Advances in Knowledge-Based Technologies

Proceedings of the
Master and PhD Seminar
Summer term 2012, part 1

Softwarepark Hagenberg
SCCH, Room 0/2
25 April 2012

Software Competence Center Hagenberg
Softwarepark 21
A-4232 Hagenberg
Tel. +43 7236 3343 800
Fax +43 7236 3343 888
www.scch.at

Fuzzy Logic Laboratorium Linz
Softwarepark 21
A-4232 Hagenberg
Tel. +43 7236 3343 431
Fax +43 7236 3343 434
www.flll.jku.at

Program

Session 1. Chair: Roland Richter

13:00 Martin Reinhardt:

Applications of Image Processing using Monogenic Wavelet Frames

13:30 Kurt Pichler:

Detecting Cracks in Reciprocating Compressor Valves

Session 2. Chair: Bernhard Moser

14:15 Michael Affenzeller:

Fitness Landscape Analysis applied to instances of the Quadratic Assignment Problem Library

14:45 Wolfgang Heidl:

Population Decision Modeling

TECHNISCHE UNIVERSITÄT BERGAKADEMIE FREIBERG

The University of Resources. Since 1765.

Faculty of Mathematics and Computer Science
Institute of Applied Analysis

Seminar paper

Applications of Image Processing using Monogenic Wavelet Frames

Martin Reinhardt
DI Stefan Schausberger
Dr. Bettina Heise

Applied Mathematics
Specialization: Communication Technologies

April 25, 2012

Tutor/First Proofreader:
Dr. Dipl.Phys.in Bettina Heise
CDL-MS-MACH Linz/ FLLL Hagenberg

Contents

In the last years Seminar I introduced the Monogenic Wavelet and the Riesz Transform as a new possibility to new get image properties. With these properties some new applications can be performed as well as several pre-calculations to improve existing applications.
The main areas of using these properties include machine learning, Image Processing and Imaging in FF-OCM (Full Field Optical Coherence Tomography). This work shows the progress of my thesis "Wavelets in Image Processing" and gives an insight of the further work in this area.

1 Recap

The base of the Monogenic Signal is the Riesz Transform. It is defined as:

$$\widehat{\mathcal{R}_\alpha f}(\xi) = i\frac{\xi_\alpha}{\|\xi\|}\widehat{f}(\xi),\ \alpha = 1, \ldots, n \tag{1}$$

With the help of this Transform the analytical Signal based on the Hilbert Transform

$$f_a = f + i\mathcal{H}f$$

can be extended to the Monogenic Signal

$$f_m = (f, \mathcal{R}_1 f, \ldots, \mathcal{R}_n f) \tag{2}$$

This Monogenic Signal provides the extraction of new image properties like Amplitude, Phase and Phase Orientation by the following calculations:

- Amplitude: $A(x) = \sqrt{f^2(x) + (\mathcal{R}_1 f(x))^2 + \ldots + (\mathcal{R}_n f(x))^2}$
- Phase: $P(x) = \left|\operatorname{atan2}(\sqrt{(\mathcal{R}_1 f(x))^2 + \ldots + (\mathcal{R}_n f(x))^2}, f)\right|$
- Phase orientation: $O(x) = \frac{(\mathcal{R}_1 f(x), \ldots, \mathcal{R}_n f(x))}{\sqrt{(\mathcal{R}_1 f(x))^2 + \ldots + (\mathcal{R}_n f(x))^2}}$

The locale phase can be interpreted as the optical flow of an image. Here the local phase orientation denotes the direction of the flow and the local phase angle denotes the strength. The influence of the properties can be easily shown at the example of the so called "Zone Plate" test image in figure 1.

2 Applications of Monogenic Signals

Applications in Image Processing

Next to the fact that the Riesz Transform is a good working edge-detector, the Monogenic Analysis can be used to find new kinds of features for classification of images. First ideas of using the three extracted properties of textures instead the image itself exists, but haven't been implemented or tested yet. Similar ideas exists for the clustering of textures. The main effort of the use of these properties is the reaching of higher robustness in the algorithms of Machine Learning.
To show that this approach is useful, a decomposition into the new image properties were made. Due to deal with the backlight the algorithms "Equalisation of Brightness" respectively "Rolling Ball" were performed. The results of this composition are shown in figure 2.

As a result it can be seen that these similar looking textures can get new features from the phase and the orientation to provide a possibility of cluster/classify them. Several tests in the future will show the usability and the improvement of the clustering/classification.

Applications in Imaging

With the help of the SLM Device the filters can be applied in an optical setup called Full Field Optical Coherence Microscopy. Information about this technology can be provided by Dr. Bettina Heise and DI Stefan Schausberger from the CD Laboratory/Linz.
The first step is to take 2 phase shifted images and to use the difference to eliminate the backlight. The resulting image can be used in the usual way to extract the new image-properties. These are shown in figure 3.

Next to the usability of the provided ways of Image Processing the filters can be also used directly as a phase filter. Because of the fact, that in imaging just intensities can be measured and given as filters, the Riesz Transform cannot be applied in each direction. These special Filters have been implemented and tested but not evaluated and interpreted yet.

3 An Outlook on new Wavelet Frames

3.1 Notations

The Group $GL_n(K)$ or $GL(n,K)$ includes all regular $n \times n$ Matrices with coefficients out of K.
Vectors:

- $x \in \mathbb{R}^n$ is a column vector.
- $\xi \in \widehat{\mathbb{R}}^n$ is a row vector.
- a vector multiplying a matrix on the right is understood to be a column vector
- a vector multiplying a matrix on the left is understood to be a row vector

Operators for $f \in L^2(\mathbb{R}^n)$:

- The Translation-Operator T_y with $y \in \mathbb{R}^n$ is defined as $(T_y f)(x) = f(x-y)$
- The Dilation-Operator with $M \in GL_n(\mathbb{R})$ is defined as $(D_M f)(x) = |\det M|^{-\frac{1}{2}} f(M^{-1}x)$
- The Modulation-Operator with $\nu \in \mathbb{R}^n$ is defined as $(M_\nu f)(x) = \exp^{2\pi i \nu x}$

While $f \in L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$ the Fourier Transform is defined as

$$\hat{f}(\xi) = \int_{\mathbb{R}^n} f(x) \exp^{-2\pi i \xi x} \, \mathrm{d}x \tag{3}$$

With the back transform

$$\check{f}(x) = \int_{\widehat{\mathbb{R}}^n} f(\xi) \exp^{2\pi i \xi x} \, \mathrm{d}\xi \tag{4}$$

The results are the following properties of the operators

- $(T_y f)\hat{}(\xi) = \left(M_y \hat{f}\right)(\xi)$
- $(D_M f)\hat{}(\xi) = \left(\widehat{D}_M \hat{f}\right)(\xi) = |\det M|^{\frac{1}{2}} \hat{f}(\xi M)$

3.2 Affine systems with composite dilations

A general definition of affine systems with composite dilations can be given by ([FM-CFHLG09]):

$$\mathcal{A}_{AB}(\Psi) = \left\{ D_A D_B T_k \psi^l : A \in G_A, B \in G_B, k \in \mathbb{Z}^n, l = 1, \ldots, L \right\} \tag{5}$$

where $\Psi \subset \left\{\psi^1, \ldots, \psi^l\right\} \in L^2(\mathbb{R}^n), G_A \subset GL_n(\mathbb{R})$ and $G_B \subset GL_n(\mathbb{R})$ with $|\det B| = 1$ Here the Elements $A \in G_A$ dilate into at least one direction while the elements of G_B give the geometry of the system $\mathcal{A}_{AB}(\Psi)$.

G_B is a countable subset of $\widetilde{SL}_n(\mathbb{Z}) = \{B \in GL_n(\mathbb{R}) : |\det B| = 1\}$ and $G_A = \{A^i : i \in \mathbb{Z}\}$ where $A \in GL_n(\mathbb{Z})$. Also it is required that A normalises G_B ($ABA^{-1} \in G_B$ for each $B \in G_B$) and that the space of quotients $B/(ABA^{-1})$ is finite. Then the sequel $\{V_i\}_{i\in\mathbb{Z}}$ of closed subsets of $L^2(\mathbb{R}^n)$ is called AB Multiresolution Analysis, if:

(i) $D_B T_k V_0 = V_0$, for ech $B \in G_B$ and $k \in \mathbb{Z}^n$

(ii) for each $i \in \mathbb{Z}$ is $V_i \subset V_{i+1}$, where $V_i = D_A^{-i} V_0$

(iii) $\bigcap V_i = \{0\}$ und $\overline{\bigcup V_i} = L^2(\mathbb{R}^n)$

(iv) there is a $\phi \in L^2(\mathbb{R}^n)$ so that $\Phi_B = \{D_B T_k \phi : B \in G_B, k \in \mathbb{Z}^n\}$ a semi orthogonal Parseval Frame of V_0. It means that Φ_B is a Parseval Frame of V_0 with the additional fact that $D_B T_k \phi \perp D_{B'} T_{k'} \phi$ is for any $B \neq B'$, $B, B' \in G_B$, $k, k' \in \mathbb{Z}^n$.

