s c c h
software competence center
hagenberg

FUZZY.LOGIC.LAB.LINZ

Advances in Knowledge-Based Technologies

Proceedings of the Master and PhD Seminar Winter term 2008/09, part 2

Softwarepark Hagenberg
SCCH, Room 0/2
February 25th, 2009

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

14:00–15:45 Session 1 ***(Chair: Roland Richter)***

14:00 Fabrizio Durante:
Multivariate Extreme Value Models via Copulas

14:45 Enrico Foscolo:
Inferential aspects of copula modeling

15:15 Rachele Foschi:
Evolution of dependence and ageing for bivariate survival models

15:45 Coffee Break

16:00–17:30 Session 2 ***(Chair: Bernhard Moser)***

16:00 Henrike Stephani:
Enhancing the Interpretability of Terahertz Data through Unsupervised Learning

16:30 Edwin Lughofer:
Evolving Vector Quantization for Classification of On-Line Data Streams

17:00 Ulrich Brandstätter, Roland Richter:
flllame – FLLL's Advanced Machine Learning Environment

Leila Muresan *et al.*:
High resolution microarray image analysis

Multivariate Extreme Value Models via Copulas

Fabrizio Durante

Department of Knowledge-Based Mathematical Systems
Johannes Kepler University, A-4040 Linz, Austria
e-mail: fabrizio.durante@jku.at

Nowadays, several multivariate models have been considered to account for the dependence between extremes, especially in view of their possible application to risk measurement in environmental sciences — see, e.g., [4, 7, 1] and references therein. In this respect, the concept of *copula* has provided to be very useful for the construction of *Multivariate Extreme Value* models.

Here, we recall that a multivariate copula $\mathbf{C}(u_1, \ldots, u_d)$ is a joint distribution over $\mathbf{I}^d$, whose univariate margins are uniform. The link between d-copulas and multivariate distributions is provided by Sklar's Theorem [8], asserting that any d-dimensional joint distribution function F can be expressed in terms of its margins $F_1, \ldots, F_d$, and a suitable *copula* C, implicitly defined by:

$$F(x_1, \ldots, x_d) = C\left(F_1(x_1), \ldots, F_n(x_d)\right). \tag{1}$$

Copulas fully describe the overall structure of dependence of the variables of interest, and provide a global model for their stochastic behaviour. They are also fundamental for the definition of several rank-based *measures of association*, like *Kendall's* τ and *Spearman's* ρ, which are frequently used in applications. For a thorough theoretical and practical introduction about copulas see, respectively, [3, 6] and [5, 7].

Now, suppose that F is a *multivariate extreme value* (shortly, MEV) distribution, i.e. F is the distributional limit of (stationary) sequences of component-wise maxima of d-dimensional random vectors. Then, it is well known that all the margins F_i's of F are *Generalized Extreme Value* laws, and the corresponding copula C is MEV, i.e. it satisfies the relation

$$C\left(u_1^t, \ldots, u_d^t\right) = \left[C\left(u_1, \ldots, u_d\right)\right]^t \tag{2}$$

for all $\mathbf{u} \in \mathbf{I}^d$ and $t > 0$ [7]. Two MEV copulas are of particular interest, $\Pi_d(\mathbf{u}) = u_1 \cdots u_d$ and $M_d(\mathbf{u}) = \min\{u_1, \ldots, u_d\}$. The former one models independent variates, while the latter one models comonotone dependent ones, where each variable is a monotone increasing function of the others.

The main target pursued in this talk is to present a general multivariate framework for modeling the maxima sampled via a network of non-independent gauge stations. To this aim, we present a new multi-parameter MEV copula with some distinguished properties:

1. the parameters can be interpreted in terms of multivariate measures of association and/or dependence;

2. the simulations are simple and fast;

3. the procedures for fitting empirical data may use an analogue of the Method of Moments technique; this may be convenient when the joint distributions are not absolutely continuous.

This model has been recently proposed in [2].

In order to illustrate the practical utility of this new family of MEV copulas, we consider in the talk a set of three certified gauge stations recording annual maximum flood data in northwestern Apennines and Thyrrhenian Liguria basins (Italy): (*1*) Airole, (*2*) Merelli, and (*3*) Poggi. Then, we compare this new MEV model with other models already presented in the literature.

References

[1] G. Balkema and P. Embrechts. *High risk scenarios and extremes*. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich, 2007.

[2] F. Durante and G. Salvadori. On the construction of multivariate extreme value models via copulas. *Environmetrics*, in press, 2008.

[3] H. Joe. *Multivariate models and dependence concepts*, volume 73 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1997.

[4] Y. Malevergne and D. Sornette. *Extreme financial risks*. Springer-Verlag, Berlin, 2006. From dependence to risk management.

[5] A.J. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management. Concepts, techniques and tools*. Princeton Series in Finance. Princeton University Press, Princeton, NJ, 2005.

[6] R.B. Nelsen. *An introduction to copulas*. Springer Series in Statistics. Springer, New York, second edition, 2006.

[7] G. Salvadori, C. De Michele, N.T. Kottegoda, and R. Rosso. *Extremes in Nature. An Approach Using Copulas*, volume 56 of *Water Science and Technology Library*. Springer, Dordrecht (NL), 2007.

[8] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

Inferential Aspects of Copula Modeling

Enrico Foscolo*

25th February 2009

Abstract

In this paper we propose a brief review of the main estimation methods regarding copulas parameters. Firstly, we show fully parametric method through the so-called *inference functions for marginals* (IFM), then we enter into a set of semiparametric procedures, among which we mention the *pseudo maximum likelihood* (MPL) and the *minimum distance* (MD) methods. We highlight in particular the issue of robustness, useful if one can not be able to determine the appropriate copula for our data. Finally, we show a series of results obtainned in large simulation studies, in order to comment the performance of each estimator.

1 Introduction

There are many fields in which we face multivariate phenomena and tools are required in order to detect joint behaviour of several random variables. Let us focus our attention on financial, insurance or environmental (in particular, hydrology) area: it becomes aware of the fact that phenomena, which are treated, are mostly multidimensional. Over the last ten years it has been made a large use of copula-based models because of their useful properties. Without loss of generality, let us consider here only the bivariate case and parametric families of copulas with one-dimensional parameter θ.

From a formal point of view we can define copulas C as two-place functions with uniform marginals: copulas are precisely distribution functions that express the dependence between random variables.

Let (X, Y) be a continuous bivariate random vector with probability distribution $H_{X,Y} : \Re^2 \mapsto [0, 1]$ and marginals $F_X(x) := P(X \leq x)$ and

*Department of Statistical Science, University of Bologna, Italy.
E-mail: enrico.foscolo2@unibo.it

$G_Y(y) := P(Y \leq y)$. Specifically, a function C is called *copula* if it satisfies the following conditions:

(C1) for all $u, v \in [0,1]$,

$$C(u,0) = 0 = C(0,v) \tag{1}$$

and

$$C(u,1) = u = C(1,u) \tag{2}$$

(C2) for every $u_1, u_2, v_1, v_2 \in [0,1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$,

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0. \tag{3}$$

The well-known Sklar's theorem (see Sklar (1959)) elucidates the role that copulas play in the relationship between a bivariate distribution function and its univariate marginals.

Theorem 1.1 (Sklar, 1959) *Let H a bivariate distribution function with marginals F and G; then a copula exists such that for all x, y in $\Re$*

$$H_{XY}(x,y) = C[F(x), G(y)] \tag{4}$$

Moreover, if F and G are continuous, then the copula $C(\cdot,\cdot)$ is unique.

Let us note that, if C is a copula and F and G are distribution functions, then $C[F(x), G(y)]$ is a joint distribution function with marginals F and G.

We could sentencing that the main advantage of copula-based models is to separate marginal behaviour from joint behaviour:

- by varying $C(\cdot)$ one alters the dependence structure,
- by varying $F(\cdot)$ and $G(\cdot)$ one alters the marginal behaviour.

Both operations are permitted. In this way, the selection of an appropriate model for the dependence between X and Y, represented by copula, can proceed independently from the choice of the marginal distributions. Copulas are an effective way to describe multivariate dependence with any kind of marginal distribution.

Despite the large literature, there are still many problems to be solved. From a statistical point of view, as Genest and Favre (2007) point out, *"inference for copulas is* [...] *still under development"* and *"formal methodology for testing the goodness-of-fit of copula models is just emerging"*.

So the aim of this paper is to classify the most common estimating procedures for copulas parameters and compare the results of simulation studies in order to verify the performance in term of mean squared error.

Since there is not available a well-recognized tool for identifying the correct parametric copula for a specific data set, great attention will be given to the issue of robustness. With this goal we shall show the results presented in Mendes et al. (2007) and Foscolo et al. (2008), where some estimators are tested in different situations of contagion of data generating process.

2 The Maximum Likelihood and Maximum Pseudo Likelihood Methods

The classical statistical inference theory that can be applied here is the *asymptotic maximum likelihood estimation (MLE)*.

Now, suppose that a parametric family of copulas $\{C_\theta : \theta \in \Theta\}$ has been considered as a model for the dependence between two random variables X and Y.

Given a random sample (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), from $H_{XY}(x, y) = C_\theta \{F(x), G(y)\}$, we wish to estimate θ.

Let us denote the density associated with F, G and C as f_δ, g_η, c_θ respectively. The density of $H_{XY}(x, y)$ is given by

$$\begin{aligned} h(x,y) &= \frac{\partial^2}{\partial x \partial y} H(x,y) = \frac{\partial^2}{\partial x \partial y} C_\theta (F_\delta(x), G_\eta(y)) = \\ &= f_\delta(x)\, g_\eta(y)\, c_\theta (F_\delta(x), G_\eta(y)) \end{aligned} \tag{5}$$

The *Maximum Likelihood Estimator* (or *Exact Maximum Likelihood Method*, see Cherubini et al. (2004)) of (δ, η, θ) is then defined by maximizing

$$l(\delta, \eta, \theta) = \sum_{i=1}^{n} \ln f_\delta(x_i) + \sum_{i=1}^{n} \ln g_\eta(y_i) + \sum_{i=1}^{n} \ln c_\theta (F_\delta(x_i), G_\eta(y_i)) \tag{6}$$

Multivariate optimization is computationally intensive and sometimes there is a non-negligible risk that the true optimum is not found. For these reasons, Joe (1997) suggests to proceed in two steps (*Inference Functions For marginals (IFM)*):

- estimate δ and η separately from the marginal log-likelihoods,

$$\sum_{i=1}^{n} \ln f_{\delta}(x_i), \quad \sum_{i=1}^{n} \ln g_{\eta}(y_i) \tag{7}$$

- for $i \in \{1, \ldots, n\}$ set $\widehat{U}_i = F_{\widehat{\delta}}(x_i)$ and $\widehat{V}_i = G_{\widehat{\eta}}(y_i)$;
- now choose $\widehat{\theta}$ as the estimated value of θ that maximizes the score function

$$\sum_{i=1}^{n} \ln c_{\theta}\left(\widehat{U}_i, \widehat{V}_i\right) \tag{8}$$

The main advantage of this method is that it is easy to implement, but the main drawback is that an inadequate selection of marginal distributions could reflect on the estimation of the dependence parameter θ, even when $\mathcal{C} \in C_{\theta}$ holds true.

Moreover, Joe (2005) showed that the two-stage estimator is asymptotically normal and that, under reasonable regularity conditions, it is as efficient as one-stage maximum likelihood estimator (see also Zhao and Harry (2005)).

In addition, there are other possible ad hoc estimation methods that have been proposed for overwhelming the hard computational efforts to get exact MLE or IFM. These methods share, and also mix, concepts from nonparametric statistical inference and simulation techniques.

Semiparametric copula models are based on parametric copulas and on nonparametric estimation of marginal distribution functions. As the marginals are unknown, they must be estimated in some way. The safest way is nonparametric estimation of the marginals:

$$\widehat{F}(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{I}(X_j \leq x) \tag{9}$$

$$\widehat{G}(y) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{I}(Y_j \leq y) \tag{10}$$

To avoid boundary problems, $\widehat{F}(x)$ and $\widehat{G}(x)$ are often rescaled by replacing n by $(n+1)$ in the denominator.

