion to predict the direction of the economy is of obvious interest to many people. As such there has been significant amount of work that uses text from various online sources for prediction of economic indexes and stock market trends (see [9, 18, 5] for instance). In general the framework of these papers is to obtain natural language text from the Web, such as news stories, message board data, Twitter feeds, etc., and to use language specific features, often sentiment based, to train a classification algorithm to predict the future direction or value of the index/market. Learning algorithms range from simple two-class Naive Bayes and Support Vector Machines to more sophisticated algorithms with varying results and claims.

An overview and comparison of a number of such predictive systems tailored specifically to the stock market is given in [12] and [14]. Some of the reported work describes trading strategies based on system predictions that perform well beyond market expectations. However, the authors suggest the systems

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

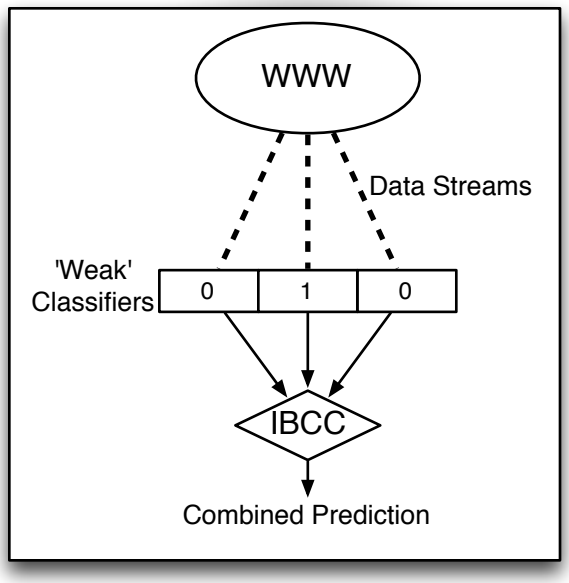


Fig. 1. Framework for prediction. We aggregate independent predictions from multiple heterogeneous data streams from the WWW into a single combined trend prediction using a Bayesian combination framework.

they review suffer from a lack of proper testing and unrealistic market expectations. As well, most of the systems reviewed in these summaries use a "bag of words" model to compute the features for the document-level classification. The authors argue this approach is far too general and accuracy from this is impacted due to the loss of contextual information from each document.

Current work has focussed on the use of big textual data to predict economic and market trends (see [4], [10], [16]). An example of note is [1]. Here the authors regressed from multidimensional sentiment moods (i.e., "Calm", "Happy") obtained from a stream of Tweets to the market and found some weak correlation with a single dimension of sentiment. While this research generated a significant buzz in the media and financial sectors its application to real-world trading remains unclear.

Other interesting work using text features for various predictions include the work described in the overview from [20] and [2]. Here a group of "text-driven forecasting" models are described that are used to predict phenomenon ranging from the volatility of yearly returns from financial reports, box office revenues from film critics reviews, and menu prices from the sentiment of costumer restaurant reviews. More recently [15] used *Google Trends* to find more significant correlations with changes in volumes of search queries of particular finanical terms and the lagged market trend.

3 Streaming Prediction Framework

Our goal is to efficiently use the big data freely available on the WWW to make predictions of economic variables of interest. However, for a given domain there is an overwhelming amount of data available from any number of sources. A simplifying conceptual approach for making sense of the abundance of Web data is to treat each online source of data as a separate *stream* of data. Each data stream has its own underlying distribution and throughput, the rate at which the source produces raw data, and hence its own independent level of predictive accuracy. If we treat each stream as a classifier in its own right we can make use of ensemble methods to combine the independent predictions into a single best prediction. As well, since each stream is considered independently this approach enables us to fuse multiple heterogeneous sources of data together into a single model of prediction. In this

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

section we describe a framework for data stream extraction and aggregate prediction using IBCC from independent "weak" classifiers built from multiple WWW streams.
As Figure 1 depicts our framework for stream-based prediction is divided into three parts:

1. Extracting the relevant streams from the Web in a structured and efficient manner.
2. Training an ensemble of base classifiers – one for each data stream – using features and models specific to each stream.
3. Aggregate multiple, stream-specific classifications into a single globally optimised prediction.

Below we describe in further detail each part of this framework.

3.1 Structured Stream Extraction

Since we aim to predict the trend of the economic index, the NFP, we want to find streams that contain useful information for predicting the economy. An immediate question we must answer is how to find and extract *only* the data relevant to the predictive task at hand from the massive amount of data available online. Consider that even within a stream from a single source there may be data that pertains to an arbitrary number of domains. For example, a stream of text from a website that broadcast news in real-time will contain stories ranging from the economy to celebrity surgery and everything in-between. We may want to use the pertinent articles on the economy from such a source but indiscriminate collection of the stream will mean most of the text we collect will be irrelevant to our predictive task.
Hence we use a mechanism based on *Oxpath*, a query language for web data extraction that enables the automation of user-driven queries of a given source and then structured retrieval of the returned data [6]. For instance, suppose we aim to collect articles pertaining to the NFP from the online archives of various newspapers and magazines. Using Oxpath we can setup an automated process to periodically query multiple sources for particular terms over specific dates, daily for instance, and save the data returned as structured entries into a local repository. This enables us to capture details present on a web page such as the author, title, date, etc., of an article. This means we do not have to download, process, and classify raw HTML pages from the web which is a tedious and error prone process. Instead we have direct structured access to the desired content of the stream.

3.2 Stream-specific Classifiers

Once we have access to the pertinent data from a particular data source we need to train a predictive model specific to that stream to forecast the NFP, our dependent variable. Here any of the standard machine learning models in the literature are viable. For example, since we are predicting the directional trend of an economic index we use simple binary logistic regression models where a class of 1 means "up" and 0 means "down". However, to use any predictive models we first must derive features from the raw streams to use as training data to our classifier.
In this work we use both real-valued time series and text data streams. For the time series data we can use standard multivariate features such as smoothed moving averages etc. For the text data we try something simple but new. First we use sentiment composition to score individual sentences with a distribution over positive, negative, or neutral sentiment [13]. Afterwards we combine these sentence-level sentiment features in some informative way as input into our training algorithms. In Section 5 we report experiments on various approaches for combining the sentiment distribution from individual sentences as input features for model training. Next we describe how we combine these stream-specific predictions into a single best prediction.

3.3 Binary IBCC Model

Due to the differences in their underlying distributions, each of the individual data stream's predictive accuracies may very enormously in reliability. Classifier combination methods are well suited to situations such as these and serve to make best use of the outputs of an ensemble of imperfect base classifiers to enable higher accuracy classifications. Using a *Bayesian* approach to classifier combination provides a principled mathematical framework for aggregation where poor predictors can be mitigated and in which multiple data streams, with very different distributions and training features, can be combined to provide complementary information [7]. Here we describe a binary, two-class variation of the IBCC model of [19]. [5]

[5] The full model for an arbitrary number of classes ≥ 2 is described in [19].

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

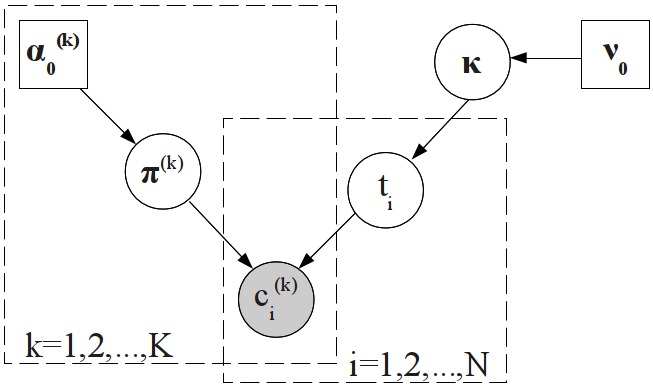


Fig. 2. Graphical model for IBCC. The arrows indicate dependencies while the shaded node represents observed variables, the square nodes are hyper-parameters. All other variables must be inferred. Here the predictions $\boldsymbol{c}_i^k$ of each base classifier k are generated dependent on the confusion matrices $\boldsymbol{\pi}^k$ and the true label t_i.

