Advances in Knowledge-Based Technologies

Proceedings of the
Master and PhD Seminar
Sommer term 2014, part 1

Softwarepark Hagenberg
SCCH, Room 0/2
4 April 2014

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: Susanne Saminger-Platz

13:00	Elisa Perrone:
	Design of experiments for Copula Models
13:30	Ciprian Zavoianu:
	Enhanced Evolutionary Algorithms for Solving Computationally Intensive Multi-Objective Optimization Problems
14:00	Kurt Pichler:
	Monitoring System for Reciprocating Compressor Valves – A data-driven Approach

Session 2. Chair: Bernhard Moser

14:45	Carlos Cernuda:
	Advanced data mining and machine learning techniques in chemometric modelling
15:15	Birgit Zauner:
	Practical Reliability Measures for Chemometric Models

Design of experiments for Copula Models*

Elisa Perrone

Abstract In applications modeling dependencies by traditional covariance functions is often of limited use. Then stochastic dependence can easily and elegantly modeled by so-called copulas, functions with very special properties that have a strong connection with arbitrary marginal distributions (See[4]). The idea is to look into the relationship between the optimal design theory and the copula theory in order to find out what could be the best combination between the design model and the copula family. A first application of copulas to the optimal design theory was treated in [1]. In this work we give a general formulation for the application of copulas to the optimal design and we show a first example in order to give a more general view to what could be the strengths and the weakness of this approach.

Key words: Copulas, Optimal Experimental Design, Fisher Information Matrix.

1 Brief description

The collection of data requires a certain amount of effort such as time. A proper design potentially allows to make use of the resources in the most efficient way.

The classical optimal design problem is the estimation of the model parameters subject to the condition that a design criterion is optimized.

The choice of the design criterion will turn out to be a crucial part of an optimal design problem.

1.1 Introduction to the Optimal Design Theory

Let us consider a vector $\mathbf{x}^T = (x_1, \dots, x_r) \in \mathscr{X}$ of control variables, where $\mathscr{X} \subset \mathbb{R}^r$ is a compact set.

The result of the observations is the vector:

Elisa Perrone

Institut für Angewandte Statistik, JKU - Linz, Austria. e-mail: elisa.perrone@jku.at

[*] PhD Thesis. Supervisor: Prof. W. G. Müller. Co-Supervisor: Prof. E. P. Klement.

$$\mathbf{y}(\mathbf{x}) = (y_1(\mathbf{x}),,\dots,y_m(\mathbf{x})),$$

with

$$\mathbf{E}[\mathbf{Y}(x)] = \eta(\mathbf{x},\beta) = (\eta_1(\mathbf{x},\beta),\dots,\eta_m(\mathbf{x},\beta)),$$

where $\beta = (\beta_1,\dots,\beta_k)$ is a certain unknown parameter vector to be estimated and η_i are known functions.

In this work we will focus on the case $m = 2$.

Let us call $c_{\mathbf{Y}}(\mathbf{y}(\mathbf{x},\beta),\alpha)$ the joint probability density function of the random vector $\mathbf{Y}$, where $\alpha = (\alpha_1,\dots,\alpha_l)$ are unknown parameters.

Definition 1. For a single observation the matrix $J(\mathbf{x},\beta,\alpha)$, a $(k+l)\times(k+l)$ matrix defined as follows

$$J(\mathbf{x},\beta,\alpha) = \begin{pmatrix} J_{\beta\beta}(\mathbf{x}) & J_{\beta\alpha}(\mathbf{x}) \\ J^T_{\beta\alpha}(\mathbf{x}) & J_{\alpha\alpha}(\mathbf{x}) \end{pmatrix} \tag{1}$$

where the matrix $J_{\beta\beta}(\mathbf{x})$ is the $(k\times k)$ matrix with the (i,j)th element defined as

$$\begin{aligned} &\mathbf{E}\left(-\frac{\partial^2}{\partial\beta_i\partial\beta_j}\log c_{\mathbf{Y}}(\mathbf{y}(\mathbf{x},\beta),\alpha)\right) = \\ &= \mathbf{E}\left(\left(\frac{\partial}{\partial\beta}\log c_{\mathbf{Y}}(\mathbf{y}(\mathbf{x},\beta),\alpha)\right)\left(\frac{\partial}{\partial\beta}\log c_{\mathbf{Y}}(\mathbf{y}(\mathbf{x},\beta),\alpha)\right)^T\right) \end{aligned} \tag{2}$$

and so are also the matrices $J_{\beta\alpha}(\mathbf{x})$ and $J_{\alpha\alpha}(\mathbf{x})$, is called the *Fisher Information Matrix*.