The space V_0 is called AB Scale Space and the function ϕ is called AB Scale Function für V_0. Furthermore ϕ is called orthonormal AB Scaling Function, if Φ is a ONB for V_0.

Outgoing from the theory about affine systems from section 3.3 of the book [FM-CFHLG09] and the definitions from [LLKW05] there is the possibility to define affine systems with composite dilations in $\mathbb{R}^2$ with

$$\mathcal{A}_{AB}(\psi) = \left\{ \psi_{i,j,k} = |\det A|^{\frac{j}{2}} \psi\left(B^j A^i x - k\right) : i, j \in \mathbb{Z}, k \in \mathbb{Z}^2 \right\} \tag{6}$$

Here is $\mathcal{A}_{M_{as}}(\psi) = \mathcal{A}_{AB}(\psi) = \mathcal{A}_{a,s,t}(\psi)$ with

$$M_{as} = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix} = \begin{pmatrix} a & s\sqrt{a} \\ 0 & \sqrt{a} \end{pmatrix}$$

and

$$\mathcal{A}_{a,s,t}(\psi) = \left\{ \psi_{a,s,t} = a^{-\frac{3}{4}} \psi\left(M_{as}^{-1}(x - t)\right) : a \in \mathbb{R}^+, s \in \mathbb{R}, t \in \mathbb{R}^2 \right\} \tag{7}$$

The generating functions ψ should be well localised what means that they decrease fast in frequency and spatial domain.

These systems have the basis of an affine Group. In the case of two dimensional Shearlets this is the group

$$G = \left\{ (M, t) : M \in \mathcal{D}_\alpha, t \in \mathbb{R}^2 \right\} \tag{8}$$

with $0 < \alpha < 1$ and $\mathcal{D}_\alpha \subset GL_2(\mathbb{R})$ (General Linear Group) with

$$\mathcal{D}_\alpha \left\{ M = M_{\alpha s} = \begin{pmatrix} a & -a^\alpha s \\ 0 & a^\alpha \end{pmatrix}, a > 0, s \in \mathbb{R} \right\} \tag{9}$$

$$f \mapsto \left\{ \mathcal{SH}_{\psi}^{\alpha} f\left(a,s,t\right) = \langle f, \psi_{ast} \rangle, a > 0, s \in \mathbb{R}, t \in \mathbb{R}^2 \right\} \tag{10}$$

$\mathcal{SH}$ maps $f \in L^2(\mathbb{R}^2)$ on a transform independent space which depends on the scale a, the shearing parameter s and the localisation t.
With the elements out of G a continious Shearlet System can be formed from ψ_{ast}:

$$\psi_{ast}\left(x\right) = \left|\det M_{as}\right|^{-\frac{1}{2}} \psi\left(M_{as}^{-1}\left(x-t\right)\right) \tag{11}$$

These shearlet systems are not tested or implemented in a discrete way yet. The effort in these systems will be the separation of different frequency areas in one specified scale of a Multiresolution Analysis of Shearlets. More further work will be done in research of a possible monogenic Shearlet system which could offer an extraction of the new image properties in the scales and in the specified support areas of the used Wavelets.

References

[FS01] Felsberg, M. and Sommer, G. "The monogenic signal". In: *IEEE Transactions on Signal Processing* 49 (Dec. 2001).

[FMCFHLG09] Forster, B., Massopust, P., Christensen, O., Forster-Heinlein, B., Labate, D., and Grochenig, K.C.O.N. *Four Short Courses on Harmonic Analysis: Wavelets, Frames, Time-Frequency Methods, and Applications to Signal and Image Analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2009. ISBN: 9780817648909. URL: http://books.google.at/books?id=iw-Ch7JtsFoC.

[GL11] Guo, K. and Labate, D. "The Construction of Smooth Parseval Frames of Shearlets". 2011.

[GL07] Guo, Kanghui and Labate, Demetrio. "Optimally sparse multidimensional representation using shearlets". In: *SIAM Journal on Mathematical Analysis* 39. 1 (2007), pp. 298–318. URL: http://link.aip.org/link/SJMAAH/v39/i1/p298/s1&Agg=doi.

[HSMF10] Held, Stefan, Storath, Martin, Massopust, Peter, and Forster, Brigitte. "Steerable Wavelet Frames Based on Riesz Transform". In: *IEEE Transactions on Image Processing* 19 (Mar. 2010).

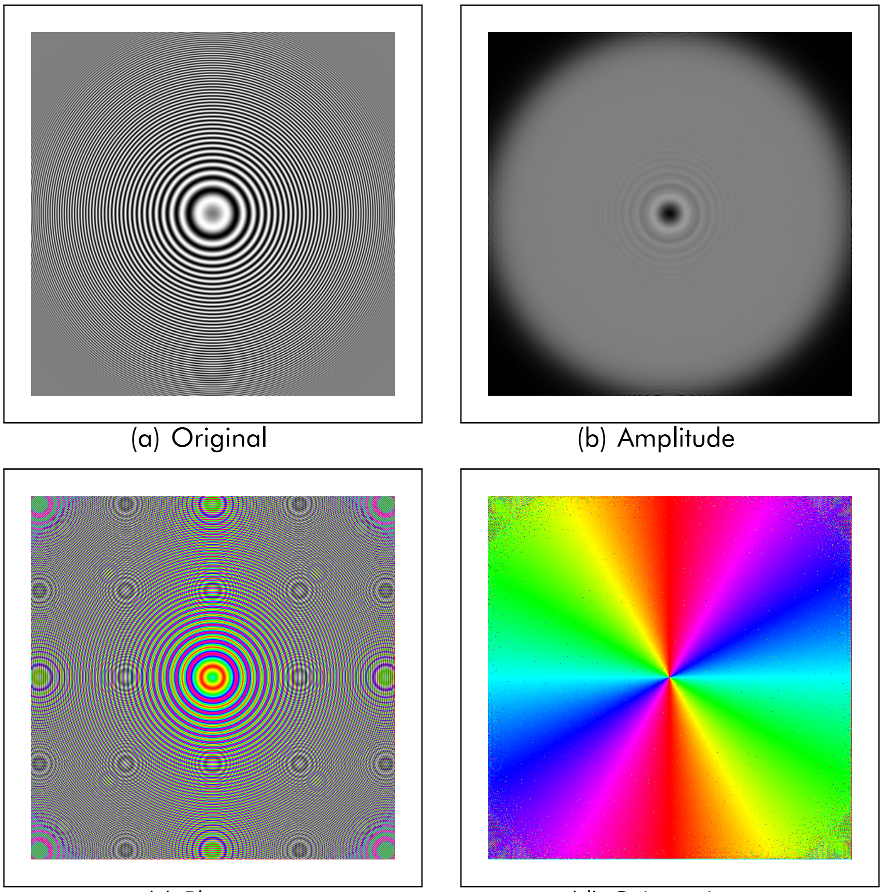
[KS05] Kutyniok, Gitta and Sauer, Tomas. "From Wavelets to Shearlets and back again". In: *Transform* (2005), pp. 1–6. URL: http://www.shearlet.uni-osnabrueck.de/papers/FWtSaba.pdf.

[LLKW05] Labate, Demetrio, Lim, Wang-Q, Kutyniok, Gitta, and Weiss, Guido. "Sparse multidimensional representation using shearlets". In: *Analysis* 5914 (2005). Ed. by Papadakis, M, Laine, A, and Unser, MEditors, 59140U–59140U–9. URL: http://link.aip.org/link/?PSI/5914/59140U/1&Agg=doi.

[LMR98] Louis, Alfred Karl, Maaß, Peter, and Rieder, Andreas. *Wavelets. Theorie und Anwendungen*. Teubner Verlag, 1998. ISBN: 3-519-12094-1.

[MP04] Ma, Jianwei and Plonka, Gerlind. "A Review of Curvelets and Recent Applications". In: *IEEE Signal Processing Magazine* 73. 1 (2004), pp. 1–34. URL: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+Review+of+Curvelets+and+Recent+Applications#0.

[UV10] Unser, Michael and Ville, Dimitri Van De. "Wavelet steerability and the higher-order Riesz transform". In: *IEEE Transactions on Image Processing* 10 (2010).