We want to predict the trend of the NFP over some number of months, or more generally *epochs*, indexed from $i \in \{1, \ldots, N\}$. We assume the trend $\boldsymbol{T}$ of the NFP is generated from an underlying binomial distribution with parameters $\boldsymbol{\kappa}$. Each epoch has a value $t_i \in \{0, 1\}$ where the ith epoch has a label $t_i = 0$ if the NFP index decreased from the prior epoch and $t_i = 1$ if it increased. The prior probabilities of the trends t_i are given by $\boldsymbol{\kappa}$ **:** $p(t_i = j|\boldsymbol{\kappa}) = \kappa_j$, where j iterates over the class labels $\{0, 1\}$.

We denote the number of base classifiers, or data streams, as K. Each stream's base classifier $k \in \{1, \ldots, K\}$ produces a real-valued output matrix $\hat{\boldsymbol{C}}^k$ of size $N \times j$. The output vector $\hat{\boldsymbol{c}}_i^k \in [0, 1]$ for epoch i denotes the probabilities given by classifier k of assigning a discrete trend label $\boldsymbol{c}_i^k \in \{0, 1\}$. The jth element of the trend label, $c_{ij}^k = 1$, while all other elements are zero, indicates that classifier k has assigned label j to epoch i. We assume the vector $\boldsymbol{c}_i^k$ is drawn from a binomial distribution dependent on the true label t_i, with probabilities $\boldsymbol{\pi}_j^k = p(\boldsymbol{c}_i^k|t_i = j, \boldsymbol{\pi}_j^k)$. Both parameters $\boldsymbol{\pi}^k$ and $\boldsymbol{\kappa}$ have Beta-distributed priors.

The joint distribution over all variables for the binary IBCC model is

$$p(\boldsymbol{\kappa}, \boldsymbol{\Pi}, T, \boldsymbol{C}|\mathbf{A}_0, \boldsymbol{\nu}) = \prod_{i=1}^{N}\{\kappa_{t_i} \prod_{k=1}^{K} \boldsymbol{\pi}_{t_i}^k \cdot \boldsymbol{c}_i^k\} p(\boldsymbol{\kappa}|\boldsymbol{\nu}) p(\boldsymbol{\Pi}|\mathbf{A}) \tag{1}$$

where $\boldsymbol{\Pi} = \{\boldsymbol{\pi}_j^k | j \in \{1, 0\}, k = 1 \ldots K\}$ denotes all base classifier probabilities, $\mathbf{A}_0 = \{\boldsymbol{\alpha_{0j}}^k | j \in \{1, 0\}, k = 1 \ldots K\}$ the corresponding set of hyper-parameters, and $\boldsymbol{\nu_0} = [\nu_0, \nu_1]$ are the hyper-parameters for $\boldsymbol{\kappa}$. A graphical model of IBCC is shown in Figure 2.

The probability of a test point t_i at epoch i being assigned class j is given by

$$p(t_i = j) = \frac{\rho_{ij}}{\sum_{y=1}^{J} \rho_{iy}} \tag{2}$$

where

$$\rho_{ij} = \kappa_j * \prod_{k=1}^{K} (\boldsymbol{\pi}_j^k \cdot \boldsymbol{c}_i^k) \tag{3}$$

which accounts for the probability of the class κ_j weighted by the combined prediction probabilities $\boldsymbol{\pi}_j^k$ of each stream's independent predictions $\boldsymbol{c}_i^k$.

A key feature of IBCC is that each base classifier k is modelled by $\boldsymbol{\pi}^k$, which intuitively represents a *confusion matrix* that quantifies the decision-making abilities of the indivdual base classifier k. The goal of inference for the model is to optimise the distributions over the unknown variables $\boldsymbol{T}$, $\boldsymbol{\Pi}$, and $\boldsymbol{\kappa}$ such that the probability of t_i for each epoch i is maximized for epochs with true increases in the NFP and minimised for epochs i where the NFP decreased. In [19] this approach has been shown to outperform a number of baseline combination methods for classification tasks.

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

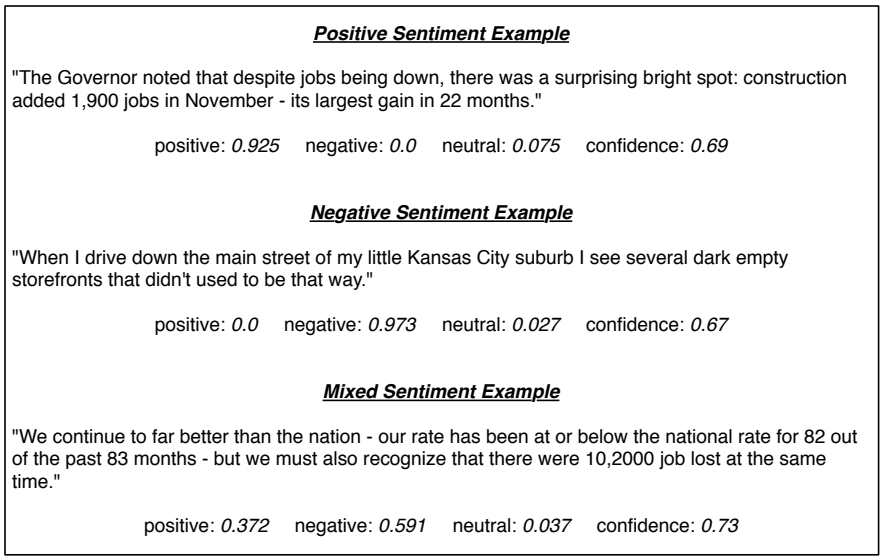


Fig. 3. Examples of sentences with different sentiment distributions accounting for the positive, negative, and neutral dimensions of a sentence.

4 Data

In this section we describe the data we collected to test our streaming prediction framework. Our timeline for training and testing spanned the NFP index monthly from January, 2000 through December, 2012. We made use of both time series data and text data from a variety of online sources described below.

We collected time series data online from a variety of online sources including the Federal Reserve Economic Data website [6], a Federal Reserve bank resource which compiles and maintains a large number of economic time series and data sets. Other sources included the Bureau of Labor Statistics [7] and the Conference Board [8] which both publish various economic indexes. We collected 33 different time series from such online sources to use as independent variables for predicting the NFP. We describe in detail our predictive models in the following section.

We also collected a number of textual data streams from multiple sources. For this we ran pointed queries against a large news database [9] and collected archived test data from nearly 700 distinct online text sources such as the Associated Press, Dow Jones, Wall Street Journal, etc. Altogether we collected over 6.6 million sentences of raw text from the streams.

After we collected the text data we processed the text at the sentence level for individual sentiment analysis using the model in [13] [10]. After sentiment analysis each sentence is represented as a distribution over three dimensions of sentiment: positive, negative, and neutral. Figure 3 shows some example results from the sentiment analysis system. In the next section we detail our experiments for prediction based on text, timeseries and their combination. using these sentiment dimensions as features.

5 Experiments

5.1 Experiment Setup

As described in Section 4 we collected data over a timeline of 13 years from 2000-2013 which contained 156 monthly epochs. We used the last 24 epochs as test points and the rest of the epochs in the timeline

[6] `http://research.stlouisfed.org/fred2/`
[7] `http://www.bls.gov/`
[8] `http://www.conference-board.org/`
[9] `http://www.dowjones.com/factiva/index.asp`
[10] The sentiment model we used is available as an API service from `http://theysay.io/`.

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

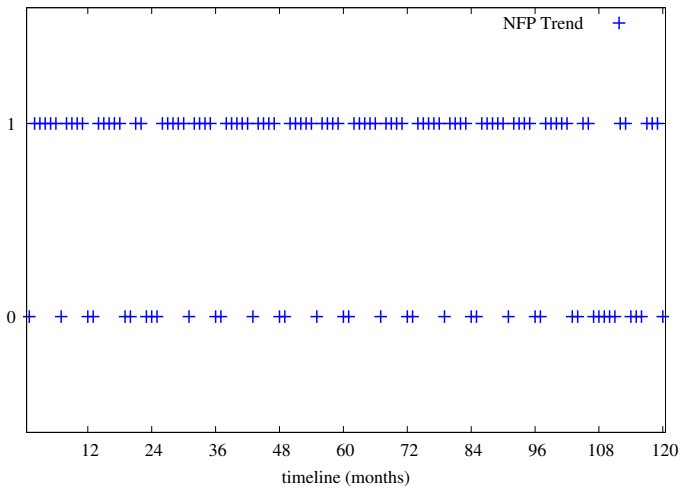


Fig. 4. The full NFP index trends over a 10 year timeline where 1=up, 0=down.