For r independent observations at $x_1,\dots,x_r$, the corresponding Information matrix is

$$\mathbf{M}(\xi,\beta,\alpha) = \sum_{i=1}^{r} w_i J(x_i,\beta,\alpha)$$

where $\sum_{i=1}^{r} w_i = 1$ and

$$\xi = \begin{Bmatrix} \mathbf{x_1} & \mathbf{x_2} & \dots & \mathbf{x_n} \\ w_1 & w_2 & \dots & w_n \end{Bmatrix}.$$

Definition 2. A probability distribution function ξ on the actual design space Ξ , which is the class of all the probability distributions on the Borel set $\mathscr{X}$, is called a *design measure*.

The Information Matrix on a general design measure is:

$$M(\xi,\beta,\alpha) = E(J(\tilde{x},\beta,\alpha))$$

where $\tilde{x}$ is a random vector with distribution ξ.

The aim is to approximate theory is concerned with finding $\xi^*(\beta,\alpha)$ such that maximizes some function $\phi(M(\xi,\beta,\alpha))$.

We will consider as optimal criterion a function $\phi(M) = \log \det M$, if M is non singular. This criterion is called *D-optimality* and a design that maximize this function is called *D-optimal design*.

1.2 Copulas generalities

Definition 3. Let $\mathbb{I} = [0,1]$. A *two-dimensional copula* (or *2-copula*) is a bivariate function $C : \mathbb{I} \times \mathbb{I} \longrightarrow \mathbb{I}$ with the following properties:

1. for every u_1, $u_2 \in \mathbb{I}$
$$C(u_1,0) = 0,\ C(u_1,1) = u_1,\ C(0,u_2) = 0,\ C(1,u_2) = u_2; \tag{3}$$
2. for every u_1, u_2, u_3, $u_4 \in \mathbb{I}$ such that $u_1 \leq u_3$ and $u_2 \leq u_4$,
$$C(u_3,u_4) - C(u_3,u_2) - C(u_1,u_4) + C(u_1,u_2) \geq 0. \tag{4}$$

Theorem 1. ***Sklar's Theorem***
Let $\mathbf{F}_{Y_1Y_2}$ be a joint distribution function with marginals F_{Y_1} and F_{Y_2}. Then there exists a 2-copula C such that

$$\mathbf{F}_{Y_1Y_2}(y_1,y_2) = C(F_{Y_1}(y_1), F_{Y_2}(y_2)) \tag{5}$$

for all reals y_1, y_2.
If F_{Y_1} and F_{Y_2} are continuous, then C is unique; otherwise, C is uniquely defined on $Ran(F_{Y_1}) \times Ran(F_{Y_2})$.

Conversely, if C is a 2-copula and F_{Y_1} and F_{Y_2} are distribution functions, then the function $F_{Y_1Y_2}$ given by (5) is a joint distribution with marginals F_{Y_1} and F_{Y_2}.

1.3 The connection between Copulas and Optimal Designs

According to the Sklar's theorem, in the case $m = 2$, the joint probability density function written in Equation (2) is exactly the density of the copula function such that
$$F_{Y_1,Y_2}(y_1,y_2;\alpha) = \int c_{\mathbf{Y}}(\mathbf{y}(\mathbf{x},\beta),\alpha)d\mathbf{y} =$$
$$= C(F_{Y_1}(y_1), F_{Y_2}(y_2);\alpha).$$

The general idea of this work, hence, is to use a copula function as joint distribution function of the random vector $\mathbf{Y}$ and to investigate the dependence of the design with respect to the copula choice and to the copula parameter.

References

1. Denman N.G. , McGree J.M. , Eccleston J.A., Duffull S.B.:Design of experiments for bivariate binary responses modelled by Copula functions. Computational Statistics and Data Analysis (2011).
2. V. V. Fedorov. The design of experiments in the multiresponse case. Theory of Probability and its applications, vol XVI, (1971), Nr.2.
3. O. Krafft and M. Schaefer. D-Optimal Designs for a Multivariate Regression Model.Journal of Multivariate Analysis, vol 42, (1992), pag. 130-140.
4. Nelsen R.B. : An introduction to copulas. Springer-Verlag, New York, second edition, 2006.
5. S. D. Silvey: Optimal Design. CHAPMAN & HALL/CRC, 1980.