In the bivariate case, the *Maximum Pseudo Likelihood Method* (MPL) requires that C_{θ} be absolutely continuous with density c_{θ} and it simply consists of doing an optimization of the multivariate likelihood as a function of the dependence parameter θ (see Genest et al. (1995) and Genest and Favre (2007)):

$$l(\theta) = \sum_{i=1}^{n} \ln c_\theta \left(\widehat{F}(x_i), \widehat{G}(y_i) \right) \tag{11}$$

Only rank-based estimators are considered in the sequel for F and G:

$$\widehat{F}(x_i) = \frac{R_i}{n+1} \tag{12}$$

$$\widehat{G}(y_i) = \frac{S_i}{n+1} \tag{13}$$

where (R_i, S_i) are the pairs of ranks associated with the sample. This methodological choice is justified by the fact that ranks are the best summary of the joint behaviour of the random pairs. So the method involves maximizing a rank-based log-likelihood score function.

The MPL method is just easier to implement as Joe's one, because you only need to take the ranks. As a copula function is invariant by monotone increasing transformations of the marginals, it seems natural to expect the same of its parameter estimates. The estimator

$$\widehat{\theta}_n := arg \max_{\theta} \sum_{i=1}^{n} \ln c_\theta \left(\frac{R_i}{n+1}, \frac{S_i}{n+1} \right) \tag{14}$$

of θ does just that.

There is a clear loss of information in assuming nonparametric estimation for the marginals. If the parametric form of the marginals was known, Joe's method would obviously be preferable.

Genest et al. (1995) proved that the estimator $\widehat{\theta}_n$ is asymptotically normal and that, under reasonable regularity conditions, it is as efficient as one-stage maximum likelihood. These calculations do not make any parametric assumption about marginal distributions. Thus,

$$\widehat{\theta}_n \approx N \left(\theta, \frac{\nu^2}{n} \right) \tag{15}$$

where ν^2 is the sample variance computed from pseudo-observations (to compute the variance and the pseudo-observations we remind the reader to Genest and Favre (2007)).

2.1 The Weighted Version of Maximum Likelihood Estimators

It is reasonable to assume that the parametric and semiparametric approach work well when the choice of the marginals and copula itself is done in a correct way. Let us suppose that it does not happen, that is there is a situation of misspecification or contamination of the data generating process. In these cases we would need to use some robust tools in order to estimate the dependence parameter θ.

Mendes et al. (2007) obtain an automatic robust procedure that should work well when there are or not contamination in the data. The *Weighted Maximum Likelihood Estimators (WMLE)* consist in a robustification of the MLE, where points previously identify as outliers, i.e. far from the main mass of points.

Mendes et al. (2007) introduce an estimation method for copulas parameters in two steps. At first step they are not concerned with efficiency. The goal is to choose points which do not seem to follow the dependence structure defined by the majority of the points. Identification of points is based on the Mahalanobis distances computed using robust estimates for covariance and location, being the cutoff point the 0.975-quantile of a chi-square random variable with 2 degree of freedom. Points with robust distances greater than cutoff are identified as *atypical* and are given zero weight. In the second step they obtain the maximum likelihood estimates of $\widehat{\theta}$ of copula parameters θ, using just those data points with assigned weights equal to 1.

These estimates are expected to posses good bias properties but larger variances when compared to the MLE under true model. Under contaminated models, Mendes et al. (2007) show through simulations that for the majority of scenarios considered they possess small bias and variance and outperform the MLE.

3 The Minimum Distance Estimators

Suppose that the copula associated with the two-dimensional distribution function H is $\mathcal{C} \in C_\theta$. On the other hand, we have the nonparametric estimation of copula, called *Empirical Copula* (see Deheuvels (1979)), formally defined by

$$C_n(u,v) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{I}\left(F(x_i) \leq u, G(y_i) \leq v\right) \tag{16}$$

where $\mathbf{I}(A)$ denoting the indicator function of set A.

Copula measures of goodness-of-fit, named *Minimum Distance Estimation (MDE)*, may be obtainned by computing distance between empirical copula C_n and the parametric copula $\mathcal{C}$ fitted to the data. Let us define the minimum distance functional T on the space of the copula by

$$T(D) := arg\min_{\theta} \rho\left(C_n, \mathcal{C}\right) \tag{17}$$

Here ρ is a distance between probabilities on $[0,1]^2$. Let us consider the empirical process $\mathbb{C}_n := \sqrt{n}\left(C_n - \mathcal{C}\right)$. In Tsukahara (2005), Mendes et al. (2007) and Foscolo et al. (2008) it has been considered

- the *Cramér-von Mises distance*,

$$\rho\left(C_n, \mathcal{C}\right) = \int_{[0,1]^2} \mathbb{C}_n^2\left(\mathbf{u}\right) d\mathbb{C}_n\left(\mathbf{u}\right) \tag{18}$$

- the *Kolomogorov-Smirnov distance*,

$$\rho\left(C_n, \mathcal{C}\right) = \sup_{\mathbf{u}\in[0,1]^2} \left|\mathbb{C}_n\right| \tag{19}$$

The empirical version of these distance are given by, respectively,

- $$\rho_{emp}\left(C_n, \mathcal{C}\right) = \sum_{i=1}^{n} \left\{C_n\left(u_i, v_i\right) - \mathcal{C}\left(u_i, v_i\right)\right\}^2 \tag{20}$$

- $$\rho_{emp}\left(C_n, \mathcal{C}\right) = \max_{u,v\in[0,1]^2} \left|C_n\left(u_i, v_i\right) - \mathcal{C}\left(u_i, v_i\right)\right| \tag{21}$$

The MDE estimate for θ is the solution $\widehat{\theta}$ which minimizes over all $\theta \in \Theta$ the selected empirical copula based goodness-of-fit statistic:

$$\widehat{\theta} = arg\min_{\theta} \rho_{emp}\left(C_n, \mathcal{C}\right) \tag{22}$$

For a precise formulation of the robustness of MDE and for the proof of asymptotic normality we remind the reader to Tsukahara (2005).

3.1 The Weighted Version of Minimum Distance Estimators

Mendes et al. (2007) propose different weight functions which emphasize different regions on the unite squared and are able to handle different locations of model violation: so they obtain the so-called *Weighted Minimum Distance Estimators (WMDE)*.

In order to emphasize deviations in the *tails* (the corners of the unit squared), they apply the weight function

$$w\left(u_i, v_i\right) = \frac{1}{\sqrt{\left[\mathcal{C}\left(u_i, v_i\right)\right]\left[1 - \mathcal{C}\left(u_i, v_i\right)\right]}} \tag{23}$$

or the squared weight function

$$w\left(u_i, v_i\right) = \frac{1}{\left[\mathcal{C}\left(u_i, v_i\right)\right]\left[1 - \mathcal{C}\left(u_i, v_i\right)\right]} \tag{24}$$

However, they note that this goal may be better achieved by using the weight function

$$w_1\left(u_i, v_i\right) = \frac{1}{\sqrt{\left[u_i + v_i - \mathcal{C}\left(u_i, v_i\right)\right]\left[1 - \mathcal{C}\left(u_i, v_i\right)\right]}} \tag{25}$$

which emphasizes just the points in the lower left (called LL) and the upper right (called UR) corners (denoted by LL+UR).

When using function (25), points in the lower right quadrant (called LR), in the upper left quadrant (called UL), in the middle have the same lack of influence on the resulting statistics. Thus, they propose a weight function which assign more weight to points located in the middle and in the LL corner, the UR corner, and in both corners, respectively:

$$w_{2,LL}\left(u_i, v_i\right) = \sqrt{1 - u_i - v_i + \mathcal{C}\left(u_i, v_i\right)} \tag{26}$$

$$w_{2,UR}\left(u_i, v_i\right) = \sqrt{\mathcal{C}\left(u_i, v_i\right)} \tag{27}$$

$$w_2\left(u_i, v_i\right) = \sqrt{1 - u_i - v_i + 2\mathcal{C}\left(u_i, v_i\right)} \tag{28}$$

The weight function (26) represents the square root of the probability mass (as given by $\mathcal{C}$) in the rectangle $[x, 1] \times [y, 1]$, and it increases from 0 to 1 as the point (x, y) moves towards $(1, 1)$. The weight function (27) represents the square root of the probability mass in the rectangle $[0, x] \times [0, y]$, and it increases from 0 to 1 as the point (x, y) moves towards $(0, 0)$. The weight

function (28) represents the square root of the sum of the above probabilities and possesses the nice property of downweighting just points located at the LR and UL corners.

Mendes et al. (2007) observe that the weight function given by (23) is too flat in the middle and gives much more weight to LL when compared to UR. Their first proposed weight function (25) is an improvement, since it enhances almost equally both the LL and the UR corners and does not emphasize the LR and the UL corners. The second weight functions proposed ((26), (27), (28)) are even more promising because they give equal weights to the LL and UR corners and to the middle points, just downweighting the points in LR and UL corners.

According to the copula type (possessing or not tail dependence) it could emphasize just the LL or the UR quadrant. The following weight functions are based on variations of (25), (26), (27), (28) (still more weight may be given to the tails if we consider second degree statistics, which use squared weights):

(LL)

$$\frac{1}{\sqrt{u_i + v_i - \mathcal{C}(u_i, v_i)}} \tag{29}$$

$$\frac{1}{u_i + v_i - \mathcal{C}(u_i, v_i)} \tag{30}$$

$$1 - u_i - v_i + \mathcal{C}(u_i, v_i) \tag{31}$$

(UR)

$$\frac{1}{\sqrt{1 - \mathcal{C}(u_i, v_i)}} \tag{32}$$

$$\frac{1}{1 - \mathcal{C}(u_i, v_i)} \tag{33}$$

$$\mathcal{C}(u_i, v_i) \tag{34}$$

In order to obtain robust estimators for the dependence parameter θ, we mention Foscolo et al. (2008). They choose the *Cramér-von Mises distance* (18) in order to estimate θ, but, in this form, they note that the estimator gives biased estimates, in particular when the size of the sample is small. As Genest et al. (1995) note, *"there is typically enough data to obtain nonparametric estimates of the marginal distributions, but insufficient information to afford nonparametric estimation of the structure of association"*.

No matter what kind of copula, they observe that, when θ increases, the points are always along the diagonal of the squared $[0,1]^2$. In this way, they identify a sort of edge which separates the *"central"* data from the *"outskirts"* ones (which, instead, have different membership according to the fact that $\mathbf{u}$ and $\mathbf{v}$ come from a copula with a certain θ). They use the following algorithm (based on *conditional distribution method*, see Nelsen (2006)):

- a series of n values u from a uniform r.v. $(0,1)$ is generated;
- always using the same u values, a series of n values v is generated through the method quoted above for a total number of times defined by the user (in this case, they choose to simulate $n = 5000$ v-values for a number of simulation m equal to 2000);
- in this way, they can build the following matrix,

	1° Simul.	...	m° Simul.
u_1	v_{11}	...	u_{1m}
...	...	...	...
u_n	v_{n1}	...	u_{nm}

- they calculate the means of the rows and their standard deviations;
- finally, they define the edges for the selection of data in the following way

$$upperBound = \overline{\boldsymbol{v_{i.}}} + costant * St.D.(\boldsymbol{v_{i.}}) \tag{35}$$

$$lowerBound = \overline{\boldsymbol{v_{i.}}} - costant * St.D.(\boldsymbol{v_{i.}}) \tag{36}$$

  where $i = 1, \ldots, n$ and *costant* is chosen by the user; the points outside the edges will turn up the searched observations.

Thus, the proposed estimator, named *Modified Minimum Distance Rank Estimator* (*MMDRE*), has the following form:

$$\rho_{emp}(C_n, \mathcal{C}) = \sum_{i=1}^{n^\star} \left\{ C_n(u_i^\star, v_i^\star) - \mathcal{C}(u_i^\star, v_i^\star) \right\}^2 \tag{37}$$

where $n^\star$ represents the number of *"outskirts"* pairs $(u_i^\star, v_i^\star)$.

4 Simulation Study

To check and compare the performance of various estimators, a simulation study is carried out in Tsukahara (2005), Mendes et al. (2007) and Foscolo et al. (2008).

In Tsukahara's simulation study four well-known one-parameter families of bivariate copulas are chosen: Clayton, Gumbel-Hougaard, Frank and Plackett copula.