Source	AUC
Random	0.52
Always Up	0.50
Back Returns	0.54

Table 1. Baseline results for predicting the NFP index.

as training points. However, as the economy normally tends towards growth, save for in periods of recession, there is an over representation of 109(70%) positive cases compared to only 47(30%) negative instances in the NFP index since 2000. This is shown clearly in Figure 4 where the 1 on the y-axis indicates an upward trend of the NFP index and a 0 indicates a downward trend. To ascertain whether our approach is valid for learning good predictions rather than just optimising for the overrepresented class we subsampled randomly from the positive class to obtain a balanced training set.
For each data source we learnt a base classifier independently and used *rolling* predictions so the features associated with a given test point became part of the training data for the next test epoch. These models were then used as base inputs for IBCC. Note that the stream-specific classifiers need not give good individual prediction results as long as each contain useful information for the IBCC model. In fact base classifiers with very poor accuracy may be useful as IBCC can account for negative results so long as there is consistent information encoded in the probabilities. We measure our results using the standard metric Area Under the Receiver Operating Characteristic Curve (AUC). The AUC is the probability of ranking a positive example higher than a negative example and takes into account both true and false positive predictions [21]. For completeness, in the results that follow we show the results for the individual streams as well as the results using the IBCC model.

5.2 Baselines

Table 1 reports some baseline measures of prediction standard for the NFP. *Random* uses a random number generator to output a real number between $[0, 1]$ which it treats as prediction probabilities. *Always Up* always predicts the NFP as rising with a probability of 1. We also used the industry standard of *Back Returns* and predict each epoch will follow the trend of the last. Each of these achieve an AUC around 0.5 - as expected since subsampling makes the empirical priors of up and down equal and these methods do not have predictive power, with back returns performing at near random performance.

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

Source	Averages	Trends
Associated Press*	0.69	0.37
Dow Jones	0.44	0.25
Reuters News	0.46	0.36
Market News Intl.*	0.70	0.23
Other Sources*	0.63	0.63
Wall Street Journal	0.63	0.53
IBCC	0.81	**0.85**

Table 2. Stream-specific and combined AUC results for predicting the NFP index. We get better prediction accuracy using multiple sources (starred) with IBCC.

Source	AUC
CPI	0.70
ISM	0.85
JOLTS	0.66
LFL	0.71
IBCC	**0.90**

Table 3. Stream-specific and combined AUC results for predicting the NFP index using time series data. Here again accuracy is improved when using IBCC.

5.3 Text Stream Prediction

In this section we report on results predicting the NFP using the text streams both as independent classifiers and as base inputs to the IBCC model. Our general approach to using the sentiment features described in Section 4 is to aggregate the sentiment distributions over all sentences in an epoch and then use this representation as feature input into a simple logistic regression classifier models.
For example, the first results column in Table 2 shows the results when we use the percentages of word-weighted positive versus negative sentiment for each epoch for NFP trend prediction. The third column of Table 2 presents results using all the dimensions of sentiment available but using the *differences* in the counts between epochs as features. The idea behind this approach is intuitive and assumes the trends of sentiment implicit in the text should correlate with the trends of the economy. A raised level of negativity in the news media compared to normal would reflect a period of economic difficulty and visa versa for positive sentiment in the news. We can see this approach achieves a good measure of correlation between the text sentiment and the trends of the NFP.
These results over text show clearly there is predictive information within economic news that we can access via selecting intuitive features from the sentiment analysis of the text. Using these sentiment features in a state-of-the-art machine learning framework gives good prediction results for the NFP that significantly beat the baselines.

5.4 Time Series Prediction

Using the same methodology as above we build a suite of independent classifiers based on the time series data we collected and described in Section 4. We collected over thirty different economic indexes but here we report only on the four series with the best independent prediction results: the Consumer Price Index (CPI), the Institute for Suppy Management Manufacturing Index (ISM), the JOLTS Nonfarm Index (JOLTS), and the Labor Force Levels (LFL). Each of these is directly or indirectly related to the unemployment rate and hence the NFP. As with the text streams, for each time series we trained a logistic regression classifer using multivariate features from the data. The features consisted of the point value plus a number of indicators of the trend and moving averages for various time frames.

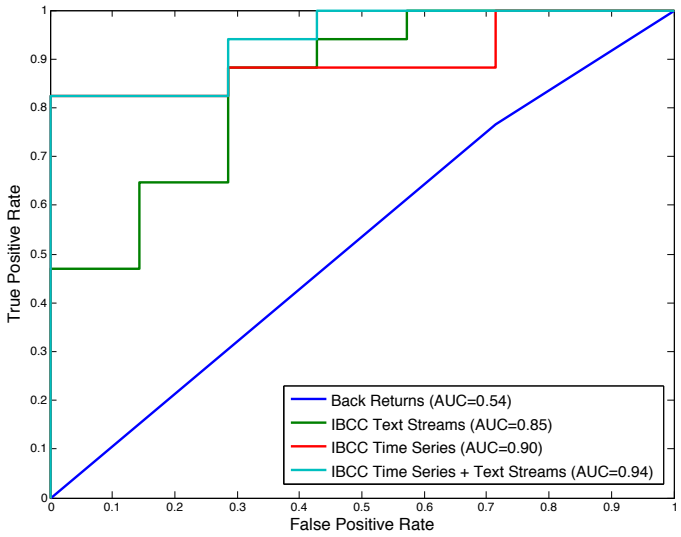


Fig. 5. The AUC results for the NFP predictions. Using a combination of text and time series data results in the best prediction accuracy.

As can be seen from Table 3 each individual time series gives significantly better results than the baselines and improves upon the text sentiment results. When we combine each of these weak classifiers using the IBCC model we get an improved overall AUC of 0.90.

5.5 Heterogeneous Data Prediction

Finally we tested combining the different data types – time series and text stream data – into a single prediction. This is straight forward since each data source is treated as an independent base classifier so IBCC cannot distinguish between the data types. As Table 4 shows, using both the data types together provides significantly improved prediction accuracy indicating that the sentiment within the text streams contains information that is complementary to the real-valued time series.

Source	AUC
Time Series + Text Averages	**0.94**
Time Series + Text Trends	0.91

Table 4. The AUC results when we combine hetereogenous data types with the IBCC model.

Figure 5 depicts the AUC results between the baselines and the IBCC results. Clearly we are learning something of interest using our streaming framework and associated combination model. As well we see the text data is providing us with a source of knowledge that is not present in the time series and, when used in a classifier combination setting, provides extra useful information that improves prediction. [11]

[11] To the authors' knowledge there is no prior published benchmark for NFP prediction against which to make a direct comparison. Our primary comparison here is against the stream-specific weak classifiers.

6 Conclusion

Using news streams and other text sources to make economic predictions is an area that has generated significant interest in the last decade. Our results show clearly there is predictive information within economic news that we can access via selecting intuitive features from the sentiment analysis of the text. Using these sentiment features in a state-of-the-art machine learning framework gives good prediction results of economic trends and variables of interest such as the NFP. However, our results show that combining these text streams with more standard time series data within a classifier combination framework such as IBCC produces highly accurate predictions. Clearly there is information within the text that is complementary to the information contained in the time series data. Using IBCC allows easy integration of multiple classifiers of arbitrary data types from a variety of sources and allows us to model the complementary information to obtain better results.

The scope of this type of economic prediction has many potential applications in both further academic research to more direct financial and market orientated ones with a host of directions for future work. For example, extending the classifier combination model to produce real-valued predictions instead of just predicting trend categories is research which we are current conducting. And while there is large scope for future work on using sentiment of big WWW text data for economic predictions, we believe the research we have reported in this paper is a step forward in the current literature in this area.