Enhanced Evolutionary Algorithms for Solving Computationally Intensive Multi-Objective Optimization Problems

Alexandru-Ciprian ZĂVOIANU

Johannes Kepler University Linz, Austria
ciprian.zavoianu@jku.com

The task of optimizing the design of electrical drives aims to simultaneously increase efficiency, improve fault tolerance and operating characteristics, and reduce costs. Therefore, this process can be viewed as a multi-objective optimization problem (MOOP) that poses at least two big challenges:

- many degrees of freedom in the variation of design parameters;
- objectives / constraints that are non-linear and difficult to model - usually requiring time-intensive finite element (FE) simulations;

As evolutionary computation methods have generally proved efficient in tackling MOOPs, we focused our efforts on improving and adapting these state-of-the-art techniques by:

- introducing on-the-fly surrogate modeling in order to reduce the dependency on FE simulations [1] [2] [5];
- developing new algorithms that require less fitness evaluations in order to converge [3] [4];
- investigating what is the best way of distributing the optimization processes over computer cluster [6];
- synthesizing new, domain specific convergence / quality measures that can aid in the algorithm synthesis and fine tuning stages [4];

Acknowledgement
This work was conducted within LCM GmbH as a part of the COMET K2 program of the Austrian government. The COMET K2 projects at LCM are kindly supported by the Austrian and Upper Austrian governments and the participating scientific partners.

References

[1] Zăvoianu, A.C., Bramerdorfer, G., Lughofer, E., Silber, S., Amrhein, W., Klement, E.P.: A hybrid soft computing approach for optimizing design parameters of electrical drives. In: Snášel, V., Abraham, A., Corchado, E.S. (eds.) Advances in Intelligent Systems and Computing, Advances in Intelligent Systems and Computing, vol. 188, pp. 347–358. Springer Berlin Heidelberg (2013)

[2] Zăvoianu, A.C., Bramerdorfer, G., Lughofer, E., Silber, S., Amrhein, W., Klement, E.P.: Hybridization of multi-objective evolutionary algorithms and artificial neural networks for optimizing the performance of electrical drives. Engineering Applications of Artificial Intelligence 26(8), 1781–1794 (2013)

[3] Zăvoianu, A.C., Lughofer, E., Amrhein, W., Klement, E.P.: Efficient multi-objective optimization using 2-population cooperative coevolution. In: Computer Aided Systems Theory - EUROCAST 2013. pp. 251–258. Lecture Notes in Computer Science, Springer Berlin Heidelberg (2013)

[4] Zăvoianu, A.C., Lughofer, E., Bramerdorfer, G., Amrhein, W., Klement, E.P.: DECMO2 - a robust hybrid and adaptive multi-objective evolutionary algorithm. Soft Computing (minor revision) (2014)

[5] Zăvoianu, A.C., Lughofer, E., Bramerdorfer, G., Amrhein, W., Klement, E.P.: An effective ensemble-based method for creating on-the-fly surrogate fitness functions for multi-objective evolutionary algorithms. In: Proceedings of the 15th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2013), pp. 237–244. IEEE Computer Society Conference Publishing Services (CPS) (2014)

[6] Zăvoianu, A.C., Lughofer, E., Koppelstätter, W., Weidenholzer, G., Amrhein, W., Klement, E.P.: On the performance of master-slave parallelization methods for multi-objective evolutionary algorithms. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) Artificial Intelligence and Soft Computing, Lecture Notes in Artificial Intelligence, vol. 7895, pp. 122–134. Springer Berlin Heidelberg (2013)

Monitoring System for Reciprocating Compressor Valves – A data-driven Approach

Kurt Pichler, PhD-Thesis

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 and pressure conditions. For changing conditions or different valve types, operators adapt the threshold values manually. Since modern reciprocating compressors are controlled by reverse flow control systems, changing load and pressure are not unusual, and the fault detection methods have to cope with that fact. In this thesis, two independent novel approaches are presented: one evaluates measurements of accelerometers, the other one cylinder pressure measurements. Both methods can handle the tasks mentioned above, i.e. varying load and pressure conditions and different valve types. Two different approaches were developed because existing compressors are equipped with different sensing systems. Hence it is easier to upgrade the monitoring system without mounting additional sensors.