For each value of θ, Tsukahara generates 500 sample of size 100 and computes five estimates: in particular pseudo maximum likelihood estimator (14), minimum Cramér-von Mises distance estimator (18) and minimum Kolmogorov-Smirnov distance estimator (19). The last two estimators are based on the link between the parameter θ and the measures of association Kendall's τ and Spearman's ρ.

Tsukahara notes that, in case of correct specification of copula $\mathcal{C}(u, v)$, the pseudo maximum likelihood estimator performs better than the others; it has the smallest mean squared error and bias in many cases. The performance of the minimum Kolmogorov-Smirnov distance estimator is almost always worst: for all four families, the estimator becomes unstable in that its estimated bias and mean squared error both get larger as the positive dependence gets stronger.

Mendes et al. (2007) compare MLE, WMLE, MDE, WMDE in a large simulation study, where the experiments consider ϵ-contaminated parametric copula families containning varying proportions ϵ of contaminating points located at different regions of the copula support. The selected families include elliptical copulas (Normal), copulas for extreme values (Gumbel, Husler-Reiss), copulas widely used in practice (Frank, Joe, Tawn) and some other families defined in Joe (1997).

They set ϵ equal to 0%, 5%, 10%: these proportions of atypical points are generated from a contaminating bivariate normal distribution $F^{\star}$ with correlation coefficient $\rho = 0.00$ and very small variances. They establish five possibilities for the location of the contaminating points: the center of the unit squared and the regions nearby the four corners. Three sample size are considered (50, 100, 300) and the number of scenarios for each one is 1000. For sake of comparisons, for all copula models they set θ such that corresponding kendall's τ would be equal to 0.00, 0.25, 0.50.

Mendes et al. (2007) give the winner(s) under no contamination and for the contaminated models. The results typically do not depend on the sample size and on strength of dependence. Moreover, the efficiency of all estimators increase with sample size. They show that for each copula family there is one or a couple of robust estimates performing very well, in the sense of small

mean squared error, despite the contamination percentage and location and the sample size. They suggest that, when fitting copulas possessing upper-tail dependence, it should use the upper right weighted version of (18) and (19). The lower left weighted versions of minimum distance estimators should be preferred for copulas possessing lower-tail dependence. When fitting elliptical copulas or those possessing no-tail dependence, it could select the weighted version of maximum likelihood estimator or the estimator (18) with squared weights (25).

Finally we report the results of Monte Carlo simulations shown in Foscolo et al. (2008). A simulation study has been carried out for Clayton, Frank and Gumbel-Hougaard copula. For each of these functions they estimate the dependence parameter θ, using (14), (18) and (37): 2000 pseudo-random samples of size $n = \{100, 300, 500\}$ are simulated from each of the three copulas with $\theta = \{2, 3, 4, 5, 6, 7, 8\}$. They analyse the case of correct specification, when the copula assumed in their estimator is just the one used to simulate, and the case of *totally misspecification*, when this ideal situation does not hold.

Because of its properties, the pseudo maximum likelihood method seems to be always the best estimator, the most correct and efficient in case of correct specification. Moreover, this comment is independent from the size of samples. On the contrary, the estimators (18) and (37) appear more biased then (14), although they make to mark good results in term of unbias and efficiency when the sample occurs with a medium/large size ($n = \{300, 500\}$). When they consider the case of totally misspecification, maximum pseudo likelihood estimator strongly underestimates the parameter, while (18) and (37) seem to converge to the correct values. The sample size again helps to improve the efficiency of the estimators.

5 Concluding Remarks

Parametric and semiparametric estimators of copulas parameters have been discussed and the issue of robustness have been introduced. Tsukahara (2005), Mendes et al. (2007) and Foscolo et al. (2008) agree that maximum likelihood procedure seems to be always the best estimator in term of mean squared error, when model is not contaminated, even though in Mendes et al. (2007) it is always possible to find a weighted minimum distance estimator as good as the maximum likelihood estimator. With the exception of minimum Kolmogorov-Smirnov distance estimator, which is almost always worst, under contamination the robust estimators always presented superior performance. We note that the efficiency of all estimators increase with sample size. As

Mendes et al. (2007) point out, it seems that the parametric family of copula is not determinant when finding the more robust estimator, but if it posses lower or upper tail dependence.

References

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance.* Wiley Finance Series. Chichester: John Wiley & Sons.

Deheuvels, P. (1979). La fonction de dependance empirique et ses propriétés. un test non paramètrique d'independance. *Bull. Cl. Sci., Acad. R. Belg.*, 65:274–292.

Foscolo, E., Ayyad, C., Porcu, E., and Mateu, J. (2008). Some new estimators for copula-based models. pages 217–228. Statistics for Spatio-Temporal Modelling. Proceedings of the Fourth International Workshop on Spatio-Temporal Modelling (METMA4), Publisher Democratica Sarda.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4):347–368.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

Joe, H. (1997). *Multivariate models and dependence concepts.* Monographs on Statistics and Applied Probability. 73. London: Chapman and Hall.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *J. Multivariate Anal.*, 94(2):401–419.

Mendes, B. V. M., de Melo, E. F. L., and Nelsen, R. B. (2007). Robust fit for copula models. *Communications in Statistics - Simulation and Computation*, 36(5):997–1017.

Nelsen, R. B. (2006). *An introduction to copulas. 2nd ed.* Springer Series in Statistics. New York, NY: Springer.

Sklar, A. (1959). Fonctions de répartition á n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris*, 8:229–231.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *Can. J. Stat.*, 33(3):357–375.

Zhao, Y. and Harry, J. (2005). Composite likelihood estimation in multivariate data analysis. *Can. J. Stat.*, 33(3):335–356.

Evolution of dependence and ageing for bivariate survival models

Rachele Foschi
Department of Mathematics
University "La Sapienza", Rome
foschi@mat.uniroma1.it

Let (X, Y) be a pair of two non-negative, continuous random variables, with survival function $\overline{F}(x, y)$. In different applied fields, as reliability, survival analysis, interacting defaults, interest arises in studying the conditional survival function of X, Y, given the event that X, Y fall over a threshold,

$$\overline{F}_t(x, y) = P\left(X > t + x, Y > t + y | X > t, Y > t\right).$$

We denote by K the *survival copula* of $\overline{F}$ and by K_t, for $t > 0$, the survival copula of $\overline{F}_t$. Since we aim at studying the *evolution* of dependence properties of $\overline{F}_t$, for $t \geq 0$, we consider the family $\{K_t\}_{t\geq 0}$ of survival copulas.

Starting with a notion of dependence, we define different notions of dependence for K, reflecting the dependence properties of the copulas K_t for some t (see [2]).

An analogous investigation concerns the ageing properties of the model $\overline{F}$, described by the ageing function B, associated with $\overline{F}$ (see [1]). Again, a family $\{B_t\}_{t\geq 0}$ is defined, representing ageing properties of the models $\overline{F}_t$, for $t \geq 0$. A property of an element of the family, B_t, can be expressed as a different property of B.

For both the families, we are interested in studying which properties are preserved for t spanning $\mathbb{R}_+$.

The relationships among dependence and bivariate ageing are also investigated.

References

[1] B. Bassan, F. Spizzichino, *Relations among univariate ageing, bivariate ageing and dependence for exchangeable lifetimes*, Journal of Multivariate Analysis 93 (2005) 313-339

[2] F. Durante, R. Foschi, F. Spizzichino, *Threshold copulas and dependence properties*, Statistics and Probability Letters 78 (2008) 2902-2909.

[3] R. Foschi, *Semigroups of semi-copulas and a general approach to hyper-dependence properties*. Submitted (2009).

[4] R. Foschi, F. Spizzichino, *Semigroups of semicopulas and evolution of dependence at increase of age*, Mathware & Soft Computing XV (2008) 95-111.

[5] R. B. Nelsen, *An Introduction to Copulas*, Springer, New York, 1999.

Enhancing the Interpretability of Hyper-Spectral Terahertz Images with Unsupervised Classification

Henrike Stephani
Fraunhofer Institute for Industrial Mathematics (ITWM)
Fraunhofer Platz 1
67663 Kaiserslautern
Email: henrike.stephani@itwm.fraunhofer.de

***Abstract*—We present the applicability of hierarchical agglomerative cluster algorithms to Terahertz (THz) spectroscopic analysis of chemical compounds as well as hyper-spectral images. We show the influence of different windowing and filtering methods in the spectral data pre-processing to enhance the clustering results. Two distance measures are compared: Classical Euclidean distance on the full frequency range and a distance working only on the minima of the spectra. We further propose an interactive clustering process for THz hyper-spectral image classification and visualization using a combination of the time-domain and frequency-domain spectral information.**

I. Introduction

With the generation of a high volume of data comes the necessity of computer aided data organization. Unsupervised classification, also known as cluster analysis, is one method to do that. The goal of cluster analysis is to categorize a given amount of data without prior knowledge about the resulting classes. We use clustering tools for the analysis of hyper-spectral imaging data.

Hyper-spectral imaging is a way of using other spectral bands than the one of visible light to analyze material. Such an image is generated by gathering a spectrum at each pixel. These spectra usually contain hundreds to thousands of measured channels. Thus, the data has high volume - being the number of pixels - as well as high dimensionality [1]. The most prominent examples for hyper-spectral imaging are ultraviolet and infrared imaging but with the development of THz spectroscopy also the spectral bandwidth of 100GHz to 10THz is used [2]. Industrial applications do mainly focus on two different properties of THz waves. Firstly, most packaging materials such as ceramics, cloth, carton, and plastic are non-absorband thus non-destructive testing is possible. Secondly, most chemical compounds have a characteristic absorption spectrum and hence can be detected.

Cluster analysis is applied here to improve the visualization of the THz imaging especially by combining information from time resolved as well as frequency resolved spectra.

II. Methods

The main influence factors for clustering are the chosen features and the distance measure. The computational cost of these algorithms primarily lies in the calculation of the distance measure depending on the dimensionality of the feature space. In spectral analysis it is possible to analyze both time domain as well as frequency domain features. The usual features of the time domain have low dimensionality - making the distances easily computable - while in the frequency domain the dimensionality is high and a number of preprocessing steps need to be executed beforehand.

A. Clustering

Classification in general deals with assigning classes to a given set of data. In supervised classification the desired classes are known beforehand and a sufficiently characterizing set of samples from each class is given. In clustering only the data is given and is being ordered with respect to a certain distance measure. The distance between samples in one class is to be minimized while the distance between clusters is to be maximized [3]. In this paper classical agglomerative hierarchical clustering shall be used. This algorithm operates in the following way: Beginning with one sample per cluster in each iteration the closest clusters are being united. Assuming a given distance between the samples, the closeness of the cluster is defined by the link function, generalized by the Lance-Williams formula [4]. When merging the clusters C_j and C_k the distance of the resulting cluster C_{jk} to another cluster C_l will be:

$$D(C_{jk}, C_l) = \alpha_j D(C_j, C_l) + \alpha_k D(C_k, C_l) +$$

$$\beta D(C_j, C_k) + \gamma |D(C_j, C_l) - D(C_k, C_l)|.$$

For the clustering in the time domain the following coefficients are used: $\{\alpha_j, \alpha_k, \beta, \gamma\} = \{1/2, 1/2, 0, 1/2\}$. The resulting distance is called complete link function. It is especially beneficial when the clusters are not sharply separated but blend in each other. If we assume that each cluster varies around a "prototype" the average link distance - $\{\frac{C_j}{C_j+C_k}, \frac{C_k}{C_j+C_k}, 0, 0\}$ - is to be preferred. We use these coefficients for clustering in the frequency domain. The hierarchical clustering procedure results in a tree-graph called dendrogram. It is particularly useful in navigating between different coarseness levels, i.e. showing the samples being clustered in different numbers of categories. The visualization of high-volume results can be interactively improved by discarding certain branches of the dendrogram altogether and displaying other branches in a finer way.

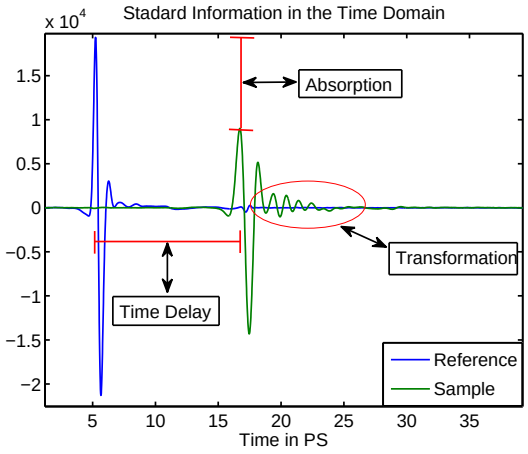


Fig. 1. Three characteristics of time domain spectra.