References

1. Bollen, J., Mao, H., Zeng, X.J.: Twitter mood predicts the stock market. J. Comput. Science 2(1), 1–8 (2011)
2. Chahuneau, V., Gimpel, K., Routledge, B.R., Scherlis, L., Smith, N.A.: Word salad: Relating food prices and descriptions. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 1357–1367. Association for Computational Linguistics, Jeju Island, Korea (July 2012), http://www.aclweb.org/anthology/D12-1124
3. Chatfield, C.: The Analysis of Time Series: An Introduction. CRC Press (1975)
4. Choi, H., Varian, H.R.: Predicting the present with google trends. The Economic Record 88(s1), 2–9 (2012), http://EconPapers.repec.org/RePEc:bla:ecorec:v:88:y:2012:i:s1:p:2-9
5. Fan, D.P.: Predicting the index of consumer sentiment when it isnt measured. In: JSM Proceedings, AAPOR. p. 60986110. American Statistical Association, Alexandria, VA (2010)
6. Furche, T., Gottlob, G., Grasso, G., Schallhart, C., Sellers, A.: Oxpath: A language for scalable data extraction, automation, and crawling on the deep web. The VLDB Journal pp. 1–26 (2012), http://dx.doi.org/10.1007/s00778-012-0286-6
7. Ghahramani, Z., Kim, H.C.: Bayesian classifier combination. Gatsby Computational Neuroscience Unit Technical Report GCNU-T., London, UK: (2003)
8. Hamilton, J.D.: Time Series Analysis. Princeton University Press (1994)
9. Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., Allan, J.: Mining of concurrent text and time series. In: In proceedings of the 6 th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining Workshop on Text Mining. pp. 37–44 (2001)
10. Mao, H., Counts, S., Bollen, J.: Predicting financial markets: Comparing survey, news, twitter and search engine data. CoRR abs/1112.1051 (2011)
11. Mills, T.C., Markellos, R.N.: The Econometric Modelling of Financial Time Series. Cambridge University Press (2008)
12. Mittermayer, M.A., Knolmayer, G.: Text mining systems for market response to news: A survey. Proceedings of the IADIS European Conference Data Mining (2007)
13. Moilanen, K., Pulman, S.: Sentiment composition. In: Proceedings of Recent Advances in Natural Language Processing (RANLP 2007). pp. 378–382 (September 27-29 2007), http://users.ox.ac.uk/~wolf2244/sentCompRANLP07Final.pdf
14. Nikfarjam, A., Emadzadeh, E., Muthaiyah, S.: Text mining approaches for stock market prediction. In: Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on. vol. 4, pp. 256–260 (2010)
15. Preis, T., Moat, H.S., Stanley, H.E.: Quantifying trading behavior in financial markets using google trends. Scientific Reports 3(1684) (April 2013)

16. Preis, T., Reith, D., Stanley, H.E.: Complex dynamics of our economic life on different scales : insights from search engine query data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences Vol.368(No.1933), 5707–5719 (2010), http://wrap.warwick.ac.uk/50936/
17. Savor, P., Wilson, M.: How much do investors care about macroeconomic risk? evidence from scheduled economic announcements. Journal of Financial and Quantitative Analysis FirstView, 1–62 (3 2013)
18. Schumaker, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: The azfin text system. ACM Trans. Inf. Syst. 27(2), 12:1–12:19 (2009)
19. Simpson, E., Roberts, S., Psorakis, I., A., S.: Dynamic bayesian combination of multiple imperfect classifiers. Decision Making and Imperfection. Intelligent Systems Reference Library 474 (2013)
20. Smith, N.A.: Text-driven forecasting. http://www.cs.cmu.edu/~nasmith/papers/smith.whitepaper10.pdf (2010), http://www.cs.cmu.edu/~nasmith/papers/smith.whitepaper10.pdf
21. Spackman, K.A.: Signal detection theory: valuable tools for evaluating inductive learning. In: Proceedings of the sixth international workshop on Machine learning. pp. 160–163. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1989)

Belief CSP : a New CSP Framework under Uncertainty

Theoretical Foundations

Aouatef Rouahi[1], Kais Ben Salah[2], and Khaled Ghédira[1]

[1] Higher Institute of Management of Tunis, SOIE Laboratory, Tunis, Tunisia
rouahi.aouatef@hotmail.fr, khaled.ghedira@isg.rnu.tn
[2] Higher Institute of Management of Sousse, SOIE Laboratory, Sousse, Tunisia
kaisuria@yahoo.fr

Abstract. The Constraint Satisfaction Problem (CSP) is acknowledged as a simple declarative formalism for modeling and solving well-defined decision problems. Still, the standard CSP has proven unsuited for reasoning under imperfection where flexibility is a key notion. With the aim of building a reliable model and by merging the belief function theory and the CSP formalism, in this paper, we introduce a new CSP extension, labeled Belief Constraint Satisfaction Problem (BCSP), fulfilling two tasks that are managing imperfection and modeling flexibility. The first task conserves the real core of the problem being tackled; the second one allows maintaining that core as the problem's environment is becoming more or less imperfect.

Keywords: CSP, Belief function theory, Belief CSP, Imperfection, Flexibility

1 Introduction

The CSP framework has carried high attention within the AI community because of its simplicity and generality. However, decision problems tackled by the classical CSP are assumed to be well-defined so that all their items are precise and known with certainty. Hence, the CSP has proven unfit for reasoning under imperfection[3] where any item may be uncertain and/or imprecise. The imperfect items may be deleted, ignored or reformed. Anyhow, we may find ourselves taking on the wrong problem and, thus, yielding irrelevant decisions. Therefore, in order to get a reliable model of the problem being tackled, we have to model that imperfection. Various CSP extensions have been made to deal with imperfection by exploiting uncertainty theories[3, 9, 10]. These proposals handle only one imperfection aspect, i.e., uncertainty or imprecision, and differently of what

[3] Uncertainty has been, mostly, used in the literature to refer to imperfection, whereas, imperfection implies both uncertainty which is an epistemic property of the relation between the information and the agent knowledge about the world, and imprecision that touches the information itself[14].

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

we propose these latter, in the best cases, use two formalisms to express both imperfection and flexibility as do the Fuzzy CSP[2] which combines fuzzy sets and possibility theories under a commensurability assumption. However, there is no attempt to take advantage of the belief function theory[1, 11, 13] which offers a sound mathematical basis that manages simultaneously both imperfection aspects and enables both imperfection and flexibility to be expressed with the same formalism. In addition, the belief function theory allows an explicit management of both partial and total ignorance using only the available knowledge and no more. This paper is devoted to introduce the theoretical foundations of a new CSP extension labeled BCSP reaping benefits from both CSP and belief function theory and fulfilling two tasks that are managing imperfect (uncertain and/or imprecise) relations and modeling soft or flexible constraints.
The paper is organized as follows: in the section 2 we, succinctly, recall the formal definition of the CSP besides some related notions. Then, we overview the basics of the belief function theory as a tool for modeling imperfection and flexibility within the CSP. In the section 3, we introduce the BCSP, where, the main concepts are illustrated by a simple example.
In this paper, we want to focus on the theoretical foundations of the BCSP, where comparison with other CSP extensions and experimental results will be introduced in a forthcoming paper.

2 Background Concepts

2.1 Constraint Satisfaction Problem (CSP)

A classical CSP is defined by a quadruplet (X, D, C, R) where $X = \{x_1, ..., x_n\}$ is a finite set of n variables, each x_i takes its values in a finite domain D_i such that $D = \{D_1, ..., D_n\}$. The simultaneous assignment of values to a set of variables is called an instantiation and denoted by θ. $C = \{C_1, ..., C_m\}$ is a finite set of m constraints where each constraint C_i is defined on a subset of variables $S_i \subseteq X$ delimited its scope and by a relation R_i that specifies the set of compatible instantiations with C_i; R_i is a subset of the cartesian product of the domains of the variables in S_i (i.e., $R_i \subseteq \times\{D_i \mid x_i \subseteq S_i\}$)[4]. A constraint C_i is said to be satisfied by an instantiation θ defined on a set of variables V iff $S_i \subseteq V$ and $\exists\theta_i \in R_i$ such that $\theta_i \subseteq \theta$. The main task is to find a solution, that is an instantiation of all variables so that all constraints are satisfied. We denote the set of all solutions of a given CSP P by $Sols(P)$. A CSP is said to be consistent iff it has at least one solution, otherwise, it is said to be inconsistent.
It is important to note that, besides handling imperfection, our proposal is based on the flexibility notion in CSPs, more precisely, in the constraints. Hence, we should distinguish between hard and soft constraints. Classically, a constraint is a yes-or-no matter where it enumerates the certainly compatible instantiations. Thus, hard constraint is considered as imperative so that every solution should satisfy, whereas, soft or flexible constraint can be satisfied to some degree.