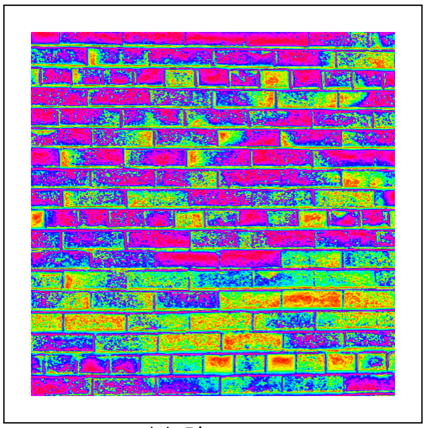
(a) Original (b) Amplitude

(c) Phase (d) Orientation

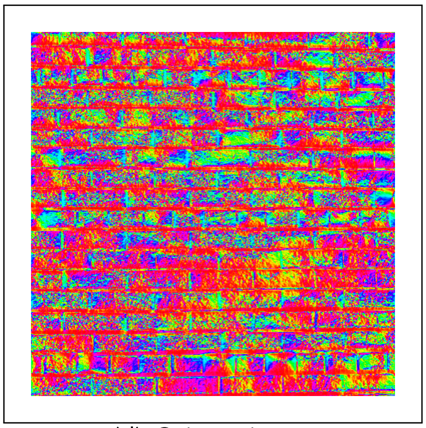
Fig. 1: Zoneplate

(a) Original

(b) Amplitude

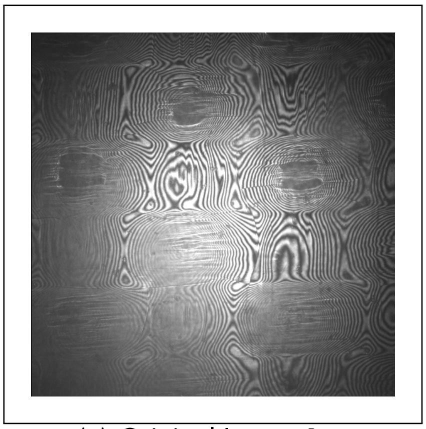


(c) Phase

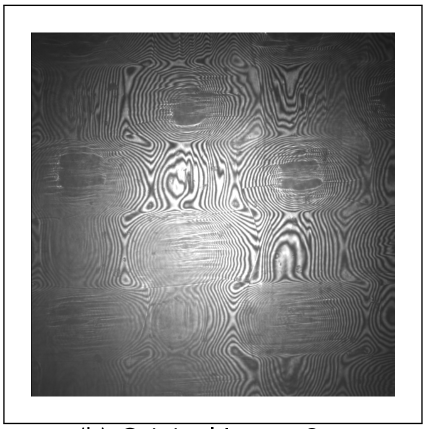


(d) Orientation

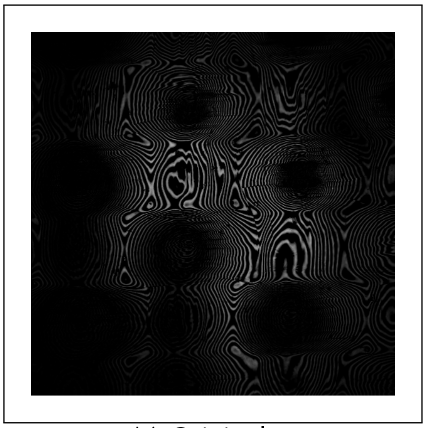
Fig. 2: Texturanalyse



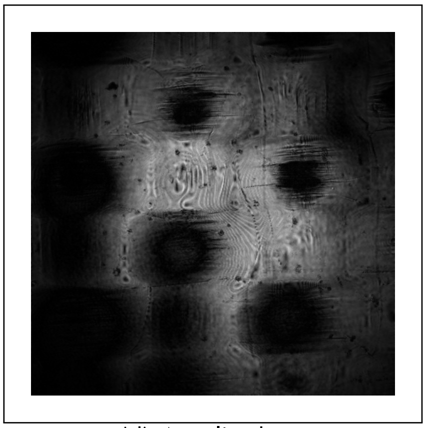
(a) Original Image 1



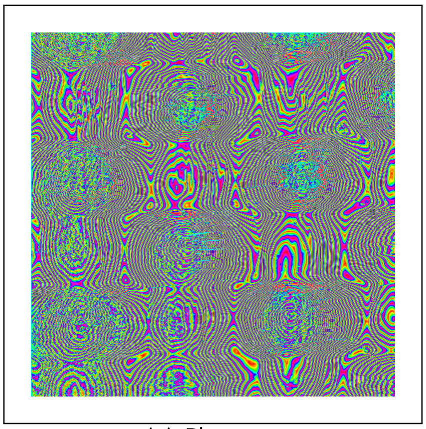
(b) Original Image 2



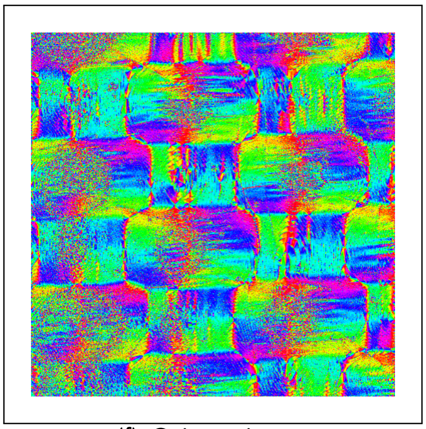
(c) Original



(d) Amplitude

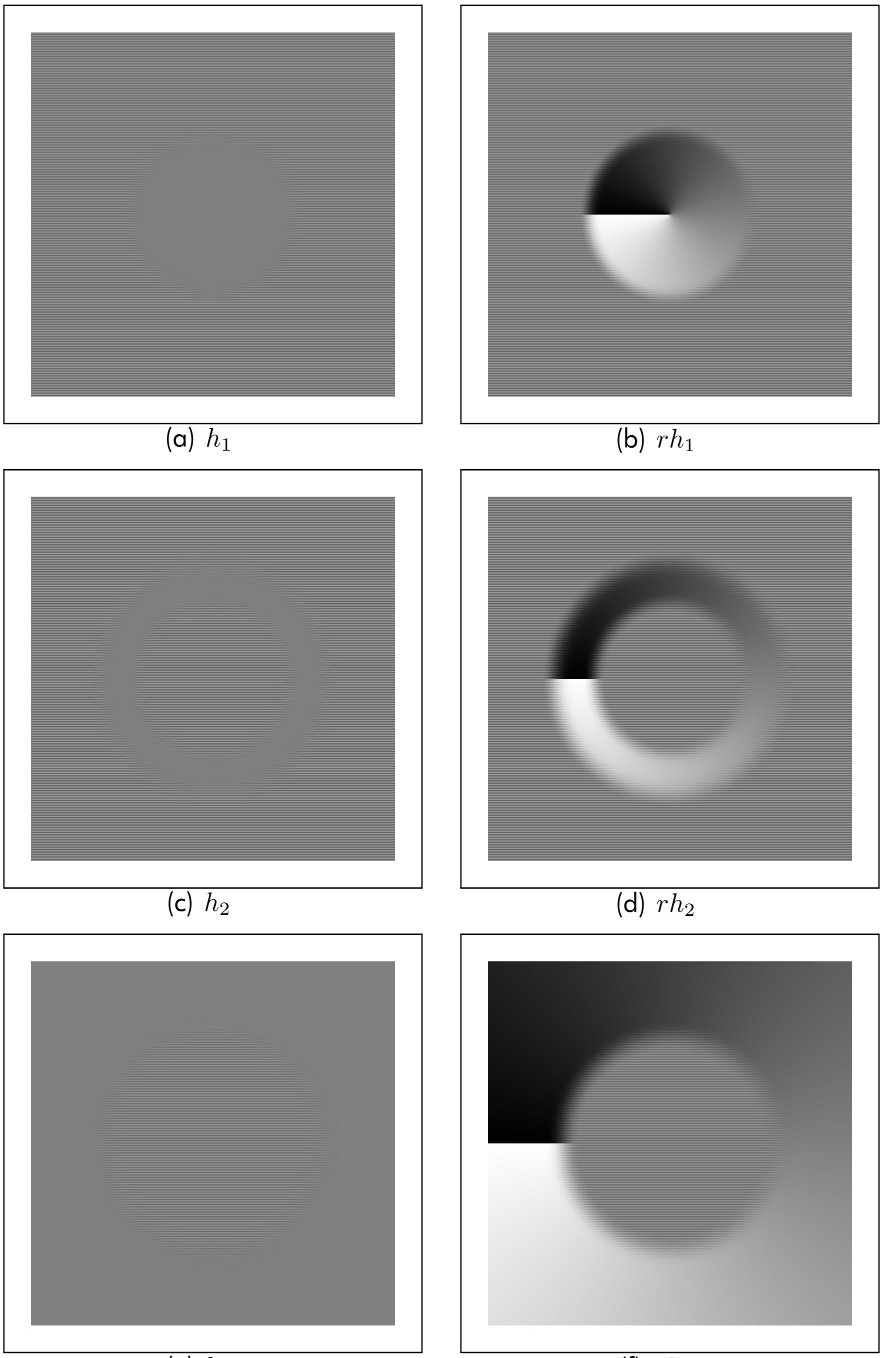


(e) Phase



(f) Orientation

Fig. 3: OCM Aufnahme



(a) h_1 (b) rh_1

(c) h_2 (d) rh_2

(e) h_3 (f) rh_3

Fig. 4: OCM - Filter

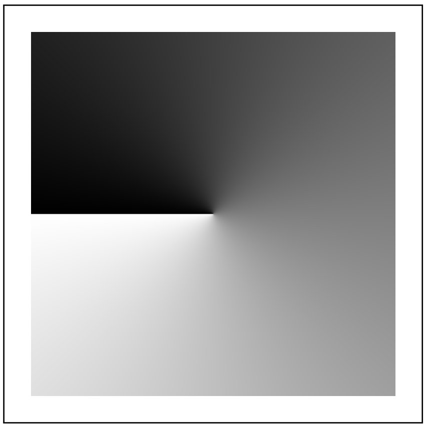


Fig. 5: OCM - Riesz

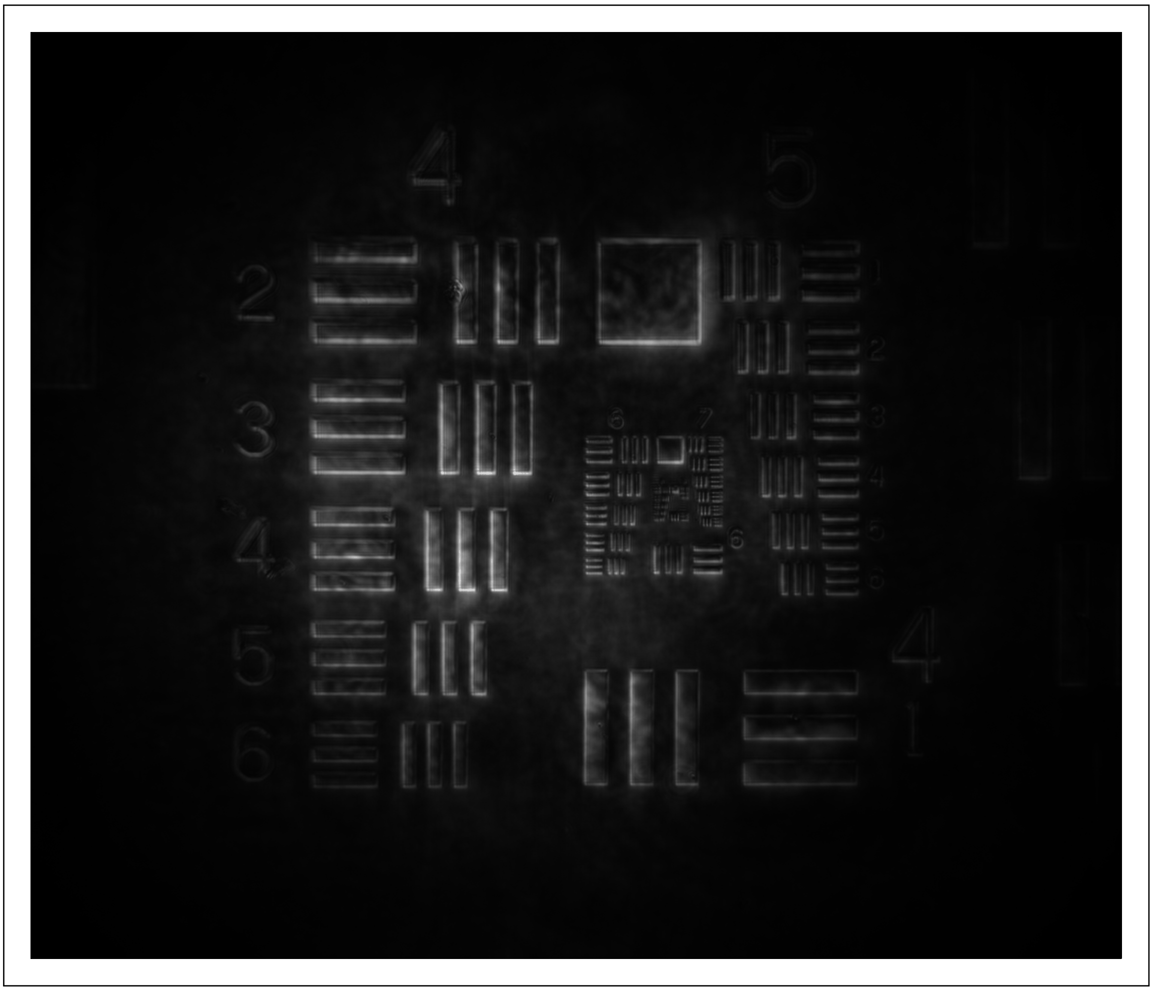


Fig. 6: OCM - Riesz Result

Detecting cracks in reciprocating compressor valves

K. Pichler[1], E. Lughofer[2], T. Buchegger[1], E.P. Klement[2], M. Huschenbett[3]

[1]Linz Center of Mechatronics GmbH, Linz, Austria
[2]Johannes Kepler University, Linz, Austria
[3]Hoerbiger Service America, Inc., Greenwood Village, Colorado, USA

Reciprocating compressors are heavily used in modern industry, for instance for gas transportation and storage. In many cases, compressors run at high capacity and without backup. Hence unexpected shutdowns lead to large losses in productivity. Furthermore, there is an economic trend towards saving labor costs by reducing the frequency of on-site inspection. Such considerations mean that compressors are run by remote control stations and monitored by automated technical systems. In this case, the system must be able to retrieve and evaluate relevant information automatically to detect faulty behavior.

The state of the art solutions for reciprocating compressor valve fault detection are designed for constant load conditions. When the load changes, operators adapt the threshold values manually. Since modern reciprocating compressors are controlled by reverse flow capacity control systems, changing load levels are not unusual, and the fault detection methods have to cope with that fact. Furthermore, most of the proposed solutions employ in-cylinder pressure measurements. We are developing a method that is based on vibration data. This makes sensor mounting easier and cheaper than the commonly used in-cylinder pressure measurements.