B. Features and Distance Measure

In time domain spectroscopy typical information that is considered can be seen in Fig. 1. By the diminishment of the main peak the absorption of the sample is measured, the refraction is measured by the time delay and the transformation of the peak contains the characteristic spectral information. The decline of the peak and the time delay are used for the first clustering. Both are normed by taking the highest and lowest peak and the first and last position as references.
In the frequency domain the most commonly used metric is the Euclidean metric on all frequencies. This approach's main disadvantage is the so-called "curse of dimensionality" [5]. Handling many dimensions the accumulations of small differences over all frequencies can have the same effect as one big differing peak. The proposed alternative approach is useful in coping with this. THz spectra have comparatively broad bandwidth features. Given a smooth shape of the spectra, it is therefore possible to find the local long-term minima of each spectrum. Those represent the difference between the clusters. Minima are computationally characterized here by being the point where the medium gradient of an interval switches from negative to positive. The choice of the size of the interval and the smoothness of the spectra is essential in finding the relevant minima instead of noise. It is helpful that noise generates sharp minima which can be ignored by choosing the intervals sufficiently wide. The distance measure is defined as follows:

Definition 1: Let $X_f = \{x_1, ..., x_r\}$ be the frequencies of the minima of Sample S_X and $Y_f = \{y_1, ..., y_k\}$ be the frequencies of the minima of sample S_Y. The distance between S_Y and S_X is then defined by:

$$D(S_X, S_Y) = \sqrt{\sum_{f \in X_f \cup Y_f} (S_X(f) - S_Y(f))^2).}$$

As most THz spectra of solids have few but broad peaks, an immense features space reduction is achieved and thereby the problems of high dimensionality are diminished. Furthermore the computational costs of calculating a distance matrix are shrunk considerably.

C. Preprocessing and Filtering

It is a common method to apply apodization functions prior to the Fourier transform to eliminate undesirable effects. The Fourier transform acts on the assumption of infinite periodic signals while experimentally given data usually is finite. Thus the effect of spectral leakage appears. Another problem are measured side effects which influence the spectrum. Such are fore example an echo of the pulse spectrum or the original pulse appearing within the sample spectrum. Two kinds of windowing functions are applied in this paper to deal with these effects. The main pulse is considered the starting point of the relevant information. Information before this pulse is suppressed generously with a Nuttall window while information coming after the pulse is hardly altered. Only to eliminate the effects of spectral leakage a Hamming window is used [6]. Two filtering methods were compared. These are Wavelet Shrinkage and Savitzky Golay Filters. The latter is one of the most prominent filters in Chemometrics. This is due to its easy and fast calculability as well as its capacity to preserve peaks [7].
Another approach is using Wavelet shrinkage [8]. The main idea of Wavelet transformation is a loss free hierarchical decomposition of a signal based on a basis of so called Wavelets which can be considered band pass filters with certain desirable properties. The main advantage over the Fourier transform is the possibility to express characteristics with respect to as well frequency as time. In that way it is possible to control the degree of influence certain frequencies have on the signal depending on the interval they occur in. Although the amount of information one gets is quite sophisticated the discrete Wavelet transformation is computationally very efficient thanks to a down-sampling mechanism that is applied during the procedure [9].

III. Application

We have to application examples. On the one hand - to illustrate the effects of preprocessing and filtering - a data set of high resolution spectra of five chemical compounds, namely PABA, acetyl salicylic acid, salicylic acid, lactose, and tartaric acid, nine spectra each compound is used. On the other hand we carry out clustering on a hyper-spectral image of a letter containing different materials among others chemical compounds.

A. Spectra of Chemical Compounds

The Windowing is applied to the time resolved spectra in the above mentioned way. In each spectrum the maximal peak is determined and depending on the resolution the windows are applied. As the single compound spectra have a very good resolution the windowing mainly serves to suppress the

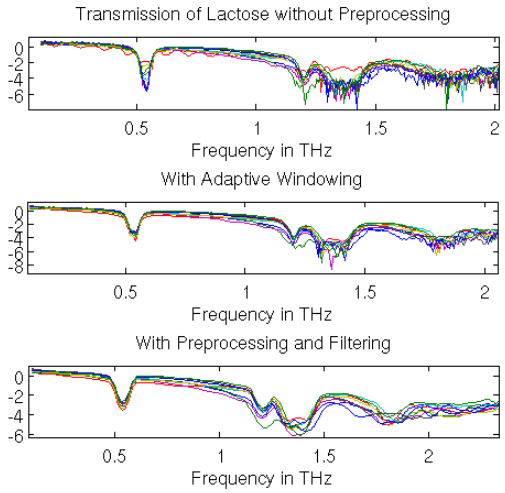


Fig. 2. Spectra of lactose. The relevant peak at 0.5 THz is well preserved by preprocessing and filtering while the spectra are smoother and more similar to each other.

measured fraction of the original pulse. As can be seen in Fig. 2 the windowed spectra have clearer peaks and contain less noise - with respect to similarity - than the unwindowed ones.

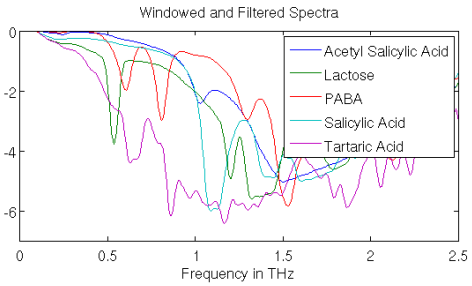


Fig. 3. Logarithmic transmission spectra of all five compounds after preprocessing and windowing.

The spectral density of the measurements declines in higher frequencies. It is advisable to discard information below a certain spectral density. To further enhance the finding of relevant minima and maxima filtering is applied. The Savitzky-Golay filter produces similar results as the Wavelet shrinkage. The peak preservance is slightly better in Wavelet shrinkage in finding minima and maxima both perform similar. The Savitzky-Golay filter is chosen for the further analysis because it is computationally faster. As proposed before, classical agglomerative clustering is applied. The evaluation is done by considering misclassifications and quality of clusters in terms of inter- and intra-cluster distance. Clustering the unpreprocessed spectra results in more than 30% misclassification depending on the clustering level.

The classification of the pre-processed and filtered spectra

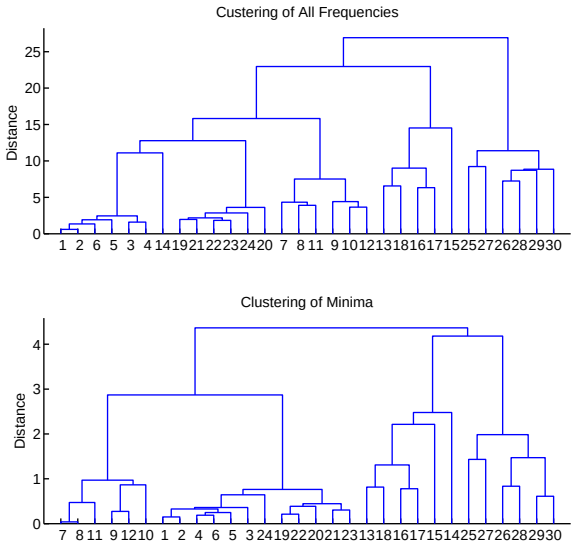


Fig. 4. Tree graph showing the parts of the clustering results. The distance between salicylic acid (samples 19-24) and acetyl salicylic acid (samples 1-6) is bigger in the top level picture, while the misclassification of sample 14 is suppressed at the bottom level and the clusters are clearer.

leads to clear classification results as can be seen in Fig. 4 on the top level dendrogram. The bottom level of Fig. 4 shows the clustering result of the distance measure proposed before. As can be seen in Fig 3 all compounds have clear minima, that were detected. Depending on the respective spectrum the number of minima varied between two and five. Therefore for each two spectra a maximum of ten frequencies was taken to calculate the distance. In comparison to that, for the usual distance calculation about 300 frequencies per spectrum are used. The clustering result were similarly good.

B. Imaging Data

The test data consists in the hyper-spectral image of an envelope containing chemicals. The goal of image analysis is to detect the chemicals within the image. First analysis is carried out in the time domain. The features displayed in Fig. 5 are used. They are normalized beforehand with respect to their maximal and minimal values to allow comparability.

On these two features the clustering is performed. The first visualization in Fig. 6 shows a segmentation in only two clusters. In the corresponding dendrogram those two groups correspond to the first branching. The branch containing the blue pixels of the image is set to being background and discarded for further analysis. A subsequent finer clustering is displayed in Fig. 7. From this clustering result the interesting regions are chosen by the user and used for the same process of spectral analysis as was performed on the chemical compounds. The chosen regions correspond to less than a quarter of the original pixels hence reduce the computational costs of the spectral analysis. As shown above preprocessing

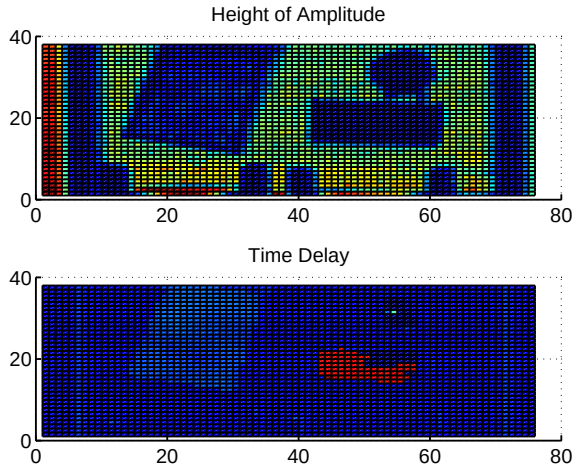


Fig. 5. A letter containing different materials. Materials with high absobance as well as chemical compounds. The plots display the different values of the main peak and the time delay.

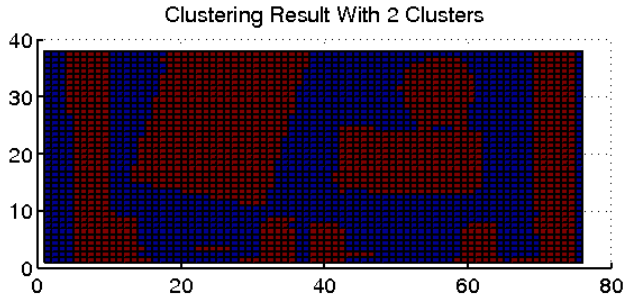


Fig. 6. A first clustering of the letter in two categories. The blue cluster represents the background and therefore the information within is discarded for further clustering.

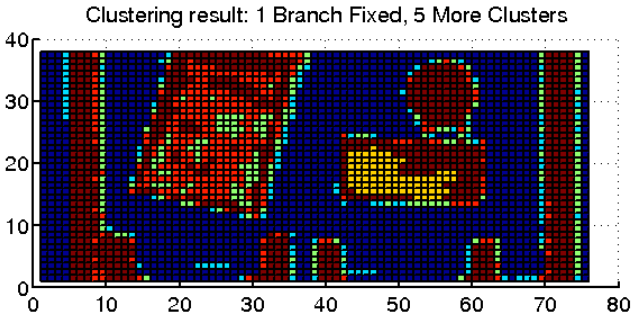


Fig. 7. A clustering with five clusters in the foreground.