The first method evaluates time-frequency representations (spectrograms) of accelerometer measurements at the valve covers. Based on previous publications, we know that a cracked or broken valve influences the amplitudes of the power spectrum in certain frequency bands. Furthermore, it is obvious that the load control system changes the timing of the valve events. Of course, both factors are reflected in the spectrogram. Keeping that knowledge in mind, we have a look at the point-wise difference of a faultless reference spectrogram and a test spectrogram. Depending on the fault state of the valve and the load levels, it shows specifically shaped structures (Fig. 1 and Fig. 2). The positions of the structures within the spectrogram are varying unpredictably with the valve type and the load. Hence, an automated detection would be hard to realize. Additionally, measurement noise makes the detection even more difficult. Both problems can be solved by applying two-dimensional autocorrelation to the point-wise spectrogram difference: the significant structures are centered and the noise effects are reduced. Thus makes it easier to define features that characterize the specific patterns. For example, Fig. 3 shows the autocorrelation for a faultless test spectrogram, but with changing load. In contrast, Fig. 4 shows the autocorrelation for a test spectrogram measured from a valve with a fissure. The different shapes can be seen clearly.

We tested the method with numerous real world test measurements. The measurements were recorded with different valve types, different accelerometers and with constant as well as varying load conditions. All of the tests proved the ability of the method to detect cracked and broken valves, cross validation using SVM classification shows very high classification accuracy.

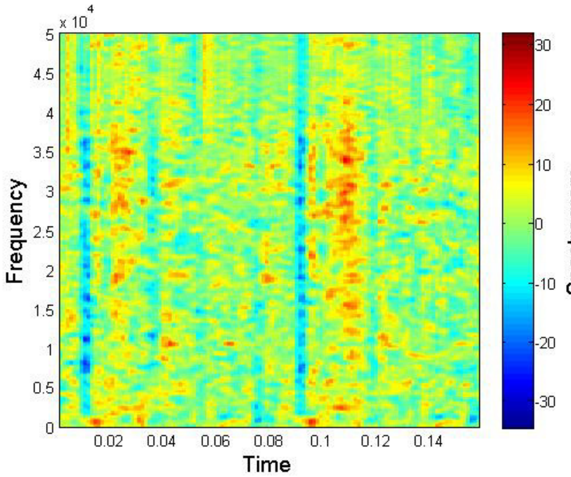

Fig. 1: Point-wise spectrogram difference for a test spectrogram from a faultless valve

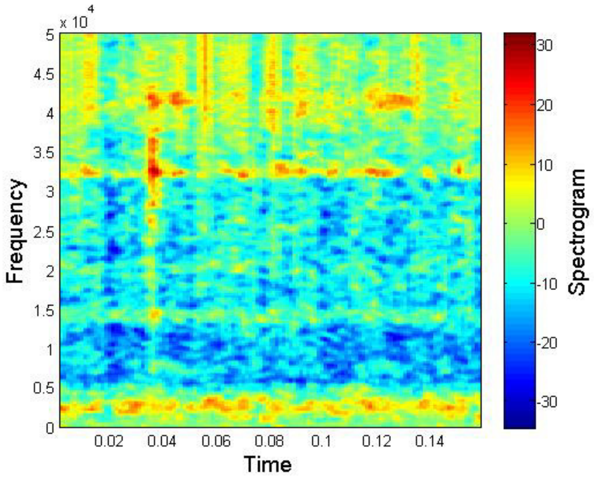

Fig. 2: Point-wise spectrogram difference for a test spectrogram from a cracked valve

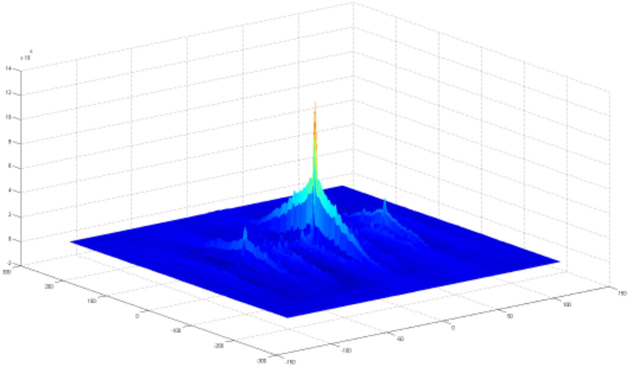

Fig. 3: Autocorrelation representation for a test measurement from a faultless valve