2.2 Belief Function Theory

The belief function theory was first initiated by Dempster[1] and then extended by Shafer[11]. Several interpretations have been introduced[18, 7]. The Transferable Belief Model(TBM) is a non-probabilistic interpretation established by Smets[15] that allows a clear separation between knowledge modeling and decision making since the reasoning process is illustrated through two levels: the credal level where the agent knowledge is represented in the static part and manipulated in the dynamic part; and the pignistic level where the decision part is considered aside. This two-level reasoning enables us to, explicitly, handle imperfection and flexibility within the CSP, all with the same formalism.
Likewise, it is important to distinguish the univariate formalism that handles one variable and the multivariate one[8] where the problem is modeled via multiple variables which, analogically, coincides with the CSP struccture.

Static part. Let $X = \{x_1, ..., x_n\}$ be a finite set of variables, where, each variable x_i is associated to a set of mutually exclusive but not necessarily exhaustive realizations, called frame of discernment[4], denoted by Θ_i. Given a non-empty subset S of X, its frame of discernment, denoted by Θ_S is obtained by the cartesian product of the frames of discernment of the involved variables (i.e., $\Theta_S = \times\{\Theta_i \in S\}$). The elements of Θ_S are called configurations of S on which we induce a valuation function called basic belief assignment (bba) that represents some knowledge (complete, partial or ignorant) about the variables in S. the bba m divides belief over singletons and subsets, i.e., the power set, of Θ_S as follows:

$$m : 2^{\Theta_S} \longrightarrow [0, 1] \quad \text{such that} \sum_{C \subseteq \Theta} m(C) = 1 \tag{1}$$

The value $m(C)$ called basic belief mass(bbm) represents the part of belief, exactly, committed to C[11]. Every configuration subset whose bbm is strictly positive is called focal element.

Dynamic Part. Two basic operations may be performed on bba function, that are, the vacuous extension which corresponds to knowledge refinement and the marginalization which coincides with knowledge coarsening[11].
The vacuous extension of a bba m defined on S, to a larger set $S^{'}$, such that $S \subseteq S^{'}$ induces a bba $m^{S\uparrow S^{'}}$ defined on $S^{'}$ as follows:

$$m^{S\uparrow S^{'}}(B) = \begin{cases} m(A) \text{ if } B = A^{\uparrow S^{'}} \\ 0 \text{ otherwise} \end{cases} \text{ for all } A \subseteq \Theta_S. \tag{2}$$

such that $A^{\uparrow S^{'}} = A \times \Theta_{S^{'}-S}$.
The marginalization is a projection of a bba $m^{'}$ defined on $S^{'}$ into a bba $m^{S^{'}\downarrow S}$ defined on a coarser set S, i.e.,$S \subseteq S^{'}$.

$$m^{S^{'}\downarrow S}(B) = \sum_{A \subseteq \Theta_{S^{'}} : A^{\downarrow S} = B} m^{'}(A) \text{ for all B } \subseteq \Theta_S \tag{3}$$

[4] The frame of discernment Θ_i corresponds, by analogy, to the domain D_i of the variable x_i in a CSP.

such that $A^{\downarrow S}$ is obtained by removing all realizations of each configuration of A which corresponds to variables in $(S' - S)$.

Decision making. As states the decision theory[6], as well as the Dutch book arguments (DBA)[5], a rationally consistent decision must be casted using probabilities. Thus, the credal bbm is reformed to a probability measure known as the pignistic probability using the pignistic transformation[13, 16]. Let m be a bbm defined on Θ, the produced pignistic probability, denoted by $BetP$, is defined as follows:

$$BetP(A) = \sum_{B \subseteq \Theta} \left(m(B) \frac{|A \cap B}{|B|(1 - m(\emptyset))} \right), \text{for all } A \in \Theta. \quad (4)$$

where $|A|$ denotes the number of elements of Θ in A.

3 Belief Constraint Satisfaction Problem (BCSP)

3.1 Definition

A Belief Constraint Satisfaction Problem (BCSP) is a quadruplet (X, D, C, R), where, $X = \{x_1, ..., x_n\}$ is a finite set of variables; $D = \{D_1, ..., D_n\}$ is the set of their domains, such that, D_i is the domain associated to the variable x_i; $C = \{C_1, ..., C_m\}$ is a finite set of belief constraints, each constraint C_i is defined by the pair (S_i, R_i), such that, $S_i \subseteq X$ represents its scope and R_i is its imperfect relation, that is an uncertain and/or imprecise relation.
As we have adopted the TBM, in the credal level imperfect relations are expressed as the static part; some useful operations on imperfect relations are represented in the dynamic part; then, in the pignistic level the belief constraints are induced and solutions are computed.

3.2 Imperfect relation

A belief constraint C_i is defined by the pair (S_i, R_i), where, S_i is a subset of the problem variables delimiting the belief constraint scope, i.e., $S_i = \{X_{i_1}, ..., X_{i_k}\} \subseteq X$, and R_i which is, in turn, defined by the pair (V_i, Θ_i) represents an imperfect (uncertain and/or imprecise) relation that associates a valuation V_i, the bba function, over the frame of discernment of the belief constraint Θ_i obtained by the cartesian product of the involved variables domains, i.e., $\Theta_i = D_{i_1} \times ... \times D_{i_k}$. The valuation V_i is defined as follows:

$$V_i = m_i : 2^{\Theta_i} \rightarrow [0, 1] \mid \sum_{I \subseteq \Theta_i} m_i(I) = 1 \quad (5)$$

where I is a singleton or a subset of instantiations.
If we define an instantiation as a logical relation between variables, the finite amount of support enclosed in the bba (i.e., $m_i(I)$) and derived from the available pieces of evidence can be interpreted as the potentiality degree of a given relation to, actually, occur so that the potentiality of the belief constraint to

be satisfied by such a variables instantiation(s), I. A second reading interprets this bbm distribution as preference levels induced over instantiations. Formally, let θ_1 and θ_2 two subsets of Θ_i (i.e., $\theta_1, \theta_2 \subseteq \Theta_i$), $m(\theta_1) > m(\theta_2)$ means that θ_1 is more believable (certain) than θ_2 and hence θ_1 is, a priori, preferred to θ_2 for the satisfaction of the belief constraint. This latter reading shows that the bba function has twofold role. The first is to quantify our belief about a given instantiation whereas the second is to induce a preorder among them.
According to the "closed world assumption" established by Shafer[11], an imperfect relation R_i is said to be normalized iff $m_i(\emptyset) = 0$. Otherwise, it is said to be unnormalized. This assumption is later relaxed by Smets[15] as the "open world assumption" where the empty set bbm quantifies the conflict amount between the beliefs on Θ_i.

An example. In order to illustrate the different notions related to the BCSP, let us tackle the problem "Pacifist or not" adapted from[12]. We would like to determine whether a person is a pacifist. The available knowledge indicates that most Republicans are not pacifists and we are 90 percent certain about this. Moreover, we are 99 percent sure that most Quakers are pacifists. The corresponding BCSP of this problem would be the quadruple (X, D, C, R) where:

- $X = \{Pa, Re, Qu\}$ the set of variables, such that, we use Pa to denote Pacifist, Re for Republican and Qu for Quaker;
- $D = \{D_{Pa}, D_{Re}, D_{Qu}\}$, such that, $D_{Pa} = \{p(Pacifist), \bar{p}(notPacifist)\}$; $D_{Re} = \{r(Republican), \bar{r}(notRepublican)\}$ and $D_{Qu} = \{q(Quaker), \bar{q}(notQuaker)\}$ which define, respectively, the domains of the variables Pa, Qu and Re ;
- $C = \{C_1, C_2\}$ the set of belief constraints;
- $R = \{R_1, R_2\}$, the set of imperfect relations, such that, $S_1 = \{Re, Pa\}$ and $S_2 = \{Qu, Pa\}$; $\Theta_1 = \{(r, p), (r, \bar{p}), (\bar{r}, p), (\bar{r}, \bar{p})\}$ and $\Theta_2 = (q, p), (q, \bar{p}), (\bar{q}, p), (\bar{q}, \bar{p})$ R_1(See Table 1), R_2(See Table 2)

Table 1. The bba function for R_1

2^{Θ_1}	m_1
$\{(r, \bar{p}), (\bar{r}, p), (\bar{r}, \bar{p})\}$	0.9
$\{(r, p), (r, \bar{p}), (\bar{r}, p), (\bar{r}, \bar{p})\}$	0.1

As we can, obviously, note both of the relations R_1 and R_2 are normalized.