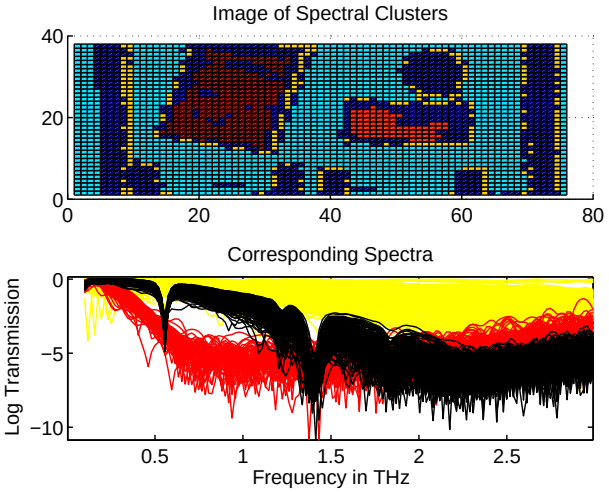


Fig. 8. Result of clutering in the frequency domain. Dark red corresponds to black, yellow and red pixel and spectra correspond to each other.

and Filtering were necessary to yield good clustering results. The chosen distance measure was Euclidean distance of all frequencies. The minima-distance worked well on parts of the data and performed considerably faster.But as can be seen in Fig. 8 some spectra have no chemical content and therefore do not contain clear minima. Therefore, some changes need to be made to yield good clustering results on imaging data with this measure. The result of the transmission clustering on all frequencies is shown in Fig. 8. In the lower part the spectra corresponding to the respective cluster are displayed. It can be seen that spectrally different compounds were automatically separated by the method. Only in the bright red cluster which has high absorbance there should some further analysis be done, as it can not clearly be determined if there are spectrally interesting pixels here.

IV. Conclusion and Further Work

In this paper it is shown that clustering algorithms are well applicable on THz data. With prior pre-processing and filtering, spectra of chemical compounds can be organized automatically. Especially in applications where the number of samples has high volume, clustering is a useful tool. The pre-processing steps used here were windowing functions, spectral density evaluation and filtering methods. All of these steps improved the clustering result. The two compared filtering methods performed comparably well. Savitzky-Golay filtering was chosen here because of its computational efficiency. Wavelet shrinkage has further advantages that were not taken into consideration yet. Those are for example possibilities of dimension reduction and better peak preservation which can be more relevant for other compounds. The proposed feature selection method can be taken to further improve the clustering itself by using incremental methods that make the algorithms more stable and computationally efficient [10], [11]. Clustering is particularly interesting in hyper-spectral THz imaging, as the amount of measured spectra is naturally higher. Here two different clustering steps were carried out. Both the clustering of the two-dimensional time domain feature space and the clustering of the frequency domain feature space led to an enhanced interpretability. It was shown that clustering provides useful possibilities for hyper-spectral image interpre-

tation which is relevant for example in applications in the security sector. The before described distance measure should be adapted to be applicable to hyper-spectral data to improve computational efficiency.

ACKNOWLEDGMENT

All data is measured by the Fraunhofer Institute for Physical Measurement Techniques (IPM), Kaiserslautern, Germany.

REFERENCES

[1] T. Warren Liao. Clustering of time series data - a survey. *Pattern Recognition*, 38:1857 – 1874, 2005.

[2] C.A. Schmuttenmaer. Exploring dynamics in the far-infrared with terahertz spectroscopy. *Chemical Reviews*, 104(4):1759–1780, 2004.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, September 1999.

[4] GN Lance and WT Williams. A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems. *The Computer Journal*, 9(4):373, 1967.

[5] P. Berkhin. *Grouping Multidimensional Data*, chapter A Survey of Clustering Data Mining Techniques, pages 25–71. Springer Berlin Heidelberg, 2006.

[6] J.G. Proakis and D.G. Manolakis. *Digital signal processing: principles, algorithms, and applications*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1996.

[7] M. Golay and A. Savitzky. Smoothing and differentiation of data by simplified least square procedures. *Analytical Chemistry*, 36:1627–1639, 1964.

[8] F. Ehrentreich, S.G. Nikolov, M. Wolkenstein, and H. Hutter. The wavelet transform: A new preprocessing method for peak recognition of infrared spectra. *Microchimica Acta*, 128(3):241–250, 1998.

[9] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[10] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data clustering method for very large databases. *SIGMOD Rec.*, 25(2):103–114, 1996.

[11] S. Guha, R. Rastogi, and K. Shim. Cure: an efficient clustering algorithm for large databases. *Information Systems*, 26(1):35–58, 2001.

Evolving Vector Quantization for Classification of On-Line Data Streams *

Edwin Lughofer
Department of Knowledge-based Mathematical Systems, Johannes Kepler University
A-4040 Linz, Austria, edwin.lughofer@jku.at

Abstract

In this talk, we present a new clustering-based classification technique (*eVQ-Class*), which is able to adapt old clusters and to evolve new ones on-line with new incoming data samples. It extends the conventional learning vector quantization approach, which is a kind of supervised version of original vector quantization, in mainly three points: 1.) it is able to evolve new clusters on demand by comparing new incoming samples with already generated clusters, 2.) it includes the label information in the training process by introducing a hit matrix and extending the feature space and 3.) it comes with a new weighted classification strategy. The novel approach will be evaluated based on high-dimensional feature data sets extracted from images recorded on-line in order to perform on-line quality control in a production process by classifying images into 'good' and 'bad' ones. The evaluation includes a comparison with well-known batch (trained and re-trained) classification techniques.

* This work was funded by the EC under grant no. 016429, project DynaVis and the Upper Austrian Technology and Research Promotion. It reflects only the authors' views.

High resolution microarray image analysis

L. Muresan[1] *, E. P. Klement[1], J. Jacak[2], G. Schütz [2], J. Hesse [3]

[1]Department of Knowledge-based Mathematical Systems, J.Kepler University, Linz, Austria
[2] Institute of Biophysics, J.Kepler University, Linz, Austria
[3]Upper Austrian Research, Linz, Austria

February 20, 2009

Abstract

Classical microarray analysis works at resolutions of two to 20 micrometers and quantifies the abundance of target molecules by determining average pixel intensities inside the microarray spots. Information on the molecular properties like brightness and position are not accessible. Precise quantification of e.g. hybridized cDNA molecules, however, can only be done when either the brightness is known or the individual molecules are visualized.

In (Hesse *et al.*, 2006) a new, high resolution approach is described, that enables the detection of individual molecules bound to microarray surfaces. Since high resolution makes the detection of single molecules possible, we propose the concentration of single molecules in each spot as a new measure of hybridization, instead average pixel intensity as used for low-resolution methods. However, the high resolution approach needs new algorithms in order to handle the three main tasks in microarray image processing: spot localization, foreground-background segmentation and computation of summary statistics that characterize the spot.

We have developed an approach for the analysis of high-resolution microarray images. It consists first of a single molecule detection step, based on undecimated wavelet transforms, and second, on separation of specific signal from unspecific background (corresponding to the segmentation step in the classical microarray analysis). Evaluation of the detection method on simulated images yielded a concentration range of 0.001 to 0.5 molecules per square micron and a minimum signal to noise ratio of 15 for reliable quantification (error of false negatives below 15%). Separation of foreground/background works if foreground exceeds background by a factor of 2. The method has also been used for analyzing real images and the results are in good agreement with the evaluation based on visual inspection.

1 Introduction

Microarray technology is used in medical diagnostics and basic research for analyzing the global transcriptional state of biological samples. The massively parallel detection approach allows the determination of several thousand expression levels in a single experiment. Novel technologies for sample preparation like Fluorescence Activated Cell Sorting and Laser Capture Microdissection allow to isolate small subpopulations of cells and enable researchers to investigate heterogeneities within their samples. For the global expression analysis of such small samples standard low-resolution methods require time consuming and possibly distorting (Nygaard *et al.*, 2005) pre-amplification steps. Recent developments in readout- (Hesse *et al.*, 2004) and platform/array-technology (Hesse *et al.*, 2006; Sonnleitner *et al.*, 2005), dramatically expand the range of directly accessible concentrations by increasing detection efficiency and the resolution to the optical diffraction limit. The images generated via this technique have sizes of typically a few GigaPixels and require novel image analysis approaches.

*Corresponding author: L.Muresan, email: leila.muresan@jku.at.

In general, microarray technology is based on specific binding of fluorescent-tagged target molecules on different locations of the array and the subsequent determination of target molecule abundance by measuring fluorescence on the respective area. Classical methods use the pixel intensity values inside the pre-determined spot regions of the microarray scans, which is an indirect measure for the presence of hybridized molecules. In subsequent analysis tasks, the location of the spot pattern is roughly identified, the sub-images are segmented into signal and background, and the information contained in the pixel intensities is summarized via appropriate statistics. Some further steps, typically background subtraction and normalization, are intended to remove all non-biological variation of the data. Several overviews of the classical microarray image analysis are available (e.g. Bajcsy (2006); Yang *et al.* (2001, 2002)).

In the case of microarrays with single molecule sensitivity, different signal detection strategies and count statistics have to be used. Each rectangular subimage of the microarray, obtained after the gridding step, contains the spot location (foreground) surrounded by a background region. Figure 6 shows images of the a simulated spot at diffraction limited resolution (200nm, left) and the same image downsampled to the scale of the classical microarray techniques (4μm middle). Due to the background noise and the fluctuations of the single molecule signal intensities in the downsampled image the foreground/background contrast is low, making the segmentation difficult. The analysis of the high resolution images, consisting of the detection of single cDNA molecules and the estimation of their concentration inside the spot, make the object of this work. The detection of single molecules is based on sparsity-adaptive wavelet thresholding, applied after a variance stabilization step. The estimation of the abundance of single molecules is performed on the detection results, and separates specific hybridization from clutter.

The approach was validated by analyzing on one hand simulated data with known ground truth and on the other hand, real microarray images, where the results were checked by visual inspection.

2 Model

Numerous models of the microarray signal were proposed Balagurunathan *et al.* (2004); Angulo (2008); Chudin *et al.* (2006); Korn *et al.* (2004); Li *et al.* (2005). They include several aspect of the acquired data such as image intensity and spot shapes.

However, having access to single molecule resolution and following (Chudin *et al.*, 2006), we consider a pixelwise compound Poisson process to model the intensity of a spot.

Given a rate $\lambda > 0$ and an arbitrary distribution Q, a *compound Poisson process* $CP(\lambda, Q)$ has the distribution of the random sum: $Z = \sum_{i=1}^{N} X_i$,where $N \sim \text{Poisson}(\lambda)$ and X_i are independent and identically distributed random variables with distribution Q, also independent of N. The tail behaviour of $CP(\lambda, Q)$ is inherited from Q, the expectation and variance are given by: $E(Z) = \lambda E(Q)$ and $\text{Var}(Z) = \lambda E(Q^2)$. In our case, the intensity of pixel i in the low-resolution microarray spot image is obtained from the high resolution data as: $Z_i = B_i + Y_i = B_i + \sum_{k=1}^{N_i} D_k$, where B_i represents background fluctuation and Y_i is a compound Poisson process, with N_i the number of single molecules inside the area corresponding to pixel i. Finally, D_k is the intensity of single molecules ($D_k \sim \text{Poisson}(\mu)$). Thus the technology of high resolution microarrays offers access to previously hidden information: instead of analyzing the low-resolution pixel values Z_i, the inference is based on N_i the number of single molecules in the image. The measure of hybridization is in our case the concentration of single molecules inside the spot of interest. The knowledge of the values of B_i and D_k has little relevance.

It is essential in microarray analysis the identification true signal and the control of the unspecific intensity variation.

We have identified the following shortcomings, which make the classical analysis less appropriate than the single molecule one:

1. Correct background estimation is difficult in low resolution microarray images

2. In case of low concentrations, due to low signal-to-noise (SNR) ratio and in the presence of artifacts low resolution microarrays cannot discriminate between signal and background (and the spots are rejected from analysis).