In this paper, the basics of the project are presented. We give an introduction about the reciprocating compressor test bench we used to acquire real world test data. A matter of special importance will be the reverse flow capacity control system to control the compressors load. This system keeps the suction valve open at the beginning of the compression phase to allow a fraction of the gas to flow back into the suction chamber. Thus it influences the timing of the valve events of the suction valves as well as the discharge valves. Then we will provide an overview of the measured data. The data contain pressure measurements, temperatures, vibrations,… Furthermore we will present a first attempt to solve the problem by employing a vibration analysis approach that determines the spectral energy in certain frequency bands. We discuss the results using this approach and some of its drawbacks. Finally, we discuss the issues to be solved for extending the method to make it independent of the load level and the valve type.

Fitness Landscape Analysis applied to instances of the Quadratic Assignment Problem Library

Michael Affenzeller

Upper Austrian University of Applied Sciences,
Heuristic and Evolutionary Algorithms Lab (HEAL)

25 April 2012

Abstract

Fitness Landscape Analysis (FLA) is a technique to extract features of problem instances that can be used for comparison and to anticipate or predict the performance of metaheuristics algorithms. This talk will present recent results on the analysis of instances of the quadratic assignment problem library (QAPLIB).

Population Decision Modeling

Wolfgang Heidl WOLFGANG.HEIDL@PROFACTOR.AT
Profactor Research, 4407 Steyr-Gleink, AUSTRIA

1. Introduction

At the beginning of each scientific analysis process stands a model. The model captures the relevant details of the system under investigation, leaves out features that are considered irrelevant and encodes relations that are assumed to exist. In short the model formalizes the assumptions we make about a system on which the analysis will be based.

In the current setting we investigate differences in visual inspection decision making (Heidl et al., 2011). Our system consists of a population of humans that carry out a repeated decision-making task (Betsch et al., 2004) over a set of stimuli images. We assume that the decision process of humans is influenced by a set of personal properties. The decision task is defined in terms of visual illustrations and is presented to each subject in the same way. The task definition is therefore a constant parameter of our system.

In a straight-forward realization of this setup the set of stimuli used for each subject to decide upon is identical and thus constant across the population. The only variables remaining in the model are the personal properties and the subject responses. With this simplified model, omitting the covariate stimuli properties, differences in decision outcomes can certainly be measured and analyzed, but lacking a decision model no further insights can be achieved.