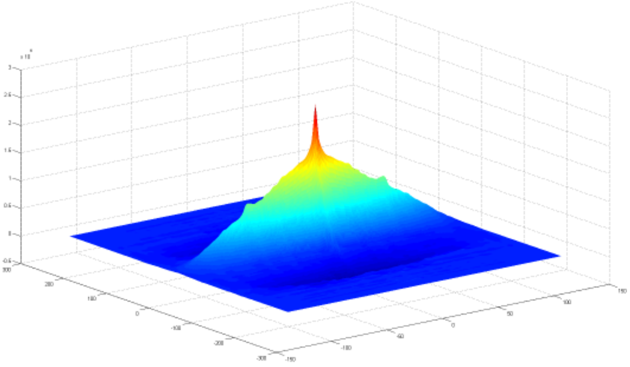

Fig. 4: Autocorrelation representation for a test measurement from a cracked valve

The second method evaluates cylinder pressure measurements in the shape of pV diagrams. When a valve of a compressor breaks, there is a leak and gas can flow through the closed valve. Of course, this affects the shape of the pV diagram of a compression cycle significantly. The pV diagram is used to describe corresponding changes in volume and pressure in a system. As the load control affects mainly the compression phase of a compression cycle, we concentrate on the evaluation of the expansion phase. This leads to a load independent method.

In the case of a broken discharge valve, the pressure in the cylinder decreases slower during expansion than in the faultless case. The reason for that is that gas flows through the closed valve from the discharge chamber into the cylinder. In the case of a broken suction valve, gas flows through the suction valve from the cylinder into the suction chamber resulting in a faster decreasing cylinder pressure. To quantify this difference, we linearize the pV diagram by switching to logarithmic scales. Then we can easily use the gradient of the expansion phase as an indicator for pressure reduction velocity in the compression cylinder during expansion (Figure 5). Since the pressure reduction velocity is also affected by the pressure conditions (suction and discharge pressures) we have to consider the pressure conditions in the feature space as well (Figure 6). But even faultless valves are not 100% leak tight. Depending on the valve type (design and material), they have different leakage factors. This is reflected in an offset in the feature space. The features can be classified using SVM classification. We validated the proposed method with real world data from a reciprocating compressor test bench. Compared with another feature extraction method from pV diagrams proposed in literature (F. Wang, L. Song, L. Zhan and H. Li, 2010, “Fault diagnosis for reciprocating air compressor valve using p-V indicator diagram and SVM”), our features show higher validation accuracy, especially in the case of small faults.

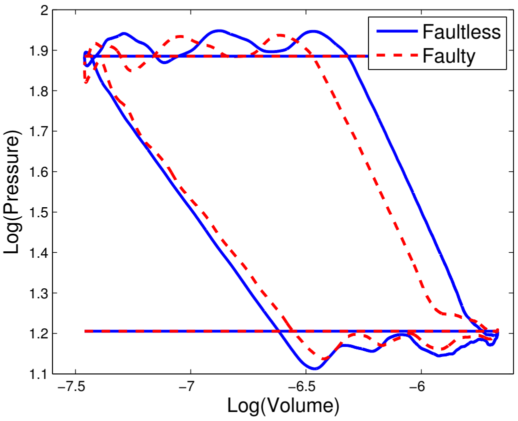

Figure 5. Logarithmic pV diagrams of measurements from faultless and faulty valves

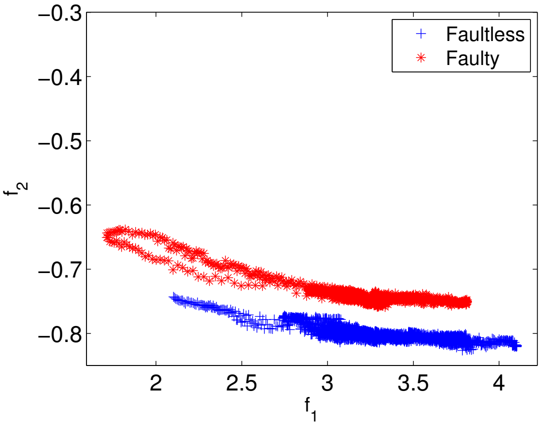

Figure 6. Features extracted from the logarithmic pV diagrams

Advanced data mining and machine learning techniques in chemometric modelling

Carlos Cernuda[1]