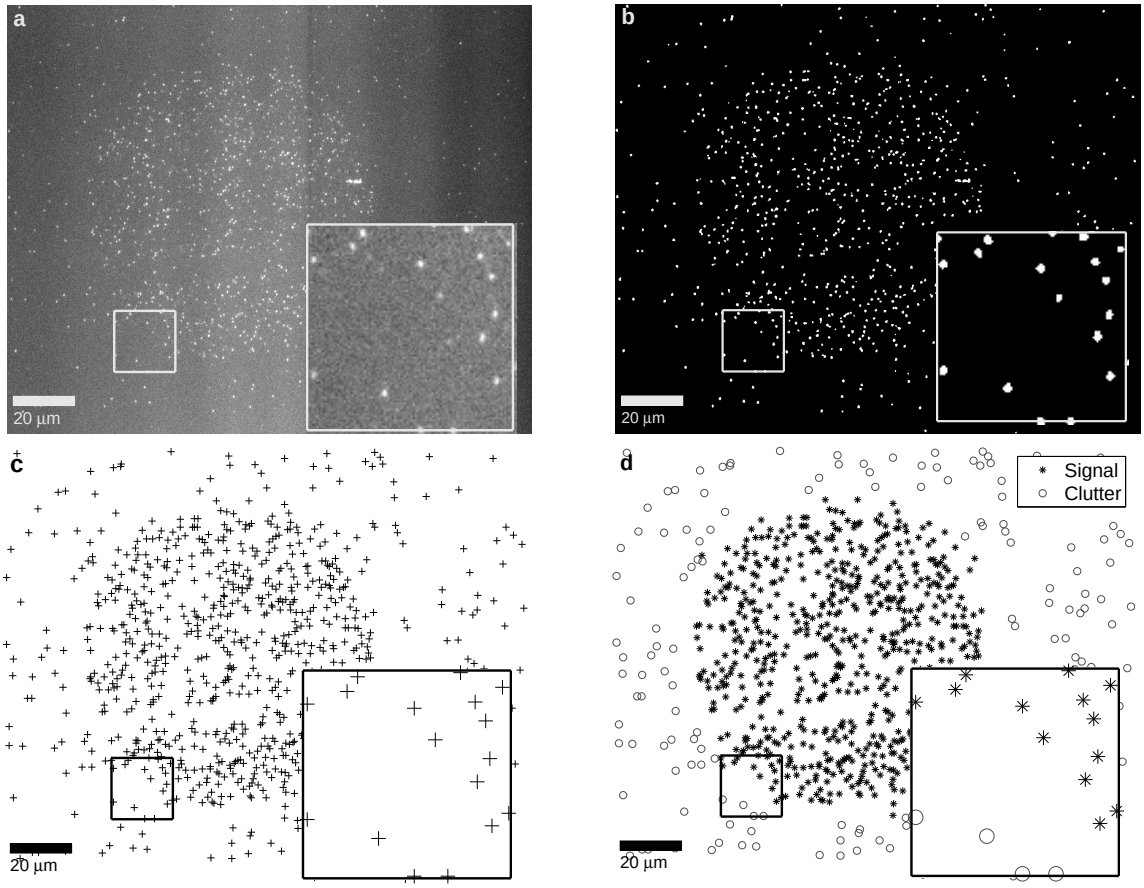


Figure 1: Analysis of a spot in a high-resolution microarray image. (a) Original image, bright features correspond to molecules bound to the chip. (b) Detection after undecimated wavelet thresholding. (c) Selection of single molecule locations (local maxima on denoised image inside the detection support in (b)), (d) Separation of hybridization signal from clutter.

3. The probability density function of a single molecule's photon intensity distorts the analysis results
4. The variability of the number of fluorophores per molecule is not modeled (usually requires dye swap normalization).

Besides offering a way to analyze very low concentration samples, the high resolution technique removes bias due to background heterogeneity and removes the necessity of the normalization step and dye swap.

3 Methods

We describe here an original framework to measure hybridization on high-resolution microarray data. Our approach relies on two independent steps. First we present a per formant method to detect blobs associated to singles molecules. Second, we introduce a concentration estimation approach based on spatial statistics.

3.1 Detection of single molecules

High resolution imaging of microarrays provides very large data sets. Discrete Wavelet transform involving filter banks are a reasonable approach for fast analysis of such data. Moreover, wavelet based approaches

have shown interesting properties in the detection of small bright features (Starck *et al.*, 1998; Olivo-Marin, 2002).

3.1.1 Isotropic undecimated wavelet transform (IUWT)

The wavelet transform is based on dilations and translations of a "father" and "mother" wavelet: $\phi_{jk}(t) = 2^{-j/2}\phi(2^{-j}t - k)$ and $\psi_{jk}(t) = 2^{-j/2}\psi(2^{-j}t - k)$ for $j, k \in Z$. The family

$$\left\{\phi_{Jk}, k = 0, 1, \ldots, 2^J - 1; \psi_{jk}, j \geq J, k = 0, 1, \ldots, 2^j - 1\right\}$$

forms an orthonormal basis of $L^2([0,1])$. Any function $f(t) \in L^2([0,1])$ can be arbitrarily well approximated by a wavelet series:

$$f(t) = \sum_k a_{Jk}\phi_{Jk}(t) + \sum_{j=-\infty}^{J}\sum_k d_{jk}\psi_{jk}(t),$$

where

$$a_{jk} = \langle f, \phi_{jk}\rangle, k = 0, 1, \ldots, 2^J - 1 \quad (1)$$

$$d_{jk} = \langle f, \psi_{jk}\rangle, \ j \geq J, k = 0, 1, \ldots, 2^j - 1 \quad (2)$$

represent the approximations and detail coefficients, respectively.

The functions ϕ and ψ fulfill the dilation equations (see Mallat (1999)):

$$\phi(x/2) = \sqrt{2}\sum_k h_k\phi(x-k), \ \psi(y/2) = \sqrt{2}\sum_k g_k\phi(x-k) \quad (3)$$

with h_k a discrete low-pass and g_k a discrete band-pass filter (followed by down-sampling), the approximation and detail coefficients can be computed recursively:

$$a_{j+1,k} = \sum_i h_{i-2k}a_{ji}, \ d_{j+1,k} = \sum_i g_{i-2k}a_{ji}. \quad (4)$$

Note that the described wavelet transform is anisotropic 1D and not translation invariant. However, these two properties are essential to a good detection scheme. We propose then to consider the isotropic undecimated wavelet transform (IUWT). The "à trous" scheme is thus used (Starck *et al.*, 1998) and wavelet coefficients are now computed over the entire grid as:

$$a_{jk} = \langle f, \phi_{jk}\rangle, \qquad \phi_{jk} = 2^{-j/2}\phi(2^{-j}(t-k)) \quad (5)$$

$$d_{jk} = \langle f, \psi_{jk}\rangle, \qquad \psi_{jk} = 2^{-j/2}\psi(2^{-j}(t-k)). \quad (6)$$

The recursive computation of the dyadic wavelet transform becomes:

$$a_{j+1,k} = a_{jk} * \tilde{h}_{jk}, \quad d_{j+1,k} = a_{jk} * \tilde{g}_{jk}. \quad (7)$$

where $\tilde{h}_{jk}$ ($\tilde{g}_{jk}$, respectively) is obtained by inserting $2^j - 1$ zeros between each sample of h_k (g_k).

In order to preserve isotropy the filters h and g and the father and mother function ϕ and ψ have to be nearly isotropic. A popular choice is based on the the B_3 spline scaling function, $h_k = [1/16,\ 1/4,\ 3/8,\ 1/4,\ 1/16]$, and for the $2D$ case a separable filter $h_{(k,l)} = h_k h_l$ and $g_{(k,l)} = \delta_{k,l} - h_{(k,l)}$, where $\delta_{k,l} = 1$ if $(k,l) = (0,0)$ and 0 otherwise.

The wavelet detail coefficients are given by: $d_{j+1,(k,l)} = a_{j,(k,l)} - a_{j+1,(k,l)}$ and the reconstruction is the sum of all details and the coarsest approximation:

$$f_{(k,l)} = a_{J,(k,l)} + \sum_{j=0}^{J} d_{j+1,(k,l)}. \quad (8)$$

When there is no confusion, a single index will be used to denote the $2D$ index (k,l). The first index, j, denotes the scale. Further details on IUWT can be found in (Starck *et al.*, 2007).

3.1.2 Thresholding based on False Discovery Rate (FDR)

The wavelet transform provides a sparse representation of signals as the number of significant coefficients is small. The remaining coefficients of low amplitude can then be considered as noise and eliminated via thresholding. Hard thresholding of wavelet coefficients d_{jk} can be written as

$$d^*_{jk} = \begin{cases} \hat{d}_{jk}, & \left|\hat{d}_{jk}\right| \geq T \\ 0, & \left|\hat{d}_{jk}\right| < T. \end{cases} \tag{9}$$

Since signals produce significant wavelet coefficients, correlated across wavelet planes, while noise is supposed to be uncorrelated, a pixel i is considered signal if all its corresponding wavelet coefficients $d_{ji}, j = 1, \ldots, J$ are exceeding the threshold T, e.g. $T = c \cdot \hat{\sigma_j}$, where $\hat{\sigma_k}$ is the robust estimate of the noise variance at scale k and c is a constant. As estimate of the noise variance Donoho and Johnstone (1994) propose: $\hat{\sigma_j} = \text{median}\left|d_{ji} - \text{median}\left(d_{ji}\right)\right| / 0.674$ and c is appropriately chosen, e.g $c = 3$.

The difficulty of the detection task lies in the fact that it has to be robust for a whole range of single molecule concentrations. The unknown concentration of single molecules in the image influences the sparsity of the signal, and implicitly the value of the parameter T that has to be chosen in order to obtain a correct detection. Therefore the detection method has to be driven by the (unknown) sparsity of the data. Some recent thresholding algorithms are sparsity adaptive as for instance Stein Unbiased Risk Estimator (SURE), FDR (Benjamini and Hochberg, 1995; Abramovich and Benjamini, 1996) and empirical Bayes methods (Johnstone and Silverman, 2004, 2005).

The wavelet coefficient thresholding approach reformulated from a multiple hypothesis testing point of view attaches to each wavelet coefficient of the true, unknown function f a ('no-signal') hypothesis $H_{jk} : d_{jk} = 0$. Ideally, only the signal coefficients should be kept.

The False Discovery Rate is defined as the expectation of the proportion of erroneously kept coefficients among all the coefficients kept in the representation.

Applying the Benjamini-Hochberg method as described in (Abramovich and Benjamini, 1996) one maximizes the number of kept coefficients controlling meanwhile the FDR to a predefined level q. The algorithm consists of the following steps:

1. For each $\hat{d}_{jk}$ calculate the two-sided p-value:

$$p_{jk} = 2\left(1 - \Phi\left(\left|\hat{d}_{jk}\right| / \sigma\right)\right)$$

2. Order ascendingly the computed p_{jk}-s, $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$

3. Find the largest i such that $p_{(i)} \leq (i/m)q$ and denote it i_0. Compute $\lambda_{i_0}(j) = \sigma\Phi^{-1}(1 - p_{(i_0)}/2)$.

4. Threshold all coefficients at level $\lambda_{i_0}(j)$.

Again the significant pixels are those that have non-zero coefficients in all the J detail levels (except the finest, which usually is contaminated by noise). The binary image obtained from the J detail levels, $B = \Pi_{j=1}^{J}\left\{d_{jk} > \lambda_{i_0}(j)\right\}$ is an indicator image for the support of the detected single molecules.

A denoised image is additionally obtained after applying the reconstruction step 8 with thresholded detail coefficients. The wavelet detection algorithm has a certain "resolution", two molecules that are spatially close together will be detected as one. In order to correct for the underestimation of the number of molecules, the binary image obtained after the detection step is combined with the denoised image such that all the local maxima of the denoised image inside the support of the binary mask are considered distinct single molecules (see fig. 1, c).

3.1.3 Variance stabilization

Wavelet methods are typically designed for additive Gaussian noise: $X_i = \mu_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma)$. However low intensities (small photon counts) collected by the sensor are not well modeled by Gaussian noise. A combination of Poisson (shot-noise) and Gaussian noise is more appropriate to describe photon count variations and read-out noise. The main difference is the heteroskedasticity of the new model (the variance of the noise depends on the signal).

In order to take into account the characteristics of the noise, variance stabilizing transforms are applied prior to wavelet detection to the input image which transform the heteroskedastic noise into Gaussian noise of variance approximately equal to one. In case of a Poisson noise model (suitable to describe the photon count model) the well known Anscombe transform can be used: $t(X_i) = 2\sqrt{X_i + 3/8}$ (it underestimates the intensity for values under 30). Modeling both the photon count noise as well as the read-out noise, one obtains the mixed Poisson-Gaussian image model (Starck *et al.*, 1998), $X_i = \alpha \cdot N_i + \epsilon_i$, where $\alpha > 0$ represents the gain of the detector, $N_i \sim Poi(\mu_i)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma)$ that can be stabilized via the generalized Anscombe transform (GAT): $t_G(x) = \sqrt{\alpha x + \frac{3}{8}\alpha^2 + \sigma^2 - \alpha\mu}$. The parameters α, μ, σ are determined from the image itself via robust fitting as described in (Boulanger *et al.*, 2008).

3.2 Foreground/background separation and estimation of single molecule concentration

Not all the peaks detected in the subimage belong to the spot of interest (see fig 1, c). The background might be heavily contaminated by unspecifically bound signal, impurities, etc. (clutter), which unaccounted for could seriously distort the hybridization results. Therefore, peaks detected in the subimages have to assigned either to foreground or background. In the concentration estimation step, we thus model both the spot and the background concentration.

In order to distinguish between peaks within the spot, representing true hybridization signal and those representing clutter a spatial mixture model is used. A similar approach used for segmentation of classical microarrays, but based on Gaussian mixture models for pixel intensity values, was described in Blekas *et al.* (2005), where the mixtures had two (signal/background) or three (signal/background/artifacts) components.

The peak locations obtained after the wavelet transform correspond either to peaks situated in the spot or to peaks in the background. Assuming that in case of strong hybridization there are much more peaks inside the spot, we shall discriminate between foreground and background via the concentration of the peaks in the two regions. The model we adopt is that of mixture of two Poisson processes with piecewise constant intensities λ_1 and λ_2(for foreground and background, respectively).

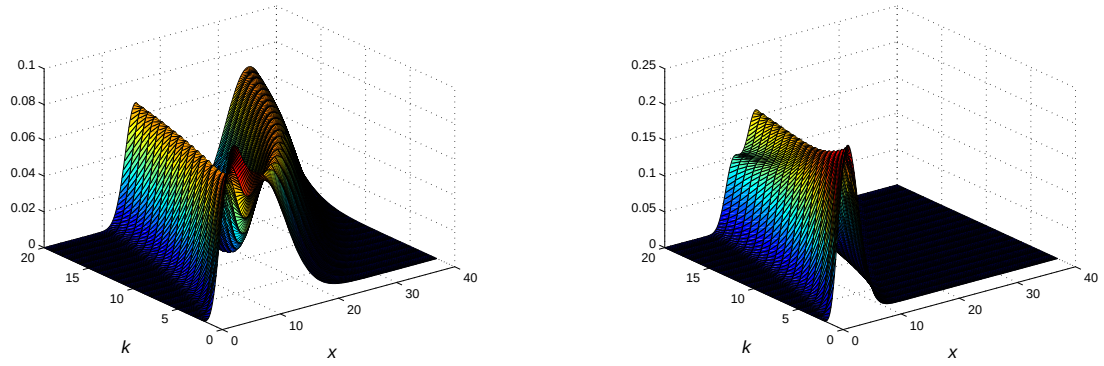
3.2.1 Estimation of concentrations based on method of moments (MOM)

A first approach for concentration estimation is to consider the count of the detected peaks inside non-overlapping, systematic quadrats, covering the subimage: y_i, where i is an index over the lattice structure. The counts are modeled as a mixture of two Poisson distributions, with constant rates λ_1, λ_2 (expressed in counts per quadrat): $p(y_i|\lambda_1, \lambda_2, \eta_1) = \eta_1 \mathcal{P}(\lambda_1) + (1 - \eta_1)\mathcal{P}(\lambda_2)$, where η_1 denotes the weight of the first component. This simple model doesn't account for correlations between neighbouring quadrats.

The three parameters λ_1, λ_2 and η_1 are determined via the method of moments for Poisson mixtures distribution discussed by Everitt, as described in (Frühwirth-Schnatter, 2006).

Although inefficient compared to other estimators, this simple method offers a closed form solution in the case of mixture of two Poisson distributions, which is crucial for the speed of the analysis for such a large quantity of data.

Let Y_i be the random variables representing the counts in the quadrat i and y_i the measured value of these variables. The first three factorial moments $\mathrm{E}(H_j(Y)|\lambda_1, \lambda_2, \eta_1)$ of the random variable Y, $H_j(Y) = \frac{Y!}{(Y-j)!}$, $j = 1, 2, 3$ are matched with empirical moments obtained from y_i, $\bar{H}_j = \frac{1}{N}\sum_{i=1}^{N} H_j(\mathbf{y_i})$.



(a) Easily separable components of the mixture $\lambda_1 = 0.025, \lambda_2 = 0.005$

(b) Mixture components difficult to separate $\lambda_1 = 0.065, \lambda_2 = 0.045$

Figure 2: Pseudo-surface representing the density functions of kNN distance $D_k^{\frac{1}{2}}(\lambda_1, \lambda_2)$ for the spatial Poisson mixture with concentrations λ_1 inside and λ_2 outside the spot, respectively. . To each value k on the kNN axis, corresponds a density function curve.

Since $\mathrm{E}(H_j(Y)|\lambda_k) = \lambda_k$ the equation system for λ_1, λ_2 and η_1 becomes:

$$\begin{aligned}\eta_1(\lambda_1 - \lambda_2) + \lambda_2 &= v_1\\ \eta_1(\lambda_1^2 - \lambda_2^2) + \lambda_2^2 &= v_2\\ \eta_1(\lambda_1^3 - \lambda_2^3) + \lambda_2^3 &= v_3\end{aligned}$$

with $v_j = \frac{1}{N}\sum_{i:y_i \geq j} y_i(y_i - 1)\ldots(y_i - (j-1))$.

We tested several quadrat sizes, but the results obtained on real images were robust for quadrats of size 20×20 pixels and above. However further study is necessary to select the optimal quadrat size.

3.2.2 Concentration estimation based on Expectation-Maximization approach

As a second approach we adopt the method of Byers and Raftery, used in minefield detection (Byers and Raftery, 1998). The location of the detected peaks are treated as a mixture of two spatial Poisson processes with different rates for foreground and background.

In the case of a single spatial Poisson processes with constant rate λ the distribution D_k of the distance from a point of the Poisson process to it's k-th nearest neighbor (kNN) can be written as

$$P(D_k \geq x) = \sum_{q=0}^{k-1} \frac{e^{-\lambda\pi x^2}\left(\lambda\pi x^2\right)^q}{q!} = 1 - F_{D_k}(x). \tag{10}$$

This leads to the density function:

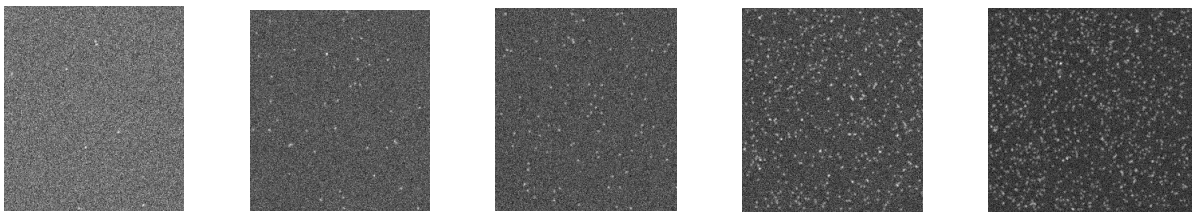
$$f_{D_k}(x) = \frac{\mathrm{d}F_{D_k}(x)}{\mathrm{d}x} = \frac{e^{-\lambda\pi x^2} 2\left(\lambda\pi\right)^k x^{2k-1}}{(k-1)!}$$

meaning that the D_k^2 follows a transformed Gamma distribution, $D_k^2 \sim \Gamma(k, \lambda\pi)$. Here λ is measured in counts of single molecules per pixels squared.

The maximum likelihood estimate of the rate of the Poisson process is:

$$\hat{\lambda} = \frac{k}{\pi\sum_{i=1}^n d_i^2} \tag{11}$$

where $d_i, i = 1, \ldots, n$ are the realizations of the k-th nearest neighbour distances.



(a) $N = 10$ (b) $N = 50$ (c) $N = 100$ (d) $N = 500$ (e) $N = 1000$

Figure 3: Detection results for FDR thresholding, at different concentrations and SNR levels. A set of simulated images is shown in (a)-(e) for different concentrations: N represents the number of peaks in the 512×512 pixel image (corresponding to $102.4\mu m \times 102.4\mu m$). SNR = 5.02 (additional Gaussian noise with $\sigma = 2.2$). The images are scaled for better visibility. The same pixel intensity scale is used for the five images.

In the case of a mixture of two Poisson processes with two intensity rates λ_1 and λ_2, the model for D_k can be written as: $D_k = D_k(\lambda_1, \lambda_1) \propto p\,\Gamma^{\frac{1}{2}}(k, \lambda_1\pi) + (1-p)\Gamma^{\frac{1}{2}}(k, \lambda_2\pi)$.

As opposed to the method of moments' η, p represents a proportion of the samples D_k. The three unknown parameters that describe the distribution D_K: p, λ_1, λ_2, are estimated via the Expectation Maximization (EM) algorithm, together with the assignments to components ("missing data") $\delta_i \in \{0, 1\}$, where $\delta_i = 1$ if the i-th point belongs to the first component (signal), and $\delta_i = 0$ otherwise.

The expectation step is:

$$E\left(\hat{\delta}_i^{(t+1)}\right) = \frac{p^{(t)} f_{D_k}\left(d_i, \hat{\lambda}_1^{(t)}\right)}{p^{(t)} f_{D_k}\left(d_i, \hat{\lambda}_1^{(t)}\right) + (1-p^{(t)}) f_{D_k}\left(d_i, \hat{\lambda}_2^{(t)}\right)}$$

and the maximization:

$$\hat{\lambda}_1^{(t+1)} = \frac{k\sum_{i=1}^n \hat{\delta}_i^{(t+1)}}{\pi\sum_{i=1}^n d_i^2 \hat{\delta}_i^(t+1)}$$

$$\hat{\lambda}_2^{(t+1)} = \frac{k\sum_{i=1}^n \left(1-\hat{\delta}_i^{(t+1)}\right)}{\pi\sum_{i=1}^n d_i^2 \left(1-\hat{\delta}_i^{(t+1)}\right)}$$

$$p^{(t+1)} = \sum_{i=1}^n \frac{\hat{\delta}_i^{(t+1)}}{n}.$$

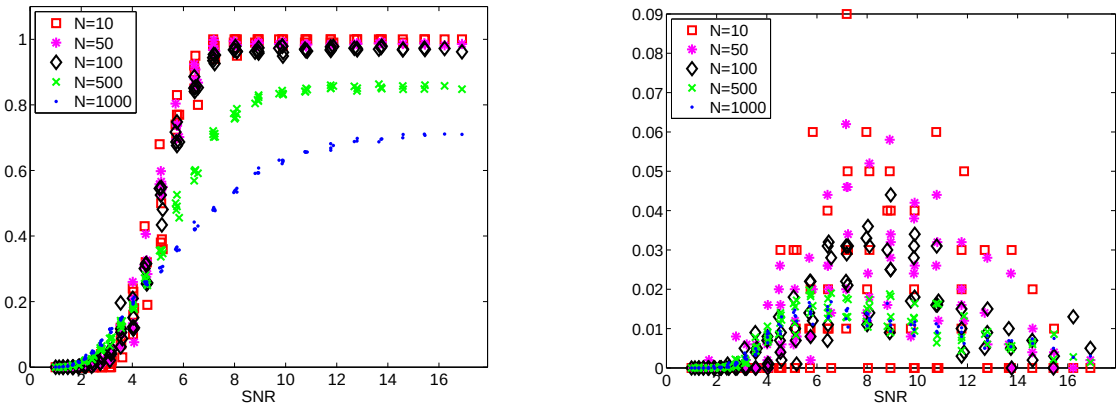
As initial values for the three parameters one can use the results obtained through the method of moments method. The bigger the difference between the signal and clutter concentration the easier is to separate the two components of the mixture. Also, as one can see in fig. 2 for high concentrations the task is more challenging than for lower concentrations. The figure 2 is a pseudo-surface representing the distribution function of the k-th nearest neighbour $D_k(\lambda_1, \lambda_2)$ in the spatial mixture model for each fixed value k on the axis kNN, and for fixed rates λ_1 and λ_2.

4 Results and discussion

The performance of the high resolution microarray image analysis was extensively tested on simulations as well as real images.

4.1 Evaluation of the detection method

The detection algorithm was tested on a set of simulation images with varied image quality parameters, as measured by the signal-to-noise ratio (SNR), as well as several molecule concentrations. Each image is of



(a) True positive ratios (5% Gaussian noise added)

(b) False positive ratios (5% Gaussian noise added)

Figure 4: The results of detection on the simulations are summarized in figures: (a) ratio of true positives and (b)ratio of false negatives with respect to the true number of simulated single molecules

dimension 512×512 pixels and contains $10, 50, 100, 500$ or 1000 randomly placed molecules.

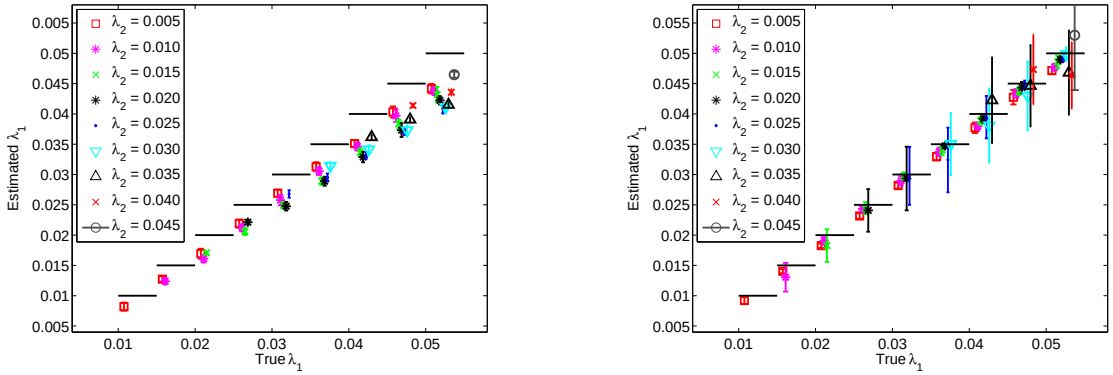
To a single molecule corresponds a diffraction limited spot, approximated by a two-dimensional Gaussian shape, with width s corresponding to the point spread function of the optical system (1.1 in our simulations). Both the constant background intensity and the peak intensity S were chosen on a logarithmic scale between 10 and 100. Noise is generated for each pixel as described in 3.1.3: the photon count noise was modeled by draws from Poisson distributions, and finally Gaussian noise is added to each pixel from $\mathcal{N}(0, \sigma)$, where σ was chosen as $0, 5, 10, 15$ and 20% of the maximum peak intensity. For this special case of Poisson-Gaussian model described , we use to following signal-to-noise (SNR) definition as in Murphy (2001): SNR $= S/\sqrt{B + \sigma^2}$, where S represents the single molecule intensity, B the (local) background of the image and σ the standard deviation of the read-out noise. (For simplicity S is the maximum intensity of the single molecule profile).

The SNR for our simulations is between 0.9 and 31.6.. For each set of parameters 10 images were generated and analyzed. One set of simulations is presented in figure 3 $(a - e)$.

The results are summarized in figure 4 (a and b). For SNR above 10 the detection for all five concentrations levels, with over 80% true positives for less than 500 molecules, but only above 60% for high concentrations ($N = 1000$). However, at high concentrations the spots can be analyzed also with conventional methods designed for low resolution microarrays. The SNR typical for our system, is usually at least 15, (at this level true positives exceed 85% of the number of simulated peaks for $N = 500$, and 70% for$N = 1000$, respectively). The rate of false positives is under 9% even for very low concentrations ($N = 10$) and substantially lower for $N > 100$. The detection performance was similar for simulation with the same SNR, independent of the weight of Poisson and Gaussian noise in the generation of the simualted image.

4.2 Evaluation of concentration estimation

The concentration estimation algorithms were tested on data representing the position of single molecules and clutter, respectively. The assumption is that signal (molecules' position) has a higher concentration than clutter. For each data set, two spatial Poisson processes are simulated: one of intensity λ_1 inside a disk of radius R (150 pixels in our case) and a second one, of intensity λ_2, independent of the first one, in a rectangle excluding this area. The values of λ_1 are increasing with a step of 0.005 from 0.01 to 0.05 peaks per pixel squared, while λ_2 varies from 0.005 to $\lambda_1 - 0.005$, for each value of λ_1. For each (λ_1, λ_2) parameter pair, ten data sets were generated. The results of the estimation of the signal concentration λ_1 is presented in figure 5. The results in the case of MOM estimation are strongly biased downward compared to the true value, while in the case of EM this bias is less strong. When the two concentration are close together one can see a stronger bias, due to the failure of separating the two components of the mixture.



(a) MOM (b) EM

Figure 5: Background/foreground separation for three different concentrations via Method of Moments and Expectation Maximization methods. The true λ values are represented as a stair-case function and for better visibility, the estimation results were slightly shifted on the abscissa.

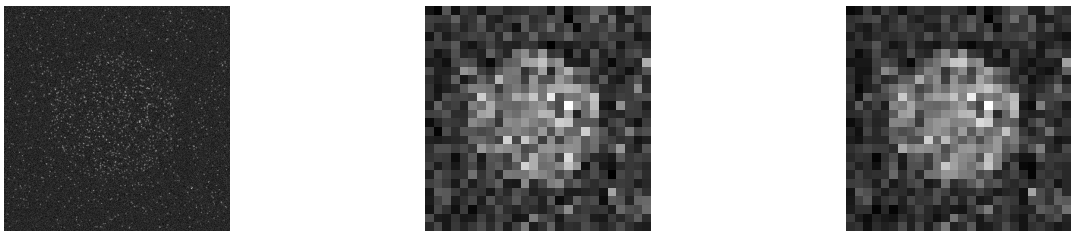


Figure 6: Correlations between the true and the estimated signal concentrations were computed for simulated images (left), for the downscaled ones (middle) and for the denoised and downscaled images (right).

4.3 Correlation tests

Finally we have compared how the results of our analysis correlate with the original concentrations, and compared these correlation results with those obtained via a classical technique on the downscaled images of the same data. The downscaled images corresponds to the resolution of commercially available systems. For this purpose, 60 images were generated with SNR between 2.85 and 31.6. In each image, single molecules were simulated with concentrations λ_1 between 0.001 and 0.011 peaks/pixels squared inside a disk of radius 150 pixels. The clutter outside this disk, λ_2 was varied between 0.001 and 0.09, such that $\lambda_2 \leq \lambda_1 + 0.02$. The images of the single molecules were generated as described in 4.1. As classical analysis we have chosen the maximum likelihood (MLE) approach to estimate the parameters of a Gaussian mixture with two components applied on the pixel intensities of the downscaled image (from each 20×20 pixels patch of simulated image we have generated by summation one pixel in the downscaled image). The MLE was applied directly to the downscaled data as well as to the downscaled result of the denoised images by wavelet thresholding.

As it is shown in figure 7 the single molecule analysis always shows a better correlation than the analysis of the downscaled images. The lowest correlation value for single molecule analysis, 0.837, is obtained for SNR = 4.81. For more typical SNR values, above 10, the correlation coefficient is higher than 0.935.

5 Conclusion

In this paper we have presented the analysis of microarray images at single molecule resolution. A model for microarray image formation was given, in order to explain the advantages of the single molecule technique. The analysis consists of estimating the two concentration of hybridization signal, which implies single molecule detection and signal clutter separation. We have tested the detection algorithm and the signal

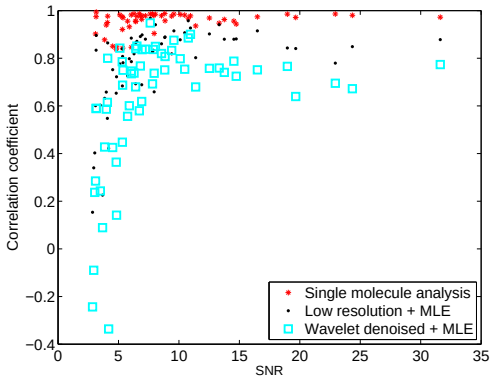


Figure 7: Correlations between the estimated and the true signal concentrations. The single molecule analysis always performs better than the analysis on the downscaled data (original and denoised via wavelet thresholding).

concentration based on MOM and EM methods on simulated images and we have shown that they perform well for the parameters typical for our system. Furthermore we have shown that our approach provides good correlation results for concentrations and SNR values where the low resolution based methods fail.

Funding

This work was supported by the GEN-AU program of the Austrian Federal Ministry of Education, Science and Culture, by the Austrian Research Fund (FWF Project L422-N20) and the European Commisson (FP6 Project Autoscreen).

References

Abramovich, F. and Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis*, **22**, 351–361.

Angulo, J. (2008). Polar modelling and segmentation of genomic microarray spots using mathematical morphology. *Image Anal. Stereol.*, **27**, 107–124.

Bajcsy, P. (2006). An overview of DNA microarray grid alignment and foreground separation approaches. *EURASIP Journ. on Applied Signal Processing*, **2006**, 1–13.

Balagurunathan, Y., Wang, N., Dougherty, E. R., Nguyen, D., Chen, Y., Bittner, M., Tent, J., and Carroll, R. (2004). Noise factor analysis for cdna microarrays. *Journal of Biomedical Optics*, **9**(4), 663–678.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**(1), 289–300.

Blekas, K., Galatsanos, N. P., Likas, A., and Lagaris, I. E. (2005). Mixture model analysis of DNA microarray images. *IEEE Transactions on Medical Imaging*, **24**, 901–909.

Boulanger, J., Sibarita, J.-B., Kervrann, C., and Bouthemy, P. (2008). Non-parametric regression for patch-based fluorescence microscopy image sequence denoising. In *Proc. IEEE Int. Symp. on Biomedical Imaging: from nano to macro (ISBI)*.

Byers, S. and Raftery, A. E. (1998). Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Staistical Association*, **93**, 577–584.

Chudin, E., Kruglyak, S., Baker, S., Oeser, S., Barker, D., and McDaniel, T. (2006). A model of technical variation of microarray signals. *Journal of Computational Biology*, **13**(4), 996–1003.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, **81**, 425–455.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer New York.

Hesse, J., Sonnleitner, M., Sonnleitner, A., Freudenthaler, G., Jacak, J., Höglinger, O., Schindler, H., and Schütz, G. (2004). Single-molecule reader for high-throughput bioanalysis. *Anal. Chem.*, **76**, 5960–5964.

Hesse, J., Jacak, J., Kasper, M., Regl, G., Eichberger, T., Winklmayr, M., Aberger, F., Sonnleitner, M., Schlapak, R., Howorka, S., Muresan, L., Frischauf, A., and Schütz, G. J. (2006). RNA expression profiling at the single molecule level. *Genome Research*, **16**, 1041–1045.

Johnstone, I. and Silverman, B. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics*, **32**(4), 1594–1649.

Johnstone, I. and Silverman, B. (2005). Empirical bayes selection of wavelet thresholds. *The Annals of Statistics*, **33**(4), 1700–1752.

Korn, E., Habermann, J., Upender, M., Ried, T., and McShane, L. (2004). Objective method of comparing dna microarray image analysis systems. *Bioimaging*, **36**(6), 960–967.

Li, Q., Fraley, C., Bumgarner, R., Yeung, K., and Raftery, A. (2005). Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics*, **21**(12), 2875–2882.

Mallat, S. (1999). *A wavelet tour of signal processing*. Academic Press.

Murphy, D. B. (2001). *Fundamentals of light microscopy and electronic imaging*. Willey-LISS.

Nygaard, V., Holden, M., Løland, A., Langaas, M., Myklebost, O., and Hovig, E. (2005). Limitations of mrna amplification from small-size cell samples. *BMC Genomics*, **6**, 147.

Olivo-Marin, J.-C. (2002). Extraction of spots in biological images using multiscale products. *Pattern Recognition*, **35**, 1989–1996.

Sonnleitner, M., Freudenthaler, G., Hesse, J., and Schütz, G. J. (2005). High-throughput scanning with single-molecule sensitivity. In *SPIE Proceedings*, pages 202–210.

Starck, J.-L., Murtagh, F., and Bijaoui, A. (1998). *Image and Data Analysis: The Multiscale Approach*. Cambridge University Press.

Starck, J. L., Fadili, J., and Murtagh, F. (2007). The undecimated wavelet decomposition and its reconstruction. *IEEE Trans. on Image Processing*, **16**(2).

Yang, Y. H., Buckley, M. J., and Speed, T. (2001). Analysis of cDNA microarray images. *Briefings in Bioinformatics*, **2**, 341–349.

Yang, Y. H., Buckley, J., Dudoit, S., and Speed, T. P. (2002). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, **11**, 108–136.