Instead we take a structured modeling approach and introduce a nested, 2-level model, where we separate modeling of individual subject's decision behavior at the inner level from modeling of the influence of subject properties. For description of the decision behavior we intend to utilize models from the field of machine learning that are automatically adapted to the data at hand in terms of parameters but also in their very structure. To foster interpretation of those models we introduce descriptive meta features. At the outer level we model the influence of subject properties on their decision behavior and thus on the inner model. Here we will primarily resort to linear regression type models that are amenable to direct interpretation. The inferences we draw will therefore take the form of correlations between personal properties of subjects and characteristic features of their decision models.

The following treatment starts with the description of our system in the most general and abstract way, where the goal is to identify the relevant variables (factors) and to define the cause-effect relationships (dependent and independent variables, covariates) of our hypothesis. We will then refine the abstract, real world decision stimuli and introduce a feature-based description of measurable stimuli characteristics. The decision models at the inner level are thus formalized as functions mapping from the vector space of reals to the categorical outcomes. At the outer stage we simplify from the (infinite dimensional) function space of decision models by introducing finite dimensional L2-features that describe function models.

2. Model

At the inner level L1 of our model we consider repeated decision making tasks over a finite set of alternatives $Y := \{Y_1, ..., Y_k\}$. In those tasks decisions need to be made upon a series of similar stimuli $\{\psi_1, \dots, \psi_l\} \subset \Psi$ under certain time constraints. Then for each subject p, the decison-making processes is thus modeled by the function

$$\begin{aligned} f_p &: \Psi \mapsto Y \\ y_i &= f_p(\psi_i) \end{aligned} \tag{1}$$

The population of decision-makers are then represented by points in the vector space of decision functions. We assume that the influence of personal properties $v \in V$ on the decision functions are in turn governed by a functional relationship g, such that

$$\begin{aligned} g &: V \mapsto (\Psi \mapsto Y) \\ f_p(\cdot) &= g(v_p) \end{aligned} \tag{2}$$

While being most general, the formulation given here is difficult to handle practically and yields little insight when it comes to inference. In particular, the stimuli ψ

in (1) are abstract entities that need further refinement to be accessible to statistical treatment. Even with its domain and image properly defined, the utility of the function-space formulation of (2) for inference is limited. Analysis will depend on the expressiveness of a functional distance measure and results will eventually be formulated in terms of (dis)similarity to prototypical decision functions. However, in the end we would like to infer about specific instances of decision functions. Therefore we will need to define how these properties can be extracted from the functions $f_p(\cdot)$.

As indicated before, the abstract model given by (1) and (2) needs to be refined to be of practical use. We start by quantizing the the stimuli ψ. We assume a D-vector of stimuli features $\boldsymbol{x}$ relevant to the decision process can be extracted from the stimuli by the function

$$\begin{aligned}\Phi_{\boldsymbol{x}} : \Psi \mapsto \mathbb{R}^D \\ \boldsymbol{x}_i = \Phi_{\boldsymbol{x}}(\psi_i)\end{aligned} \tag{3}$$

The choice of features will be specific to the task at hand. The features can be descriptions of the stimuli on a basic or higher perceptual level, such as the power spectral density of an audio signal or shape and texture descriptions of objects in an image or even describe the basis for decision on a conceptual level such as the problem description in a simulated problem solving task (Stevens & Soller, 2005). For a set of l stimuli we will the denote the corresponding set of feature vectors as the feature matrix $\boldsymbol{X} = (\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_l)$.

To make the decision functionals amenable to a feature-based description, we will restrict $f_p(\cdot)$ to be parametrized instances from an a-priori set function class C, i.e.

$$f_p(\cdot) := f_C(\cdot, \boldsymbol{\theta}_p) \tag{4}$$

For the sake of readability we will from now on omit the class index of parametrized functions. In such functions we will write the instance parameter at the end of the argument list, separated by a semicolon. The decision model parameters $\boldsymbol{\theta}_p$ encode the specific instance of given class of models f. The models will in general differ in *description length*, i.e. in the length of their encodings for each subject. As an example, let the decision model class be decision trees of the CART flavor. The encoding of a specific tree instance could then be a list of pairs $\{k, t_k\}$ of the tree traversed in preorder, with k being feature indices and t the corresponding thresholds. For such a tree, description length thus differs with number of tree nodes. From the viewpoint of statistical modeling, $\boldsymbol{\theta}_p$ encompass both model structure and parameters. Without loss of generality we will assume that encodings of decision models are made using vectors of real numbers with length H_p for subject p i.e. $\boldsymbol{\theta}_p \in \mathbb{R}^{H_p}$.

We can now write the decision model at L1 as

$$y_{p,i} = f(\boldsymbol{x}_{p,i}; \boldsymbol{\theta}_p) \oplus \epsilon_{p,i}\,, \tag{5}$$

where we introduce random disturbances $\epsilon_{p,i}$ to the decisions to account for various sources of model error. We use the $\oplus$ operator here, since the responses y are elements from the set of decision alternatives, where addition may be ill-defined.

At the outer level L2 of our model we describe the decision functionals (i.e. decision models) by a finite set of interpretable features. To distinguish these model-describing features from the stimuli features we will from now on denote them as *meta-features* or *L2 features*. A vector of J L2 features, $\boldsymbol{s}_p \in \mathbb{R}^J$ are extracted from the parameters $\boldsymbol{\theta}_p$ of the decision models by the extraction function

$$\begin{aligned}\Phi_{\boldsymbol{s}} : \mathbb{R}^{H_p} \mapsto \mathbb{R}^J \\ \boldsymbol{s}_p = \Phi_{\boldsymbol{s}}(\boldsymbol{\theta}_p)\,.\end{aligned} \tag{6}$$

Like the L1 features, the choice (selection) of the meta-features is of course application specific and depends on the research hypothesis under investigation. Once the set of meta-features is defined, their values can be acquired for each subject. For the outer L2 of the model those features are observables. Modeling of L2 can thus draw from the full repertoire of statistical modeling.

The L2 model links the L2 features to J personal properties $\boldsymbol{v}$ of the subjects, such as age, sex and education, that will typically be acquired by means of a questionnaire. It takes the form

$$\begin{aligned}g : \mathbb{R}^J \mapsto \mathbb{R}^J \\ \boldsymbol{s}_p = g(\boldsymbol{v}_p; \boldsymbol{\beta}) + \boldsymbol{\zeta}_p\end{aligned} \tag{7}$$

where it is assumed that g defines a class of functions that are parametrized by the L2-parameters $\boldsymbol{\beta}$. Again, we introduce random disturbances $\boldsymbol{\zeta}_p$ to account for the various sources of model error.

3. Estimation

We treat decision making as a subjective process that we assume to vary in parameters *and* structure. Our approach to adequately capture those variations is to use separate instances of models for each subject. The model instances are chosen from a class of models that can adapt their complexity to fit the empirical data at hand. As such, the L1 model instances for each subject are self-sufficient and the experiments need to be designed such that parameters therein can adequately be identified. In some sense our model can be seen as longitudinal (Rabe-Hesketh et al., 2004) in that multiple measurements are available for each subject. However, the main focus of our investigation is not variation over time but variation over stimuli. Thus, at the inner level of the model we perform a true experiment, where 'treatments' (stimuli) are actively controlled and responses measured. This is in contrast to the predominant approach taken in the classical examples of multilevel statistical models, which follow an observational approach (Goldstein, 2010). Since we describe the decision making of subjects with separate, potentially differently structured models, our overall model is a composite of the L1 instances and a linking L2 model. Therefore, we seek a way of jointly estimating the parameters of the model components, while still leaving the local estimation to the specialized methods.

For this joint estimation to be meaningful, we need to set a common goal, a target against which the progress and quality of the estimation can be measured. The usual way to proceed is *maximum likelihood estimation* (MLE[1]) where we seek parameter values such that likelihood of the observed data is maximized. Given a model (assumed distribution) of the residuals a loss function for the estimation can be specified and then minimized. The most common choice are independent residuals with zero mean and equal variance, for which estimation boils down to ordinary least squares. However, in our multi-level, composite setting the simplification of equal variance is not warranted because residuals arise in different levels of the model and concern different concepts/traits.

In the following sections we will develop the reduced form distribution that specifies the joint likelihood function of our model. Assuming conditional independence of the responses of different subjects we then factor the joint likelihood into an integral over component likelihoods at the L1 and L2 levels. Since no closed-form solution exist to this integral we introduce a nonparametric sampling-based estimation procedure that combines maximum likelihood component estimates. Finally, we introduce weighted regression as a simplified, parametric solution, when the model errors in L2 are assumed to be Gaussian random variables.

[1]We will use the abbreviation MLE also for the resulting *maximum likelihood estimates*.

3.1. Reduced Form

The reduced form of our model expresses the relationship between covariates, responses and random variables in a single equation. If the meta-feature extraction function $\Phi_{\boldsymbol{s}}$ is invertible, i.e. if the decision model encoding θ_p can be fully reconstructed from the set of meta-features $\boldsymbol{s}_p$, we can combine the Equations (5), (6) and (7) obtain the reduced form

$$y_{p,i} = f\Big(\boldsymbol{x}_{p,i},\ {\Phi_{\boldsymbol{s}}}^{-1}\left(g(\boldsymbol{v}_p, \boldsymbol{\beta}) + \boldsymbol{\zeta}_p\right)\Big) \oplus \epsilon_{p,i}\,. \tag{8}$$

Now the *reduced form distribution*, the conditional distribution of the observed responses $\boldsymbol{y}$ given the covariates, $\boldsymbol{X}, \boldsymbol{v}$ and the parameters $\boldsymbol{\beta}$ is the product

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{v}; \boldsymbol{\beta}) = \prod_{p=1}^{P} p^{(2)}(\boldsymbol{y}_p|\boldsymbol{X}_p, \boldsymbol{v}_p; \boldsymbol{\beta}) \tag{9}$$

where the L2 contributions of each subject are obtained by *latent variable integration* (Rabe-Hesketh et al., 2004):

$$p^{(2)}(\boldsymbol{y}_p|\boldsymbol{X}_p, \boldsymbol{v}_p; \boldsymbol{\beta}) = \int p(\boldsymbol{\zeta}_p)\, p^{(1)}(\boldsymbol{y}_p|\boldsymbol{X}_p \boldsymbol{v}_p, \boldsymbol{\zeta}_p, \boldsymbol{\beta}_p)\, d\boldsymbol{\zeta}_p\,. \tag{10}$$

Notice that the L2 contributions are just the expected values of the L1 distributions over the L2 disturbances. The contributions at the decision level L1 are

$$p^{(1)}(\boldsymbol{y}_p|\boldsymbol{X}_p; \boldsymbol{v}_p, \boldsymbol{\zeta}_p, \boldsymbol{\beta}_p) = \prod_{i=1}^{l_p} p(\epsilon_{p,i}) \tag{11}$$

where the L1 disturbances are computed using (8),

$$\epsilon_{p,i} = y_{pi} \ominus f\Big(\boldsymbol{x}_{p,i};\ {\Phi_{\boldsymbol{s}}}^{-1}\left(g(\boldsymbol{v}_p; \boldsymbol{\beta}) + \boldsymbol{\zeta}_p\right)\Big)\,. \tag{12}$$

What is needed now for the evaluation of (9) are a feasible way to integrate over $\boldsymbol{\zeta}$, and assumed distributions $p(\boldsymbol{\zeta})$ and $p(\epsilon)$ of the disturbances. In the general case, and specifically with data-driven decision models with adaptive structure and thus varying parameter count at L1, $\Phi_{\boldsymbol{s}}$ is not invertible and the integral can neither be solved in closed form nor evaluated with Laplace approximations or quadrature-type numerical integration (Rabe-Hesketh et al., 2004). Our strategy will be to circumvent the inversion of $\Phi_{\boldsymbol{s}}$ by changing the integration variable to $\boldsymbol{\theta}$, and to use Monte Carlo sampling to evaluate the integral.

3.2. Decomposition into component likelihoods

For notational convenience we omit dependence on the covariates $\boldsymbol{X}_p$ and $\boldsymbol{v}_p$. Now we write the L2 contributions of the reduced form distribution as the marginalization over the latent variables $\boldsymbol{\theta}_p$,

$$p^{(2)}(\boldsymbol{y}_p|\boldsymbol{\beta}) = \int p(\boldsymbol{y}_p, \boldsymbol{\theta}_p|\boldsymbol{\beta})\, d\boldsymbol{\theta}_p\,. \tag{13}$$

We then factor the integrand into

$$\begin{aligned} p(\boldsymbol{y}_p, \boldsymbol{\theta}_p|\boldsymbol{\beta}) &= p(\boldsymbol{y}_p, \boldsymbol{\theta}_p, \boldsymbol{\beta})/p(\boldsymbol{\beta}) \\ &= p(\boldsymbol{y}_p|\boldsymbol{\theta}_p, \boldsymbol{\beta}) \cdot p(\boldsymbol{\theta}_p, \boldsymbol{\beta})/p(\boldsymbol{\beta}) \\ &= p(\boldsymbol{y}_p|\boldsymbol{\theta}_p, \boldsymbol{\beta}) \cdot p(\boldsymbol{\theta}_p|\boldsymbol{\beta}) \\ &= p(\boldsymbol{y}_p|\boldsymbol{\theta}_p) \cdot p(\boldsymbol{\theta}_p|\boldsymbol{\beta}) \end{aligned} \tag{14}$$

where we used conditional independence of responses $\boldsymbol{y}_p$ given the latent variables $\boldsymbol{\theta}_p$ in the last step. Therefore,

$$p^{(2)}(\boldsymbol{y}_p|\boldsymbol{\beta}) = \int p(\boldsymbol{y}_p|\boldsymbol{\theta}_p)\, p(\boldsymbol{\theta}_p|\boldsymbol{\beta})\, d\boldsymbol{\theta}_p\,, \tag{15}$$

where the first term are the L1 component likelihoods $p^{(1)}(\boldsymbol{y}_p|\boldsymbol{X}_p; \boldsymbol{\theta}_p)$ and the second term the L2 likelihoods $p^{(2)}(\boldsymbol{\theta}_p|\boldsymbol{v}_p; \boldsymbol{\beta})$. While the inversion of $\Phi_{\boldsymbol{s}}$ is no longer needed to compute the integrands of (15), the integral can still not be evaluated in closed form for all but the simplest models in L1 and L2 (Newton & Raftery, 1994). However, we can use Monte-Carlo methods to estimate the integral.

3.3. Monte-Carlo estimation by sampling from the L1 posterior

Standard Monte-Carlo estimation of (15) requires samples $\boldsymbol{\theta}_p^{(r)}$ drawn from $p(\boldsymbol{\theta}_p|\boldsymbol{\beta})$. Such samples can not be obtained from the L2 posterior distribution, $p^{(2)}(\boldsymbol{s}_p|\boldsymbol{\beta})$ since this would again require the extraction function $\Phi_{\boldsymbol{s}}$ to be invertible.

However, under quite general conditions (Newton & Raftery, 1994), *importance sampling* allows for samples to be drawn from a different, easier to handle distribution but still produce a simulation consistent estimate. To compensate for not sampling from $p^{(2)}(\boldsymbol{\theta}_p|\boldsymbol{\beta})$ density conversion weights

$$w^{(r)} = \frac{p^{(2)}(\boldsymbol{\theta}_p^{(r)}|\boldsymbol{\beta})}{p^*(\boldsymbol{\theta}_p^{(r)})} \tag{16}$$

are introduced, where $p^*(\boldsymbol{\theta}_p^{(r)})$ is the density of the *importance sampling function*. The efficiency of this approach depends on the overlap of the samples with the support of the integrand. In general, the $\boldsymbol{\theta}_p$ support of $p(\boldsymbol{y}_p|\boldsymbol{\theta}_p)$ will be highly sparse and not reside in a fixed-dimensional Eucledian space.

Analogous to (Newton & Raftery, 1994), our strategy is to use the L1-posterior $p^{(1)}(\boldsymbol{\theta}_p|\boldsymbol{y}_p)$ as our importance sampling function. In general this would be a poor importance sampling function to substitute for $p^{(2)}(\boldsymbol{\theta}_p^{(r)}|\boldsymbol{\beta})$, but here we integrate over $p^{(1)}(\boldsymbol{y}_p|\boldsymbol{\theta}_p^{(r)})$, making $p^{(1)}(\boldsymbol{\theta}_p|\boldsymbol{y}_p)$ a good choice. Several nonparametric methods exist to obtain samples (approximately) from the posterior distribution, ranging from a "poor man's" approximation (Hastie et al., 2009) using non-parametric bootstrap (Efron, 1979) estimates, to more elaborate sampling schemes like weighted likelihood bootstrap (Newton & Raftery, 1994) and Markov chain Monte Carlo sampling (Gelfand & Smith, 1990). Therefore, we have $p^*(\boldsymbol{\theta}_p^{(r)}) \approx p^{(1)}(\boldsymbol{\theta}_p|\boldsymbol{y}_p)$ and the density conversion weights become

$$\begin{aligned} w^{(r)} &\approx \frac{p^{(2)}(\boldsymbol{\theta}_p^{(r)}|\boldsymbol{\beta})}{p^{(1)}(\boldsymbol{\theta}_p|\boldsymbol{y}_p)} \\ &= \frac{p^{(2)}(\boldsymbol{\theta}_p^{(r)}|\boldsymbol{\beta}) \cdot p(\boldsymbol{y}_p)}{p^{(1)}(\boldsymbol{y}_p|\boldsymbol{\theta}_p^{(r)}) \cdot p(\boldsymbol{\theta}_p^{(r)})}\,. \end{aligned} \tag{17}$$

The importance sampling Monte Carlo estimates of the L2 contributions to the reduced form distribution are then

$$\begin{aligned} p^{(2)}(\boldsymbol{y}_p|\boldsymbol{\beta}) &\approx \sum_{r=1}^{R} p^{(1)}(\boldsymbol{y}_p|\boldsymbol{\theta}_p^{(r)}) \cdot w^{(r)} \Big/ \sum_{r=1}^{R} w^{(r)} \\ &= \sum_{r=1}^{R} \frac{p^{(2)}(\boldsymbol{\theta}_p^{(r)}|\boldsymbol{\beta}) \cdot p(\boldsymbol{y}_p)}{p(\boldsymbol{\theta}_p^{(r)})} \Big/ \sum_{r=1}^{R} \frac{p^{(2)}(\boldsymbol{\theta}_p^{(r)}|\boldsymbol{\beta}) \cdot p(\boldsymbol{y}_p)}{p^{(1)}(\boldsymbol{y}_p|\boldsymbol{\theta}_p^{(r)}) \cdot p(\boldsymbol{\theta}_p^{(r)})}\,. \end{aligned}$$

When we assume uninformative (uniform) priors $p(\boldsymbol{\theta}_p^{(r)})$ and remove the additional common factor $p(\boldsymbol{y}_p)$ we get

$$p^{(2)}(\boldsymbol{y}_p|\boldsymbol{\beta}) \approx \sum_{r=1}^{R} p^{(2)}(\boldsymbol{\theta}_p^{(r)}|\boldsymbol{\beta}) \Big/ \sum_{r=1}^{R} \frac{p^{(2)}(\boldsymbol{\theta}_p^{(r)}|\boldsymbol{\beta})}{p^{(1)}(\boldsymbol{y}_p|\boldsymbol{\theta}_p^{(r)})} \tag{18}$$

being L2-weighted *harmonic* means of the sampled L1 likelihoods.

Having generated booststrap estimates $\boldsymbol{\theta}_p^{(r)}$ for each subject p the (approximate) MLE is optained by using (18) in (9) and maximizing over $\boldsymbol{\beta}$. The maximization can start from a L2-model fit on non-boostrap L1-estimates $\boldsymbol{\theta}_p$ and then progress by (numeric) gradient decent. Note that L1-posteriors do not depend on $\boldsymbol{\beta}$ and hence need not be recomputed during the course of the optimization.

3.4. Weighted Regression

When we formulate the approximate reduced form distribution and its maximization, our aim is to properly address the varying uncertainties associated with the L1 parameter estimates and thus in the corresponding L2 responses $\boldsymbol{s}_p$ (7). For this purpose it is usually sufficient to model the uncertainties by multivariate normal distributions. The random effects (uncertainties) in the L2-model are represented by the disturbances $\boldsymbol{\zeta}_p$ for which we can now estimate per-subject covariances $\widehat{\Sigma_{\boldsymbol{\zeta}_p}}$ that can be used for weighted regression over $\boldsymbol{s}_p$. We transform the bootstrap estimates of $\boldsymbol{\theta}_p$ into L2 responses $\boldsymbol{s}_p^{(r)} = \Phi_{\boldsymbol{s}}\big(\boldsymbol{\theta}_p^{(r)}\big)$ and compute

$$\widehat{\Sigma_{\boldsymbol{\zeta}_p}} = \frac{1}{R-1}\sum_{r=1}^{R}(\boldsymbol{s}_p^{(r)} - \overline{\boldsymbol{s}_p})(\boldsymbol{s}_p^{(r)} - \overline{\boldsymbol{s}_p})^{\mathrm{T}}\,, \tag{19}$$

with

$$\overline{\boldsymbol{s}_p} = \frac{1}{R}\sum_{r=1}^{R}\boldsymbol{s}_p^{(r)}\,.$$

As indicated in the previous chapter, fitting of trained models typically requires a procedure to determine complexity controlling hyper parameters. These parameters are chosen such as to minimize the expected prediction error (EPE) of the model on future data. Commonly cross validation is employed to estimate the EPE. For a given hyper parameter set, the model is fit on the entire training data, which will in our case yield per-subject parameter estimates $\boldsymbol{\theta}_p$. To obtain the required bootstrap estimates $\boldsymbol{\theta}_p^{(r)}$ we draw bootstrap samples from the entire training data. From the view point of computational effort it might be desirable to combine the usual cross validation sampling scheme for hyper-parameter estimation with bootstrap sampling. Indeed, bootstrap schemes have also been suggested for EPE estimation (Efron, 1983) (Efron & Tibshirani, 1997). Therefore, determination of hyper parameters and bootstrap estimates of the actual parameters can be obtained from the same set of bootstrap samples.

In either case, the bootstrap estimates $\boldsymbol{\theta}_p^{(r)}$ will suffer from training-set-size bias, because the number of distinct observations in each bootstrap sample is only about 63% of the training set size (Hastie et al., 2009). This bias is negligible for hyper-parameter determination, where one a assumes that it does not influence the rank of models in the search for the minimum EPE. However, depending on the learning curve, the estimated parameters of trained models will vary systematically with training set size. This indicates another benefit of the weighted regression setting. While we can use the L2 responses computed from the whole training set as unbiased point measurements, we need resort to the biased bootstrap estimates only for the

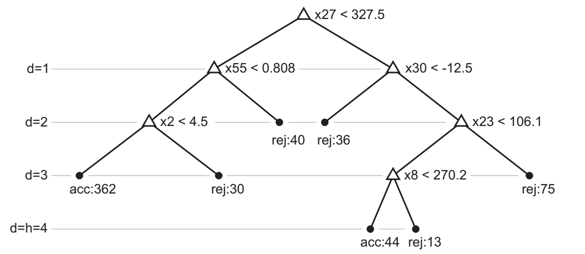


Figure 1. Decision tree induced from the responses of one subject. Triangles denote splits according to the criterion given next to it. Each leaf node is marked by a filled circle and the decision (accept/reject) associated to it. After the colon the number of samples ending up in the leaf node is given. The tree is displayed in terms of levels with equal *depth d*, with the *height* $h = 4$ being the maximum depth.

weights, that are derived from the error covariances (19).

4. Results

We illustrate the effectiveness of our proposed joint MLE approach for population decision modeling in a visual inspection study (Heidl et al., 2010) with 50 female and 50 male subjects. We will compare the statistical significance (Cohen, 1988) of population decision models obtained by ordinary regression as used in the original paper and by the weighted regression approach presented in this work. The decision behavior of each subject is modeled by CART decision trees (Breiman et al., 1993) and the extracted meta-features are regressed against subject's sex. As such we invectigate sex differences in visual inspection decision making.

To make the results more illustrative we will start with a brief introduction of decision trees and the set of model describing meta-features used in the study. We will then compare effect sizes and significance levels of the sex differences obtained by the weighted and non-weighted approaches.

4.1. Classification trees and meta-features

Figure 1 shows a typical decision tree induced from the responses of one subject in our study. To predict a subject decision y for an image with features $\boldsymbol{x}$ the decision tree is traversed downward starting from the root node n_r at the top. The split criteria at the inner nodes determine the edges to follow until a leaf node is reached. Each leaf node represents a classification result.

From Figure 1 we can for instance deduce the following rule for the leftmost branch:

IF x_{27}: Area of largest cavity < 327.5 AND x_{55}: Contrast of largest dirt in the critical zone < 0.808 AND x_2: Number of faults other than scratches < 4.5 THEN accept that part.

If we know whether the values on the right hand side of the inequations are low, high or medium (based on relative comparison), then we can transfer this rule into a more linguistic, readable one, e.g.:

IF Area of largest cavity IS NOT HIGH AND Contrast of largest dirt in the critical zone IS LOW AND Number of faults other than scratches IS LOW TO MEDIUM THEN accept that part.

As indicated in the previous subsection, trained classifiers may not in general be encoded into parameter vectors $\boldsymbol{\theta}_p$ of fixed length, making classifier comparison non-trivial. Apparently, classification trees too suffer from this problem, but *meta-features* describing the tree structure can be devised in a straight-forward manner.

Classification trees produced by the by the CART algorithm are *full binary trees*, where every node other than the leaves has exactly two children. The tree *size* N is given by the number of nodes including L leaf nodes, where $N = 2L - 1$. The *depth* d_i of a node n_i is the length of the path from the root to the node, with the maximum node depth being the tree *height* h. For trees that model decision behavior, the depth at the leaf nodes can be interpreted as effort needed to come to a decision. In addition to L and h, which depend on the graph structure alone, we can also take into account how the samples traverse the tree. If we count the number l_i of training instances traversing node n_i, and denote the set of leaf nodes $\{n_i | i \in \mathcal{L}\}$, we can compute the *average depth per sample*

$$\tilde{\mu}_d = \frac{\sum_{i\in\mathcal{L}} d_i l_i}{\sum_{i\in\mathcal{L}} l_i} , \tag{20}$$

where $\mathcal{L}$ is the set of leaf node indices. Similarly, we define the *relative depth variability* $\tilde{\sigma}'_d$ as

$$\tilde{\sigma}'_d = \frac{\tilde{\sigma}_d}{\tilde{\mu}_d}, \quad \text{with} \quad \tilde{\sigma}_d = \sqrt{\frac{\sum_{i\in\mathcal{L}} (d_i - \tilde{\mu}_d)^2 l_i}{\sum_{i\in\mathcal{L}} l_i}} . \tag{21}$$

By taking the number of traversing training instances into account we can also calculate the tree entropy

$$H = \sum_{i\in\mathcal{L}} H_i \tag{22}$$

Meta feature	Mean value (standard deviation)			Effect size (*p*-value)
	All	Female	Male	
Leaf count	6.020 (3.102)	6.860 (3.567)	5.180 (2.260)	-0.563 (0.00417)
	4.851 (2.838)	6.385 (3.106)	3.868 (2.136)	**-0.944 (0.00106)**
Average depth per sample	3.121 (0.916)	3.395 (0.882)	2.846 (0.866)	-0.628 (0.00250)
	2.982 (1.001)	3.375 (0.856)	2.610 (0.986)	**-0.828 (0.00127)**
Height	4.280 (1.638)	4.840 (1.617)	3.720 (1.457)	-0.728 (0.00064)
	3.876 (1.840)	4.795 (1.602)	3.124 (1.673)	**-1.020 (0.00042)**
Entropy	1.839 (0.552)	2.040 (0.495)	1.637 (0.532)	-0.783 (0.00019)
	1.802 (0.565)	2.049 (0.481)	1.563 (0.537)	**-0.954 (0.00005)**
Relative depth variability	0.360 (0.121)	0.407 (0.105)	0.313 (0.117)	**-0.844 (0.00006)**
	0.359 (0.117)	0.401 (0.104)	0.316 (0.112)	-0.794 (0.00013)

Table 1. Sex differences in structural meta-features of subjective classification trees. Comparison of standard approach (upper rows in the feature blocks) versus weighted approach (lower rows in the blocks). Results with higher significance of the two approaches are indicated in bold face.

with the entropy contributions of each (leaf)node

$$H_i = -p_i \log_2 p_i \,, \quad p_i = \frac{l_i}{l} \,. \tag{23}$$

4.2. Sex differences in visual inspection decision making

In Table 4.2 we compare the results obtained by the standard (non-weighted) approach to the ones obtained by the weighted approach presented in this paper. For all but the relative depth variability meta feature, the weighted approach increases the measured effect sizes and decreases the p-value by factors ranging from 1.5 for the height to around 4 for the leaf count.

The meta-feature variances for each subject have been estimated from 100 bootstrap samples of the data. The co-variance structure in the meta-features has been ignored, i.e.

$$\widehat{\Sigma_{\zeta_p}} = \begin{pmatrix} \sigma^2_{p,1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma^2_{p,J} \end{pmatrix} \tag{24}$$

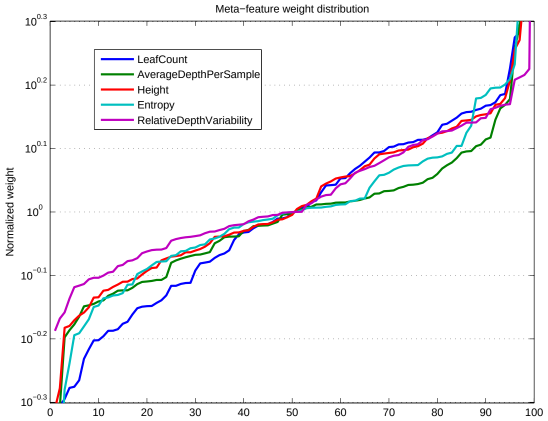


Figure 2. Distibution of the normalized meta-feature weights for 100 subjects. For each meta-feature the weights have been sorted by magnitude and normalized by their median value. The weights vary from roughly $0.5 = 10^{-0.3}$ to $2 = 10^{0.3}$ times their median.

leading to the weights $w_{p,k} = 1/\sqrt{\sigma^2_{p,k}}$ for meta-feature $\boldsymbol{s}_{p,k}$. Figure 2 shows the normalized weight distribution for all five meta-features. From the figure we can see that the weights for individual subjects vary roughly from half their median value ($10^{-0.3}$) to double their median value ($10^{0.3}$). This strongly indicates, that the uniform weight assumption of the unweighted approach is not warranted.

5. Discussion

In this paper we have shown, how machine learned models of decision behavior can be used as a measurement device to make inferences over a population of decision makers. While separate, specialized models, including their established local maximum likelihood estimation procedures are used, we can still formulate the likelihood of the composite model as the basis for a joint maximum likelihood estimation procedure. As such, the presented framework is suitable not only to the investigation of repeated decision making. It is a flexible, data-driven approach to investigate differences in how specific tasks are carried out. By choosing regression-type L1 models, the class of suitable tasks for our approach can readily be extended from ones having finite sets of possible responses to continuous responses. Also, the subjects under investigation need not be humans but can be generalized to animals and other biological and non-biological systems that respond to the application of stimuli.

Along the way of deriving our framework we made simplifications based on the following assumptions:

- The relevant aspects of subject behavior can be identified from a set of measurable stimuli, and the corresponding responses.
- The subjective decision models are instances of a chosen hypothesis-class, trained by some machine learning algorithm.
- The decision behavior of subjects is adequately represented by a finite set of meta-features that are extracted from the identified decision model instances. The L2 model rests on the extracted meta-features.

It follows from the last assumption, that the link between L1 models and the L2 model is not bijective in general. Decision models can therefore not be synthesized from personal properties of subjects.

The benefit of utilizing machine learned models at L1 is evident in the area of exploratory research. Effectively, the learning algorithm performs automatic model selection within a class of models with tunable complexity. Moreover, many machine learning algorithms are robust to non-informative features. Therefore, the researcher can safely specify an over-complete description of the stimuli when the truly relevant set of features is not yet known. For now, the application of separate instances of trained models at L1 requires experiments to be designed self-sufficient for each subject. No information is transferred between subject models that would allow joint training and reduce the required stimuli sample sizes for the individual subject. The field of transfer learning (Pan & Yang, 2010) is rapidly gaining research interest and will present a future opportunity to be applied in this work's context.

References

Betsch, T., Haberstroh, S., Molter, B., & Glckner, A. (2004). Oops, i did it again–relapse errors in routinized decision making. *Organizational Behavior and Human Decision Processes*, *93*, 62–74.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1993). *Classification and regression trees.* Boca Raton, FL: Chapman and Hall/CRC.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum Associates. 2 edition.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, *78*, 316–331.

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The .632 bootstrap method. *Journal of the American Statistical Association*, *92*, 548–560.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Goldstein, H. (2010). *Multilevel Statistical Models*. John Wiley & Sons. 4th edition.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Series in Statistics. New York: Springer. 2nd edition.

Heidl, W., Thumfart, S., Lughofer, E., Eitzinger, C., & Klement, E. P. (2010). Classifier-based analysis of visual inspection: Gender differences in decision-making. *Proceedings of SMC2010, IEEE Conference on Systems, Man and Cybernetics* (pp. 113–120). Istanbul, Turkey.

Heidl, W., Thumfart, S., Lughofer, E., Eitzinger, C., & Klement, E. P. (2011). Decision tree-based analysis suggests structural gender differences in visual inspection. *Proceedings of AIA2011, IASTED International Conference on Artificial Intelligence and Applications* (pp. 142–149). Innsbruck, Austria.

Newton, M. A., & Raftery, A. E. (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*, 3–48.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, *22*, 1345–1359.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190.

Stevens, R., & Soller, A. (2005). Machine learning models of problem space navigation: The influence of gender. *Computer Science and Information Systems/ComSIS*, *2*, 83–98.