Special relations. The most appealing feature that makes the belief function theory an efficient tool is its faithfulness in recognition our knowledge as well as our ignorance. That is why both of extreme knowledge cases, namely, the total knowledge and the total ignorance within the notion of belief constraints are easily expressed.

Table 2. The bba function for R_2

2^{Θ_2}	m_2
$\{(q,p),(\bar{q},p),(\bar{q},\bar{p})\}$	0.99
$\{(q,p),(\bar{q},p),(q,\bar{p}),(\bar{q},\bar{p})\}$	0.01

– **Complete certainty (perfect relation : certain and precise)** When the sought for instantiation is perfectly known and unique with reference to a given belief constraint, the associated relation is represented using the certain belief function where the one and only focal element is that instantiation. Consider a belief constraint C_i, which is described by the certain and precise relation R_i, the associated bba m_i is defined as follows:

$$\exists\theta \subset \Theta_i, |\theta| = 1 \text{ such that } m_i(\theta) = 1 \text{ and } m_i(\varphi) = 0, \forall\varphi \subseteq \Theta_i, \varphi \neq \theta \quad (6)$$

Obviously, such a case of total knowledge fits the classical notion of perfect relation where all tuples are known with certainty. Hence, our belief model permits, as well, the formalizing of the standard CSP. For instance,

$$\begin{cases} m\{(\bar{r},p)\} = 1 \\ m\{\varphi\} = 0, \forall\varphi \subseteq \Theta_1, \varphi \neq (\bar{r},p) \end{cases}$$

is a certain and precise relation that may describe the belief constraint C_1 of our example.

– **Complete ignorance (imperfect relation: uncertain and imprecise)** When we have no information on what the instantiation(s) satisfying a given belief constraint could be, in other words, we are not able to establish any valuation on the variables tuples, we have recourse to the vacuous belief function. If C_i is a belief constraint, R_i is its associated relation which is described by the vacuous bba m_i that is defined as follows:

$$m_i(\Theta_i) = 1 \text{ and } m_i(\theta) = 0, \forall\theta \neq \Theta_i \quad (7)$$

For example,

$$\begin{cases} m\{(r,p),(r,\bar{p}),(\bar{r},p),(\bar{r},\bar{p})\} = 1 \\ m(\theta) = 0, \forall\theta \neq \Theta_1 \end{cases}$$

is a vacuous relation that may describe the belief constraint C_1 of our example. We are indifferent toward all the tuples, so that, all are believable.

– **Partial ignorance (imperfect relation : certain and imprecise)** The intermediate case between those two extreme cases is the partial ignorance. Such a case is described using the categorical belief function. If C_i is a belief constraint, R_i is its associated relation which is described by the categorical bba m_i that is defined as follows:

$$\exists\theta \subset \Theta_i, |\theta| > 1 \text{ such that } m_i(\theta) = 1 \text{ and } m_i(\varphi) = 0, \forall\varphi \subseteq \Theta_i, \varphi \neq \theta \quad (8)$$

The following bba may be a categorical relation that describes the belief constraint C_1.

$$\begin{cases} m\{(r,\bar{p}),(\bar{r},p),(\bar{r},\bar{p})\} = 1 \\ m(\varphi) = 0, \forall\varphi \neq (r,\bar{p}),(\bar{r},p),(\bar{r},\bar{p}) \end{cases}$$

Another case which may be described by the belief functions is the case when we have some precise but uncertain information. As far as we know, there is no special belief function that expresses this case but it is still always feasible. With such knowledge, more than one focal element is believable. If we retake the belief constraint C_1, a precise and uncertain relation may be as follows:

$$\begin{cases} m\{(r, \bar{p})\} = 0.8 \\ m(r, p) = 0.2 \end{cases}$$

Operations on imperfect relations.

– **Vacuous extension** The vacuous extension of an imperfect relation R_i defined on S_i, to a larger set $S_i^{'}$, such that $S_i \subseteq S_i^{'}$, is an imperfect relation $R_i^{(\uparrow S_i^{'})}$ defined on $S_i^{'}$ and obtained as follows:

$$m_i^{(\uparrow S_i^{'})}(\varphi) = \begin{cases} m_i(\theta) \text{ if } \varphi = \theta^{\uparrow S_i^{'}} \\ 0 \text{ otherwise} \end{cases} \text{ for all } \theta \subseteq \Theta_i \tag{9}$$

Such that $\theta^{\uparrow S_i^{'}}$ denotes the cylindrical extension of the set θ to $S_i^{'}$. The vacuous extension is useful when we want to know to what extent a given instantiation may satisfy the belief constraint C_i. In fact, it corresponds to a refinement of knowledge. For a further explanation, let us take the imperfect relation R_1 describing the belief constraint C_1 defined on $S_1 = \{Re, Pa\}$ (see Table 1), its extension to $S_1^{'} = \{Re, Qu, Pa\}$ gives the imperfect relation $R_1^{(\uparrow S_1^{'})}$ as follows (see Table 3):

Table 3. The extended bba function for $R_1^{(\uparrow S_1^{'})}$

$2^{\Theta_1^{'}}$	$m_1^{(\uparrow S_1^{'})}$
$\{(r, \bar{p}, q), (\bar{r}, p, q), (\bar{r}, \bar{p}, q), (r, \bar{p}, \bar{q}), (\bar{r}, p, \bar{q}), (\bar{r}, \bar{p}, \bar{q})\}$	0.9
$\{(r, p, q), (r, \bar{p}, q), (\bar{r}, p, q), (\bar{r}, \bar{p}, q), (r, p, \bar{q}), (r, \bar{p}, \bar{q}), (\bar{r}, p, \bar{q}), (\bar{r}, \bar{p}, \bar{q})\}$	0.1

– **Marginalization** The knowledge, initially, encapsulated in the bba distribution can be refined as well as coarsened. The marginalization, which corresponds to a coarsening of knowledge, of an imperfect relation R_i defined on S_i, to a coarser set $S_i^{'}$,i.e.,$S_i \supseteq S_i^{'}$, is an imperfect relation $R_i^{(\downarrow S_i^{'})}$ defined on $S_i^{'}$ and obtained as follows:

$$m_i^{(\downarrow S_i^{'})}(\varphi) = \sum_{\theta \subseteq \Theta_i : \theta^{\downarrow S_i^{'}} = \varphi} m_i(\theta) \text{ for all } \varphi \subseteq \Theta_i \tag{10}$$

Such that $\theta^{\downarrow S_i^{'}}$ denotes the projection of the set θ to $S_i^{'}$. We can employ the marginalization when we want to know to what extent a given partial instantiation, if extended, may satisfy the belief constraint C_i. For a closer examination, let us take the imperfect relation R_2 describing the belief constraint C_2 defined on $S_2 = \{Qu, Pa\}$ (see Table 4), its marginalization to $S_2^{'} = \{Qu\}$ gives the relation $R_2^{(\downarrow S_2^{'})}$ as follows (Table 4):

Table 4. The marginalized bba function for $R_2^{(\downarrow S_2^{'})}$

$2^{\Theta_2^{'}}$	$m_2^{(\downarrow S_2^{'})}$
$\{(q, \bar{q})\}$	$0.9 + 0.1 = 1$

3.3 Belief constraint

After expressing our beliefs on the imperfect relations in the credal level, we have to extract the satisfaction degree of the belief constraints by each instantiation θ in Θ_i aside using the pignistic probabilities produced by the TBM pignistic transformation of the mass distribution.
Let C_i be a belief constraint, R_i its relation and let m_i be the associated mass distribution over Θ_i, the produced pignistic probability, denoted by $BetP_i$, is defined as follows:

$$BetP_i(\theta) = \sum_{\varphi \subseteq \Theta_i} \left(m_i(\varphi) \frac{|\varphi \cap \theta|}{|\varphi|(1 - m_i(\emptyset))} \right), \text{for all } \theta \in \Theta_i \tag{11}$$

Retaking our example "Pacifist or not", the computations are shown in the tables 5 and 6 for, respectively, C_1 and C_2.