[1] *Johannes Kepler University Linz, Austria*
*(*carlos.cernuda@jku.at*)*

Chemical information is of increasing importance in Process Analytical Chemistry. Its adequate use can assure maximum yield and product quality while minimizing energy consumption and waste production, having a direct impact on the productivity and thus competitiveness, and on the environmental issues of the respective industries.

Chemometrics is defined by the International Chemometrics Society as *the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods.*

Technology today allows us to collect huge amounts of data, thus new techniques are required to handle it efficiently. We have worked in all the steps of the chemometric modeling process: i) preprocessing by new outlier detection approaches based on the distributions of the statistics involved in Principal Component Analysis, ii) data filtering/cleaning, with enhanced variable selection procedures based on Swarm Intelligence algorithms like Genetic Algorithms, Ant Colony Optimization or Particle Swarm Optimization [1, 2], iii) off-line modeling by introducing fuzzy inference systems like Flexible Fuzzy Inference Systems in the chemometrics field [3], iv) on-line modeling by means of evolving chemometric models[4] (eChemo), v) robustness analysis by defining several local and global error bars and confidence intervals as well as ensembling strategies for handling repeated measurements [5], and vi) cost reduction, by means of decremental and incremental active learning approaches [6].

In summary, we have developed modern and robust data mining and machine learning algorithms to contribute to the improvement of chemometric modelling in general.

References

[1] C. Cernuda, E. Lughofer, P. Hintenaus, W. Märzinger,: *Enhanced genetic operators design for waveband selection in multivariate calibration based on NIR spectroscopy.* Journal of Chemometrics, **28** (2014) 13–26

[2] C. Cernuda, E. Lughofer, W. Märzinger, W. Summerer: *Hybrid evolutionary particle swarm optimization and ant colony optimization for variable selection. Aplication to near infra-red spectroscopy.* 3rd World Conference on Information Technology (WCIT-2012), Barcelona, Spain, **4** (2013) 6–13

[3] C. Cernuda, E. Lughofer, W. Märzinger, J. Kasberger: *NIR-based quantification of process parameters in polyetheracrylat (PEA) production using flexible non-linear fuzzy systems* Chemometrics and Intelligent Laboratory Systems, **109** (2011) 22–33

[4] C. Cernuda, E. Lughofer, L. Suppan, T. Röder, R. Schmuck, P. Hintenaus, W. Märzinger, J. Kasberger: *Evolving Chemometric Models for Predicting Dynamic Process Parameters in Viscose Production.* Analytica Chimica Acta, **725** (2012) 22–38

[5] C. Cernuda, E. Lughofer, P. Hintenaus, W. Märzinger, T. Reischer, M. Pawliczek, J. Kasberger: *Hybrid adaptive calibration methods and ensemble strategy for prediction of cloud point in melamine resin production* Chemometrics and Intelligent Laboratory Systems, **126** (2013) 60–75

[6] C. Cernuda, E. Lughofer, G. Mayr, T. Röder, P. Hintenaus, W. Märzinger, J. Kasberger: *Decremental Active Learning for Optimized Self-Adaptive Calibration in Viscose Production* Proceedings of the SSC 2013 Conference, Stockholm, Sweden, 2013

Reliability Measures – Ways to Estimate the Reliability of Model Predictions in the Field of Chemometrics

Birgit Zauner, Thomas Natschläger – Software Competence Center Hagenberg

In many industrial applications, e.g. the online monitoring of a chemical process, where the goal is to predict certain quality parameters, a measure of the reliability of predictions is desirable or even necessary. In our contribution, we deal with this matter and investigate several different reliability measures (RMs), developed within the context of chemometrics. These measures are all based on the assumption that the distance (measured in different ways) between the model calibration data and a new data point is related to the prediction quality for this point.

We apply the developed measures to data from an industrial partner from the field of melamine resin production and discuss obtained results. While certain advantages over commonly used reliability measures can be seen, there are also weaknesses that will be lined out and provided with a possible explanation.