Table 5. The satisfaction degrees of C_1

Θ_1	$BetP_1$
$\{(r, p)\}$	0.025
$\{(r, \bar{p})\}$	0.325
$\{(\bar{r}, p)\}$	0.325
$\{(\bar{r}, \bar{p})\}$	0.325

Hence, it can be easily seen that this notion of pignistic probability allows for expressing soft or flexible constraints starting from imperfect relations. It is of

Table 6. The satisfaction degrees of C_2

Θ_2	$BetP_2$
$\{(q,p)\}$	0.3325
$\{(q,\bar{p})\}$	0.0025
$\{(\bar{q},p)\}$	0.3325
$\{(\bar{q},\bar{p})\}$	0.3225

interest to discern the difference between hard constraint that should be certainly and fully satisfied and soft constraint whose satisfaction is not required to be neither certain nor total. Therefore, the satisfaction of a given constraint becomes, essentially, a matter of degree, such that:

$-BetP_i(\theta) = 1$ means that the instantiation θ totally satisfies the constraint C_i;
$-BetP_i(\theta) = 0$ means that the instantiation θ totally violates the constraint C_i;
$-0 < BetP_i(\theta) < 1$ means that the instantiation θ partially satisfies the constraint C_i;

The BetP also induces a preorder among instantiations. Formally, let θ and $\theta^{'}$ be two instantiations defined on the same set of variables, $BetP(\theta) > BetP(\theta^{'})$ means that θ is, a posteriori, preferred to $\theta^{'}$ for the satisfaction of the flexible belief constraint. Obviously, hard constraints are a particular case of belief constraints which are satisfied only to 1 or 0 degree. Hence the $BetP$ have also twofold role as it allows first expressing flexible constraints and second preferences among instantiations.

3.4 BCSP consistency

Belief constraint satisfiability. A belief constraint C_i, whose scope is S_i, is said to be (totally or partially) satisfied by a given instantiation $\theta \in \Theta_i$, noted $\theta \models C_i$ iff $BetP_i(\theta) > 0$. A belief constraint C_i is said to be unsatisfiable if there is no instantiation that satisfies it, i.e., $\forall \theta \in \Theta_i, BetP_i(\theta) = 0$.

Instantiation consistency. Classically, an instantiation θ of a set of variables $S \subseteq X$ is said to be consistent iff it satisfies all the constraints among that set. With the BCSP view, the constraint satisfiability is not any more a yes/no query but a matter of degree, where $BetP_i(\theta)$ indicates to what extent the instantiation θ satisfies the belief constraint C_i, and so the instantiation consistency is. Hence, a given instantiation θ is said to be consistent iff it satisfies all constraints among S to a non-nil degree. Formally, the consistency degree of an instantiation θ of

a set of variables $S \subseteq X$ is obtained as follows[5]

$$C(\theta) = BetP_{\wedge}\{R_i|S_i \subseteq S\}(\theta) = \prod_{R_i^{\uparrow S}|S_i \subseteq S} \{BetP_i\}(\theta) = \prod_{R_i|S_i \subseteq S} \{BetP_i\}(\theta^{\downarrow S}) \tag{12}$$

− If θ totally satisfies all the constraints covering S, it said to be completely consistent, i.e.,$C(\theta) = 1$.
− If θ totally violates, at least, one constraint is said to be inconsistent, i.e., $C(\theta) = 0$.
− Otherwise, it is said to be partially consistent, i.e., $0 < C(\theta) < 1$.
As the BCSP is a generalization of the classical model, if the relations are perfect (total knowledge case), a given instantiation is either consistent (1) or inconsistent (0).

BCSP solution. A solution of a BCSP P (X, D, C, R) is every consistent complete instantiation θ, i.e., an instantiation of all the variables in X whose consistency degree is greater than 0, so that, all the constraints in C are satisfied. This consistency degree, evidently, corresponds to the satisfaction degree of the BCSP P by that instantiation.

$$S_P(\theta) = C(\theta) = \prod_{C_i \in C; R_i^{\uparrow X}} \{BetP_i\}(\theta) \tag{13}$$

Accordingly, we can merely notice that the satisfaction degree of the belief CSP, as defined above, accomplishes a sort of quantitative discrimination among the several instantiations inducing then a total preorder over them. Then, the higher is the satisfaction degree, the better is the instantiation. The solution space of a BCSP P (X, D, C, R) consists of the set of all the feasible solutions, i.e.,

$$Sols(P) = \{\theta \in D_1 \times ... \times D_n | S_P(\theta) > 0\} \tag{14}$$

BCSP consistency. A BCSP P (X, D, C, R) is said to be:
− Totally consistent if and only if it has at least one solution that totally satisfies all the constraints of C, i.e., $\exists \theta \in Sols(P)|S_P(\theta) = 1$.
− Totally inconsistent if and only if all instantiations of X are inconsistent, i.e., $Sols(P) = \emptyset$ or also $\forall \theta \in D_1 \times ... \times D_n | S_P(\theta) = 0$.
− Partially consistent if and only if all solutions are somehow feasible, i.e., $Sols(P) \neq \emptyset | \forall \theta \in Sols(P), S_P(\theta) < 1$.
Toward the same view, the consistency degree of a BCSP is the satisfaction degree of its best (optimal) solution, i.e.,

[5] The Fuzzy[2] and the Possibilistic[10] CSPs suffer from the "drowning effect" because of the egalitarian min-max operators use which barely discriminates between assignments that satisfy the CSP to the same degree. To avoid falling into the same weakness, we propose using an utilitarian operator, the product, to aggregate the preference degrees.

$$C(P) = S_P(\theta^*) = \max_{\theta \in Sols(P)} (S_P(\theta)) = \max_{\theta \in Sols(P)} \left(\prod_{C_i \in C; R_i^{\uparrow X}} \{BetP_i\}(\theta) \right) \quad (15)$$

As in the classical CSP, many tasks may be performed within the BCSP framework. The first and most intuitive task to perform may be to ascertain whether or not there is a solution, in other words find out if the problem is consistent or not. Given a satisfiable BCSP, some other tasks may be required such as: Determine the consistency degree of the problem, $C(P)$.
− Find any solution, i.e., some $\theta \in Sols(P) | S_P(\theta) > 0$.
− Find all of the solutions, $Sols(P)$.
− Find all solutions having a degree greater than a given bound B, i.e., every $\theta \in Sols(P) | S_P(\theta) > B$.
− Find an optimal solution, i.e., some $\theta \in Sols(P) | S_P(\theta) = C(P)$, and so forth.
Let us get back to our example "Pacifist or not". The local consistency of the instantiation $(Pa = \bar{p}, Re = r)$ is partial with respect to the constraint C_1, i.e., $C(Pa = \bar{p}, Re = r) = BetP_1(Pa = \bar{p}, Re = r) = 0.325$. The consistency of the complete instantiation $(Pa = p, Re = r, Qu = q)$ that indicates the satisfaction degree of P giving this is $C(Pa = p, Re = r, Qu = q) = S_P(Pa = p, Re = r, Qu = q) = \prod_{R_i^{\uparrow X} | S_i \subseteq X} \{BetP_i\}(Pa = p, Re = r, Qu = q) = 0.025 * 0.3325 = 0.008$. An optimal solution may be $(Pa = p, Re = \bar{r}, Qu = q)$, where, $C(Pa = p, Re = \bar{r}, Qu = q) = S_P(Pa = p, Re = \bar{r}, Qu = q) = 0.325 * 0.3325 = 0.108$. There are other possible solutions but less consistent such as $C(Pa = p, Re = r, Qu = \bar{q}) = 0.025 * 0.3325 = 0.008$ and $C(Pa = \bar{p}, Re = r, Qu = q) = 0.325 * 0.0025 = 0.0008$. The consistency of the problem is the consistency of its best solution, i.e., $C(P) = 0.108$. The problem is partially consistent.

4 Conclusion

In this paper, we have introduced the Belief Constraint Satisfaction Problem as a new CSP extension merging the CSP formalism and the belief function theory as interpreted by the TBM in order to deal with imperfect (uncertain and/or imprecise) relations and flexible constraints with the same formalism. This paper is, basically, devoted to represent the theoretical basis of the BCSP. By making the BCSP under experiments, we have proved some important results concerning the behavior of the BCSPs under uncertainty variation, besides, some interesting facts about links between constraints tightness [6] , constraints flexibility or softness and uncertainty. Furthermore, a comparison with the Transition Phase phenomenon[17] has been made. These evaluation results, besides, the comparison with other CSP extensions will be reported in a forthcoming paper. Further research targets exploiting the other informative measures offered by the belief functions theory in order to enlarge the scope of problems that can be tackled using the BCSP formalism such as the multi-objective optimization problems.

[6] The constraint tightness is the probability of an inconsistency between two values related by that constraint.

References

[1] Dempster, A. P.: Upper and lower probabilities induced by a multivalued mapping. The Annals of Mathematical Statistics, Vol.38, No.2 (1967) 325–339

[2] Dubois, D., Fargier, H., and Fortemps, P.: Fuzzy scheduling: Modeling flexible constraints vs. coping with incomplete knowledge. European Journal of Operational Research, Vol.147, (2003) 231–252

[3] Fargier, H., and Lang, J.: Uncertainty in constraint satisfaction problems: a probabilistic approach. In Proc. of the Second European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty, ECSQARU'93, Vol.747 (1993) 97–104

[4] Ghédira, K.: Constraint Satisaction Problems: CSP Formalisms and Techniques. ISTE Ltd and John Wiley and Sons, Inc (2013)

[5] Hájek, A.: Dutch Book Arguments. In Anand, P., Pattanaik, P., and Puppe, C. (Eds.), The Oxford Handbook of Rational and Social Choice, Oxford: Oxford University Press (2008) 173–195,

[6] Jeffrey, R.: The Logic of Decision. 2nd edition, Chicago: The University of Chicago Press (1983)

[7] Kohlas, J., and Monney, P. A.: A Mathematical Theory of Hints. An Approach to the Dempster-Shafer Theory of Evidence. Lecture Notes in Economics and Mathematical Systems, Vol.425, Springer-Verlag, Berlin (1995)

[8] Kong, A.: Multivariate Belief Functions and Graphical Models. Ph.D. thesis, Department of Statistics, Harvard University (1986)

[9] Ruttkay, Z.: Fuzzy constraint satisfaction. In Proc. of the Third IEEE International Conference on Fuzzy Systems,(1994) 1263–1268

[10] Schiex, T.: Possibilistic constraint satisfaction problems or "How to handle soft constraints?". In Proc. of the 8th International Conference on Uncertainty in Artificial Intelligence, Stanford, CA, USA, (1992) 268–275

[11] Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton, NJ (1976)

[12] Shenoy, P. P.: Using Dempster-Shafer′s belief-function theory in expert systems. Advances in the Dempster-Shafer Theory of Evidence. John Wiley and Sons, New York, Yager, R.R., Fedrizzi, M., and J. Kacprzyk (Eds.) (1994) 395–414

[13] Smets, Ph.: Constructing the pignistic probability function in a context of uncertainty. In Henrion et al. (1990) 29–40

[14] Smets, Ph.: Imperfect information: Imprecision-Uncertainty. Uncertainty Management in Information Systems. From Needs to Solutions. Motro, A. and Smets, Ph. (Eds.), Kluwer Academic Publishers (1997) 225–254

[15] Smets, Ph.: The transferable belief model for quantified belief representation. In. Gabbay, D.M, Smets, Ph. (Eds.), Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol.1, Doordrecht, The Netherlands: Kluwer (1998) 267–301

[16] Smets, Ph.: Decision making in the TBM: the necessity of the pignistic transformation. International Journal of Approximate Reasoning, Vol.38, No.2 (2005) 133–147

[17] Smith, B. M.: Phase transition and the mushy region in constraint satisfaction problems. In Proc. of ECAI-94, Amestrdam, Netherlands (1994) 100–104

[18] Walley, P.: Statistical reasoning with imprecise probabilities. Chapman and Hall, London (1991)

DEMO: What Lies Beneath Players' Non-Rationality in Ultimatum Game?

Galina Avanesyan[1,3], Miroslav Kárný[1], Zuzana Knejflová[1,2], and Tatiana V. Guy[1]

[1] Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic
[2] Czech Technical University, Prague, Czech Republic
[3] University of Economics in Prague, Czech Republic

The rational strategy suggested by the game theory predicts a human playing Ultimatum Game (UG) would have tendency to decide in accordance with the assumption of self-interested rationality, i.e. to choose more for oneself and offer the least amount possible for co-players [2]. This "utilitarian" and game-theoretically correct "rational" behaviour is however rarely observed when experiments are conducted with human beings [1]. Long-term research in experimental economics indicates that humans do not behave as selfish as traditional economics assume them to do. In UG, human-responders reject offers they find too low while human-proposers often offer more than the smallest amount. An intuitively plausible interpretation of this phenomenon is that responders would rather give up some profit than be treated unfairly. This "non-rational" behaviour provides an insight into human's motivation as a *social* being.

The work challenges this view and insists on human rationality. The key hypothesis is that *humans behave rationally*, however, use different criterion than a pure economical profit. The proposed approach models a human-responder via Markov decision process with a reward function respecting both economical profit and fairness. Two types of a reward function are considered: **R1** - a reward respecting fairness towards both players, and **R2** - a reward respecting fairness towards the responder only. The influence of either economical profit or fairness is controlled by a weight. A comparison on the data gained from the games with human-responder shows that the reward **R2** leads to the strategy close to the human one while the reward **R1** often leads to the higher economical profits compare to either human strategy or strategy given by **R2**[4].

References

1. Guth, W., Schmittberger, R., Schwarze, B.: An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization 3(4), 367–388 (1982)
2. Rubinstein, A.: Perfect equilibrium in a bargaining model. Econometrica 50(1), 97–109 (1982)

[4] This research has been supported by GAČR 13-13502S

DEMO: Sparsity in Bayesian Blind Source Separation and Deconvolution

Václav Šmídl and Ondřej Tichý

Institute of Information Theory and Automation,
Academy of Sciences of the Czech Republic

The blind source separation problem deals with recovering the original signals that can be observed only via their superposition. Examples can be given: coctail party problem, astronomical images, medical image sequence, etc. The task can be specified for medical image sequence analysis: the sources are physiological tissues and the superposition is given by nature of a camera such as in scintigraphy [2]. Each pixel of an image arises as the sum of particles coming from an applied radioactive tracer from the scaned body region in a specific time-interval. Hence, each obtained image is a superposition of an unknown number of underlaying tissue images. The aim is to separate the images of biologic tissues and related time-activity curves from the sequence of images.

In the DEMO, we compare the clasical variational approach [3] with posible extension such as: (i) estimation of probabilistic regions of interest, i.e. probabilistic masks, of images [4], (ii) incorporated biologically-motivated assumption of time-activity curves such as information that each time activity-curve arise as a convolution of an input function and tissue-specific kernel [5], (iii) incorporating an assumption that the source signal is sparse using automatic relevance determination [1] and arises as a convolution.

The results suggest that further modeling within blind source separation is necessary for reasonable solution of the problem.

Acknowledgements This research has been supported by GAČR 13-29225S.

References

1. Bishop, C., Tipping, M.: Variational relevance vector machines. In: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence. pp. 46–53 (2000)
2. Buvat, I., Benali, H., Paola, R.: Statistical distribution of factors and factor images in factor analysis of medical image sequences. Physics in medicine and biology 43, 1695 (1998)
3. Miskin, J.: Ensemble learning for independent component analysis. Ph.D. thesis, University of Cambridge (2000)
4. Šmídl, V., Tichý, O.: Automatic Regions of Interest in Factor Analysis for Dynamic Medical Imaging. In: 2012 IEEE International Symposium on Biomedical Imaging (ISBI). IEEE (2012)
5. Šmídl, V., Tichý, O., Šámal, M.: Factor Analysis of Scintigraphic Image Sequences with Integrated Convolution Model of Factor Curves. In: Proceedings of the second international conference on Computational Bioscience. IASTED (2011)

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic