Centrum voor Wiskunde en Informatica

**REPORT**RAPPORT

PNA

Probability, Networks and Algorithms

*Probability, Networks and Algorithms*

# A Statistically Principled Approach to Histogram Segmentation

Greet Frederix, Eric J. Pauwels

**REPORT PNA-E0401 FEBRUARY 25, 2004**

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

# A Statistically Principled Approach to Histogram Segmentation

ABSTRACT

This paper outlines a statistically principled approach to clustering one dimensional data. Given a dataset, the idea is to fit a density function that is as simple as possible, but still compatible with the data. Simplicity is measured in terms of a standard smoothness functional. Data-compatibility is given a precise meaning in terms of distribution-free statistics based on the empirical distribution function. The main advantages of this approach are that (i) it involves a single decision-parameter which has a clear statistical interpretation, and (ii) there is no need to make a priori assumptions about the number or shape of the clusters.

*1998 ACM Computing Classification System:* I.5.3 ; I.4.6
*Keywords and Phrases:* Histogram segmentation, image segmentation, Kolmogorov-Smirnov statistic
*Note*: This work was partially supported European Commission's 5th Framework IST Project FOUNDIT, Project nr. IST-2000-28427

# A Statistically Principled Approach to Histogram Segmentation

Greet Frederix

*Hogeschool Limburg, Universitaire Campus, B-3590 Diepenbeek, Belgium*

Eric J. Pauwels

*CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands*

## Abstract

This paper outlines a statistically principled approach to clustering one dimensional data. Given a dataset, the idea is to fit a density function that is as simple as possible, but still compatible with the data. Simplicity is measured in terms of a standard smoothness functional. Data-compatibility is given a precise meaning in terms of distribution-free statistics based on the empirical distribution function.

The main advantages of this approach are that (i) it involves a single decision-parameter which has a clear statistical interpretation, and (ii) there is no need to make a priori assumptions about the number or shape of the clusters.

*Key words:* Clustering, cluster-validation, histogram segmentation, distribution-free statistics, Occam's Razor

## 1 Introduction

Clustering still is the mainstay of unsupervised learning and as such there is no shortage of methods and algorithms. Nevertheless, it remains a challenging problem as there is no general consensus on how the number and shape of clusters should be determined. For instance, fitting a mixture of Gaussians is a popular choice of methodology. However, there is no guarantee that this model is appropriate for the data at hand, e.g. the underlying distribution might be

*Email addresses:* greet.frederix@hogelimb.be (Greet Frederix), eric.pauwels@cwi.nl (Eric J. Pauwels ).

*URL:* http://www.cwi.nl/∼pauwels/research/ (Eric J. Pauwels ).

exponential. Moreover, even if Gaussians turn out to be a good choice, one still needs to take recourse to ad-hoc procedures to estimate the number of components in the mixture. There is therefore room for improvement and in this paper we propose a statistically principled approach to the problem for 1-dimensional data. Admittedly, this setting is rather restrictive, it is, however, not without merit: There are many situations in which histograms of 1-dimensional data are generated which need partitioning by grouping data based on local density-minima. In this context, the proposed method can be seen as a principled way to extract data-driven thresholds.

In essence, the method we propose is a computationally tractable version of Occam's Razor: Select the simplest density that is still compatible with the data. Compatibility is measured by statistically comparing the empirical distribution function for the data with the cumulative distribution of the proposed density. The amount of deviation between these two can be measured in precise probabilistic terms and a clear-cut quantitative decision criteria can be formulated.

The proposed solution is valid for 1-dimensional (numerical) data only as it hinges on the natural total ordering that exists on the real numbers. As such a total ordering is lacking for more-dimensional data, the algorithms detailed below cannot be implemented. Nevertheless, such data are amenable to cluster-methods that adhere, if not to the same algorithms, at least to the same philosophy. How to proceed in such a case will be expounded in an follow-up paper. For now however, we focus on the one-dimensional case.

## 2 Overview and Outline of Clustering Method

### *2.1 Brief overview of density estimation methods*

Due to lack of space, this overview is kept as concise as possible and intended only to elucidate the connection between our proposal and the extensive class of established standard approaches to the problem. For more information on alternative techniques for clustering and density estimation, we refer the reader to excellent texts such as [5,8].

A first class comprises the **parametric methods** of which Gaussian Mixture Models (GMM) are the best-known exponent. The latter approach performs superbly if and when the number of constituent Gaussian clusters is known, for then the EM algorithm [4] can be used to estimate the remaining number of parameters. However, additional (and often ad-hoc) criteria must be invoked to estimate the actual number of clusters.

**Non-parametric methods** constitute an alternative approach in which no *a priori* assumption about the underlying density is put forward. In *kernel density estimation,* the dataset is convolved by a kernel-function (again often a Gaussian) and the overall shape of the density is determined by the characteristic width of the kernel function. Now the problem is one of picking the appropriate kernel-width for which there are a number of theoretical results (e.g. Parzen windowing). However, they do depend on knowledge about the shape of the density, hence creating a recursion problem. Moreover, in many cases better results are obtained if the width of kernel-function is made location dependent. But this further complicates the parameter estimation problem.

*Spline smoothers* comprise another class of non-parametric density estimators. Most frequently, these appear under the guise of a penalized smoothing functional where for a set of observations $(x_i, y_i)$, one needs to construct the density $f$ that minimizes the functional:

$$\Psi_\lambda(f) = \int (f''(x))^2\,dx + \lambda \sum (y_i - f(x_i))^2. \tag{1}$$

Notice however, that this functional (and hence the solution) depends on the weight-factor $\lambda$ which only has a handwaving interpretation in terms of the relative importance of both penalty terms. As it turns out, the method that we propose in this paper is closely related to this penalized approach but exchanges the vagueness of the $\lambda$-parameter for an alternative with crisp probabilistic definition. In addition, there are some further subtle differences for a discussion of which we refer the interested reader to technical reports at [11].

### *2.2 Outline of Proposed Method*

In order to be as general as possible, we make minimal assumptions about the underlying density $f$. As mentioned in the introduction, there is no reason to restrict attention to a mixture of Gaussians. We will simply assume that $f$ has a square integrable derivative, so that the smoothness functional introduced below is well defined. Furthermore, once we have an estimate for $f$, the number of clusters is determined by identifying the different local maxima ("modes").

In essence, the method propounded in this paper is a computationally tractable version of Occam's Razor: For a given 1-dimensional data-set $x_1, x_2, \ldots, x_n$ we propose to construct the *simplest* density $f$ that is still *compatible with the data.*

(1) *Simplicity* is measured in terms of the standard smoothness functional $\Phi(f) = \int (f'(x))^2 dx$, which is an increasing function of the roughness of

the density.

(2) *Data-compatibility* of $f$ on the other hand is enforced by insisting that statistical tests should not be able to reject $f$ as a viable density for the observed data.

Whilst the interpretation of the first condition is straightforward, the second needs some further amplification. Basically, it insists that *if we assume* (as null-hypothesis) that $f$ is the *real* underlying data-density, then an appropriate statistical test based on the available sample $x_1, x_2, \ldots, x_n$, should not be able to reject this hypothesis (at a pre-defined significance level). As we do not want to restrict the class of densities to a parametric family, we choose the statistical test to be as general as possible. For that reason, we opt for general distribution-free tests such as Kolmogorov-Smirnov (KS) or Cramer-von Mises (CvM). The critical values for these statistics are independent of the true underlying distribution that is being tested and can therefore be computed in advance. In addition, these tests are based on the cumulative distribution function $F(x) = \int_{-\infty}^{x} f(u)du$ rather than on the density $f$ itself. This has a number of advantages: integration imparts better numerical stability, and the fact that $F$ is monotone increasing means that the problem can be translated into a spline optimization problem (see section 4).

Let us now take a closer look at the procedure we propose to estimate the density. Suppose we have a sample $x_1, x_2, \ldots, x_n$ from an unknown density $f$. First, we construct the empirical distribution function $F_n(x) = \#\{x_i \,|\, x_i \leq x\}/n$ ($F_n$ makes a $1/n$-jump at every observation $x_i$). Clearly, $F_n$ will be close to the unknown cumulative distribution function $F(x) = \int_{-\infty}^{x} f(u)du$, and appropriate distance functions $D_n = d(F, F_n)$ yield stochastic variables for which the probability density can be computed explicitly (in section 3 we will provide more details). In particular, it's possible to compute how likely it is that $D_n$ exceeds a predefined level $\delta$ and it turns out that $P(D_n > \delta)$ is a (rapidly) decreasing function of the difference $\delta$ since large deviations between $F$ and $F_n$ are exceptional. Next, we pick an acceptable level of statistical risk $\alpha$ (we will come back to what's considered acceptable in a minute). Since the probability-distribution of $D_n$ is known, one can compute for any given $0 < \alpha < 1$ the corresponding difference $\delta_\alpha$ such that $P(D_n > \delta_\alpha) = \alpha$. In words: if $F$ represents the correct underlying data-structure, then the probability that $D_n = d(F_n, F)$ *will exceed* $\delta_\alpha$ is at most $\alpha$. Hence, $\alpha$ corresponds to what in statistics is called a Type-I-error, i.e. rejecting the null-hypothesis when in fact it's correct.

Collecting the information above, the original problem can now been recast into a *constrained optimization problem:* Given data $x_1, x_2, \ldots, x_n$ construct

$F_n(x)$ and find $F$ which solves the constrained minimization problem

$$\begin{cases} \min \Psi(F) \quad \text{where} \quad \Psi(F) = \int (F''(x))^2 \, dx & (\textit{simplicity}) \\ \text{subject to} \qquad D_n = d(F, F_n) \leq \delta_\alpha & (\textit{data-compatibility}) \end{cases} \tag{2}$$

Once an optimal $F$ is found, the corresponding density $f = F'$ can be obtained and clusters identified by locating local maxima and minima.

**Overview of paper:** In order to make further progress in solving (2), we need to specify the distance function and its probability distribution. This is done in section 3 where the Kolmogorov-Smirnov and Cramer-von Mises statistics are discussed. Section 5 reformulates the minimization problem as a matrix optimization problem and presents the computational solution scheme. Finally, some experimental results are discussed in Section 6.

## 3 Distribution-Free Statistics based on the Empirical Distribution

In this section we focus on two distribution-free statistics, both measuring the deviation between the empirical distribution function $F_n$ and its underlying model-distribution $F$. Recall that a statistic is called *distribution-free* if its probability distribution does not depend on the distribution of the true underlying population. The distribution-free character of the Kolmogorov-Smirnov and Cramer-von Mises statistics is a consequence of the following elementary lemma.

**Lemma 1** *If $X$ is a stochastic variable with distribution $F$ and a continuous density $f$, then its image under $F$ is uniformly distributed on the unit-interval* [0, 1]*:*

$$U := F(X) \sim U(0,1) \qquad \textit{or again} \qquad P(F(X) \leq t) = t, \tag{3}$$

*for* $0 \leq t \leq 1$.

*In particular, any sample $X_1, \ldots, X_n$ drawn from $F$ is mapped under $F$ to an $U(0,1)$-sample: $U_i = F(X_i)$. Furthermore, the distribution function of the latter can be expressed in terms of the original as $H_n(t) = F_n(F^{-1}(t))$.*

### *3.1 The Kolmogorov-Smirnov statistic*

Our first candidate for $D_n = d(F_n, F)$ is the Kolmogorov-Smirnov statistic defined as the $L^\infty$-distance between the empirical and the proposed distribution:

$$d_{KS}(F_n, F) \equiv K_n := \sup_{x \in I\!R} |F_n(x) - F(x)|. \tag{4}$$

Invoking Lemma 1 we can make the substitution $t = F(x)$ and rewrite $K_n$ in terms that better elucidate the distribution-free character of the statistic, viz:

$$K_n = \sup_{0 \le t \le 1} |H_n(t) - t| \tag{5}$$

where as before, $H_n$ is the empirical distribution of a $U(0, 1)$-sample of size $n$. For every $\delta > 0$ one can explicitly compute the probability that $K_n$ exceeds the threshold $\delta$ (at least asymptotically for $n \longrightarrow \infty$, see eg. Durbin [6])

$$P(d_{KS}(F, F_n) > \delta) = Q_{KS}(\sqrt{n}\,\delta), \tag{6}$$

where for $x > 0$

$$Q_{KS}(x) = 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 x^2}. \tag{7}$$

### *3.2 The Cramér-von Mises statistic*

The original *Cramér-von Mises* statistic is defined as

$$d_{CvM}(F_n, F) \equiv W_n^2 := n \int_{I\!R} (F_n(x) - F(x))^2 dF(x). \tag{8}$$

Again, the distribution-free nature of this statistic is better explicified by the substitution $t = F(x)$:

$$W_n^2 = n \int_0^1 (H_n(t) - t)^2 \, dt. \tag{9}$$

As was the case for the Kolmogorov-Smirnov statistic (see (6)), it is possible to give an explicit expression for the $p$-value of the asymptotic statistic. Anderson and Darling [2] showed that $\lim_{n \to \infty} P(W_n^2 \le \delta) = P(W^2 \le \delta)$ equals

$$\frac{1}{\pi\sqrt{\delta}} \sum_{k=0}^{\infty} (-1)^k \binom{-1/2}{k} \sqrt{4k+1}\; e^{-\beta_k(\delta)} K_{1/4}(\beta_k(\delta)) \tag{10}$$

where $\beta_k(x) = (4k+1)^2/(16x)$ and $K_\nu(x)$ is the modified Bessel-function of the second kind (see [1] p. 376, # 9.6.23). The series expansion in (10) is rapidly converging so that a few terms suffice to give an sufficiently accurate value for the $p$-value.

Note that the $p$-values detailed above are asymptotic values, strictly speaking valid only when the sample-size $n$ tends to infinity. But simulation experiments show that for samples of size $n > 100$ these asymptotic values are quite accurate.

### *3.3 The choice of the threshold parameter $\alpha$*

We are now in a position to give a more detailed discussion of the choice of the $\alpha$-parameter. Its significance is most easily explained for the Kolmogorov-Smirnov statistic. Recall from (2) that we choose the threshold $\delta_\alpha$ such that

$$P(d_{KS}(F_n, F) \leq \delta_\alpha) = 1 - \alpha,$$

or equivalently:

$$P(\forall x : F_n(x) - \delta_\alpha \leq F(x) \leq F_n(x) + \delta_\alpha) = 1 - \alpha.$$

Hence the bounds $F_n(x) \pm \delta_\alpha$ provide a $(1-\alpha) \times 100\%$ *confidence interval* for the real underlying distribution $F$. A small value for $\alpha$ will result in a wide confidence-band with a high covering probability. As a consequence, one might be tempted to settle for a small $\alpha$-value (e.g. $\alpha = 0.1$). However, the requirement for high coverage confidence needs to be balanced by the need for statistical power to detect alternatives.

Indeed, a very wide confidence band will basically accommodate any choice of $F$ and we could always pick $F$ to represent a simple unimodal density. This way, the constrained optimization principle becomes vacuous. The reason is clear: small values of $\alpha$ maximize the probability that the true underlying distribution is covered, but minimize the likelihood that a real difference will be detected. This will be illustrated in the experiments.

There is another way to see this. Suppose we pick a small $\alpha$ (say 0.1) and construct $F$ that is as smooth as possible and still satisfies $d_{KS}(F, F_n) = \sup_x |F_n(x) - F(x)| = \delta_\alpha$. In particular, this entails that

$$P\left(\sup_x |F_n(x) - F(x)| > \delta_\alpha \quad | \, F_n \text{ is based on sample from } F\right) \leq \alpha = 0.1$$

This means that we put forward an underlying probability $F$ such that the observed sample is *exceptional* (in fact, has a probability of less than $\alpha = 0.1$

of being observed!). Clearly, this is an unsatisfactory state of affairs as we prefer a choice of $F$ that would make the observed sample typical rather than exceptional.

A similar argument can be used to argue against a very large value for $\alpha$ (say 0.9). For in such a case we put forward a density $F$ such that

$$P\,(\sup_x |F_n(x) - F(x)| < \delta_\alpha \quad | \, F_n \text{ is based on sample from } F) \leq 1 - \alpha = 0.1,$$

again an unlikely event. From these considerations it transpires that choosing $\alpha = 0.5$ seems most reasonable. The experiments reported in section 6.1 will further buttress this point.

## 4 Occam's Principle as a Constrained Minimization Problem

At this point we are in a position to reformulate the clustering algorithm (2) in much more precise terms.

(1) Use the data $x_1, \ldots, x_n$ to construct the empirical distribution $F_n(x)$;
(2) Pick a threshold $p$-value $\alpha$ (default value: $\alpha = 0.5$);
(3) Choose a distance function $D_n^{(q)} = d_q(F_n, F)$ $(q = 1, 2)$ and compute the corresponding $\epsilon_q$ where

- $D_n^{(1)} \equiv K_n = \sup_{x \in I\!R} |F_n(x) - F(x)|$ corresponds to the Kolmogorov-Smirnov distance (4);
- $D_n^{(2)} \equiv W_n^2 = n \int_{I\!R} (F_n(x) - F(x))^2 \, dF(x)$ refers to the Cramer-von Mises distance (8):

(4) For the chosen $q$ and $\alpha$, compute $\delta_q(\alpha)$ such that $P(D_n^{(q)} > \delta_q(\alpha)) = \alpha$ using (asymptotic) eqs. (6) or (10) as appropriate.
(5) Determine $F$ by solving the following constrained optimization problem:

$$\text{minimize} \quad \Psi(F) \quad \text{where} \quad \Psi(F) = \int_{I\!R} (F''(x))^2 \, dx \tag{11}$$

$$\text{subject to} \quad d_q(F_n, F) \leq \delta_q$$

In order to formulate a solution for equation (11) we first show how, thanks to the monotonicity of the cumulative distribution, the constraints on $F$ can be reformulated as a finite set of constraints at the distinct sample points. Indeed, for the KS distance, the condition $\sup_{x\in I\!R} |F_n(x) - F(x)| \leq \delta_1$ is equivalent to

$$v_i \leq F(x_i) \leq w_i \tag{12}$$

where $v_i = F_n(x_i) - \delta_1$ and $w_i = F_n(x_i) + \delta_1 - 1/n$. The presence of the $1/n$-term is due to the definition of the cumulative distribution $F_n$ as a right-continuous function which increases with $1/n$ at each sample-point $x_i$.

Likewise, by performing a simple integration, the constraint on the Cramer-von Mises statistic $W_n^2 \leq \delta_2$ can be recast as

$$\sum_{i=1}^{n} \left( F(x_{(i)}) - \frac{i - 1/2}{n} \right)^2 + \frac{1}{12n} \leq \delta_2 \tag{13}$$

with $x_{(i)}$ the ordered sample-points. It thus becomes a discrete constraint of the form

$$\sum_{i=1}^{n} (F(x_i) - y_i)^2 \leq \delta \tag{14}$$

by putting $y_i = (i - 1/2)/n$ and $\delta = \delta_2 - 1/(12n)$.

Since the constraints in (11) can be re-expressed as constraints in the observation points, it follows that the constrained optimization problem is actually an example of a **spline optimization problem** (which, to avoid confusion, we formulate for a function $g$):

$$\text{minimize } S(g) \equiv \int_a^b (g''(x))^2 dx \quad \text{subject to} \quad C_\omega(x_1, \ldots, x_n) \tag{15}$$

where $C_\omega(x_1, \ldots, x_n)$ is one of the following *classical constraints* at the points $a \leq x_i \leq b$, $(i = 1, \ldots, n)$:

C1. *Smoothing problem:* $\sum_{i=1}^n (g(x_i) - y_i)^2 \leq \delta$ for some predefined $\delta > 0$;

C2. *Box problem:* $v_i \leq g(x_i) \leq w_i$.

However, since we are looking for a solution in the class of distribution functions, we need to add two *additional constraints:*

**CDF1** *Monotone increasing:* $g'(x) \geq 0$;
**CDF2** *Limit behaviour:*

$$\begin{cases} \lim_{x \to -\infty} g(x) = 0 \\ \lim_{x \to +\infty} g(x) = 1; \end{cases}$$

These additional constraints will be discussed in more detail in section 5.4.

The optimization problem (15) is well-defined for $W_2[a,b]$ i.e. the class of functions defined on $[a,b]$ with absolutely continuous first derivative and square integrable second derivative.

## 5 Solving the Constrained Minimization Problem

### *5.1 Reformulation as a standard quadratic optimization problem*

It is well-known that the solution of (15) subject to (C1) or (C2) is a cubic spline. Recall that a cubic spline consists of cubic polynomials glued together at the "knots" $x_1, \ldots, x_n$ to ensure continuity of $g, g'$ and $g''$. The first and last spline-segments are linear as a result of the boundary conditions (for more details we refer to [3,7]). Furthermore, standard theory ensures us that the space of cubic splines on $n$ points constitutes a $q = n + 2$ dimensional vector space which implies that one can determine a set of $q = n + 2$ basis-functions $B_i(x)$ such that any cubic spline can be expressed as

$$g(x) = \sum_{i=1}^{q} c_i B_i(x). \tag{16}$$

Among the possible candidates for such a basis, the so-called *B-splines* are a popular choice as they have a local support and are therefore well-behaved numerically. (In fact, our implementations in Matlab are based on this choice of basis.)

Once the basis has been selected the smoothness-functional can be re-expressed as

$$\int_a^b (g''(x))^2 dx = \int_a^b \left( \sum_j c_j B_j''(x) \right)^2 dx = \mathbf{c^t \Sigma c} \tag{17}$$

where $\mathbf{c} = (c_1, \ldots, c_q)^t$ and

$$\mathbf{\Sigma} \in I\!R^{q \times q} \quad \text{with} \quad \mathbf{\Sigma}_{ij} = \int_a^b B_i''(x) B_j''(x)\, dx. \tag{18}$$

The constraints can be recast in a similar fashion by observing that the column vector $(g(x_i))_{i=1}^n$ can be written as $\mathbf{Tc}$ where $\mathbf{T} \in I\!R^{n \times q}$ and $\mathbf{T}_{ij} = B_j(x_i)$. If we denote $\mathbf{y} = (y_1, \ldots, y_n)^t$, then the optimization problem equation (15) is reduced to a matrix optimization problem:

$$\text{minimize } \mathbf{c^t \Sigma c} \quad \text{over } \mathbf{c}, \text{ subject to either} \tag{19}$$

C1. *Smoothing problem:* $\|\mathbf{Tc} - \mathbf{y}\|^2 \leq \delta$;
C2. *Box problem:* $\mathbf{v} \leq \mathbf{Tc} \leq \mathbf{w}$.

### *5.2 Solution of the matrix optimization problems*

The minimization problem (19) subject to the second constraint (C2) is a standard quadratic matrix optimization problem, i.e. a quadratic objective function subject to linear inequality constraints. Hence, only the first constraint (C1) needs further amplification.

Clearly, we can assume that the minimum is realised on the boundary of the closed ellipsoid about $\mathbf{y}$ specified by (C1). Indeed, from equation (17) it is obvious that $\mathbf{\Sigma}$ is non-negative definite. Hence, suppose that $\mathbf{c}$ is a solution within the ellipsoid, then we can further reduce the quadratic objective function (19) by taking $\mathbf{c}_* = \rho \mathbf{c}$ with $0 < \rho < 1$ such that $\mathbf{c}_*$ lies on the boundary of the ellipsoid. Put differently, the inequality in (C1) can be turned into an equality without loss of generality.

As a consequence, introducing a Lagrangian multiplier $\lambda$ turns the constraint minimization (19.C1) into a Lagrangian function

$$L(\mathbf{c},\lambda) = \mathbf{c^t\Sigma c} + \lambda\|\mathbf{Tc} - \mathbf{y}\|^2. \tag{20}$$

Each $\lambda$ gives rise to a spline $\mathbf{c}_\lambda$ with corresponding distance $\delta_\lambda = E(\mathbf{c}_\lambda) \equiv \|\mathbf{Tc}_\lambda - \mathbf{y}\|^2$. Iteratively updating $\lambda$ yields the solution corresponding to the distance $\delta$. For fixed $\lambda$, a straightforward solution to (20) is obtained by solving the linear system:

$$(\mathbf{T^tT} + \lambda^{-1}\mathbf{\Sigma})\,\mathbf{c}_\lambda = \mathbf{T^t y}. \tag{21}$$

But if there are many datapoints (say $n > 100$) there is no need to solve the above large system as a simple approximation performs equally well. This is discussed next.

### *5.3 Implementation of approximative solution*

If there are lots of data, there is no need to position a knot at every observed datapoint. In fact, computational efficiency will be improved if we approximate the original spline-solution with a spline having uniformly spaced (grid)points $t_1, \ldots, t_p$ as knots, where typically, $p$ is much smaller than the number of datapoints $n$.

Sticking to the notation $\mathbf{c}$, however this time to express the expansion of the approximate spline with respect to the $(p+2)$ basic $B$-splines defined on the uniformly spaced $t$-knots, the smoothing term still equals $\mathbf{c^t\Sigma c}$, however this time $\mathbf{\Sigma} \in I\!R^{(p+2)\times(p+2)}$. First, we discuss the optimization problem subject to constraint (C1). The expression in the constraint remains $E(\mathbf{c}) = \|\mathbf{Tc} - \mathbf{y}\|^2$ where now $\mathbf{T} \in I\!R^{n\times(p+2)}$ and $\mathbf{T}_{ij} = B_j(x_i)$ but this time the $B$-splines are defined on the set of $t$-knots. Since typically $p \ll n$ the matrix $\mathbf{T}$ is strongly rectangular, with far more rows than columns.

Applying a QR-decomposition to this matrix yields $\mathbf{T} = \mathbf{QR}$ where $\mathbf{Q}$ a $n \times (p+2)$ matrix with orthonormal columns, and $\mathbf{R}$ a square upper-triangular square matrix of size $(p+2)$. Since $\mathbf{T^tT} = \mathbf{R^tQ^tQR} = \mathbf{R^tR}$ it follows that equation (20) can be written as

$$\mathbf{c^t\Sigma c} + \lambda\|\mathbf{Rc} - \eta\|^2 \tag{22}$$

and equation (21) is reduced to simple square system of size $(p+2)$:

$$(\mathbf{R^tR} + \lambda^{\mathbf{-1}}\mathbf{\Sigma})\mathbf{c}_\lambda = \mathbf{R^t}\eta \tag{23}$$

where $\eta = \mathbf{Q^t y} \in I\!R^{p+2}$. Since the coefficient matrix is square, symmetric and strictly positive-definite, the solution is straightforward.

In the case of constraint (C2), changing from datapoints to equi-distant gridpoints is not completely adequate as constraint satisfaction on the reduced set does not imply similar compliance on the original. However, we can still use the spline defined on the gridpoints with the corresponding box-constraints and adjust the parameter $\delta$ in (25) until the original constraints are satisfied. So the reduced problem is

$$\text{minimize } \mathbf{c^t \Sigma c} \text{ subject to } \mathbf{v}^* \leq \mathbf{T}^*\mathbf{c} \leq \mathbf{w}^* \tag{24}$$

with

$$v_i^* = F_n(t_i) - \delta, \quad w_i^* = F_n(t_i) + \delta - \frac{1}{n}, \tag{25}$$

and

$$\mathbf{T}^* \in I\!R^{p \times (p+2)} \text{ where } \mathbf{T}^*{}_{ij} = B_j(t_i). \tag{26}$$

### *5.4 Enforcing the additional constraints*

Using $B$-splines as basis-functions the monotonicity condition (CDF1) can be translated into a finite number of linear constraints. We outline the main idea of this proposal which was proposed by Schwetlick and Kunert [9].

Since $g$ is a cubic spline, its derivative is a spline of degree 2. The coefficients $\mathbf{c} = (c_1, \ldots, c_{n+2})^t$ of $g$ with respect to the $B$-splines on the ordered dataset $x_1, \ldots, x_n$ are related to the coefficients $\mathbf{d} = (d_1, \ldots, d_{n+1})^t$ of its derivative $g'$ by a linear relationship: $\mathbf{d} = \mathbf{Ac}$ where $\mathbf{A} = \mathbf{LM}$ with $\mathbf{M} \in I\!R^{(n+1)\times(n+2)}$ and $\mathbf{L} \in I\!R^{(n+1)\times(n+1)}$ given by

$$\mathbf{M} = \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix}, \quad \mathbf{L} = \text{diag}\left(\frac{3}{x_i - x_{i-3}}\right)$$

with $i = 2, \ldots, n+2$. Notice that we use the standard extended dataset for cubic splines which is obtained by adding two additional points to both the left ($x_{-1}$ and $x_0$) and right ($x_{n+1}$ and $x_{n+2}$) endpoint of the original dataset. Moreover, if gridpoints rather than datapoints are used as knots, the number of points $n$ must be replaced by the number of gridpoints $p$ and $\mathbf{A}$ becomes a $(p+1) \times (p+2)$ matrix.

Since all $B$-splines are positive, the constraint $g'(x) \geq 0$ can be replaced by the stronger condition $\mathbf{d} = \mathbf{Ac} \geq \mathbf{0}$. Introducing this extra condition to the optimization problem (19) does not change its numerical (algorithmic) complexity as a quadratic optimization problem with linear constraints.

Finally, the 0-1-bounds for the cumulative distribution function can be easily incorporated in the optimization problem by adding the constraints $0 \leq g(x_1)$ and $g(x_n) \leq 1$. In matrix-notation this amounts to $0 \leq \mathbf{T_{1.}c}$ and $\mathbf{T_{n.}c} \leq 1$ where $\mathbf{T_{1.}}$ and $\mathbf{T_{n.}}$ denote the first and last $T$-row, respectively.

In summary, the minimization problem subject to constraint (C1) and these extra conditions amounts to (for $\lambda$ fixed):

$$\boxed{\begin{gathered} \min_{\mathbf{c}} \mathbf{c^t}(\mathbf{\Sigma} + \lambda \mathbf{R^t R})\mathbf{c} - \mathbf{2}\lambda\eta^{\mathbf{t}}\mathbf{Rc} \qquad (27) \\ \text{subject to} \quad \begin{cases} \mathbf{Ac} \geq \mathbf{0} \\ \mathbf{T_{1.}c} \geq 0 \\ \mathbf{T_{n.}c} \leq 1 \end{cases} \end{gathered}}$$

The quadratic optimization problem with constraint (C2) and the additional conditions is

$$\boxed{\begin{gathered} \min_{\mathbf{c}} \mathbf{c^t \Sigma c} \qquad (28) \\ \text{subject to} \quad \begin{cases} \max(\mathbf{v}^*, \mathbf{0}) \leq \mathbf{T}^*\mathbf{c} \leq \min(\mathbf{w}^*, \mathbf{1}) \\ \mathbf{Ac} \geq \mathbf{0} \end{cases} \end{gathered}}$$

The bounds $\mathbf{v}^*$ and $\mathbf{w}^*$ are adjusted to satisfy constraint (C2).

### *5.5 Overview of the computations*

**Data-compatibility based on Cramer-von Mises**

(1) Collect the data $x_1, \ldots, x_n$.
(2) Fix $\alpha$ and determine the corresponding $\delta_2$. E.g. $\alpha = 0.5$ yields $\delta_2 = 0.119$.
(3) Compute $y_1, \ldots, y_n$ and $\delta$ as defined in equation (14).
(4) Construct a grid $t_1, \ldots, t_p$ $(p = 50)$ on which the spline will be defined.
(5) Compute the matrices $\mathbf{\Sigma} \in I\!R^{(p+2)\times(p+2)}$ and $\mathbf{T} \in I\!R^{n\times(p+2)}$. Apply a QR-decomposition on $\mathbf{T}$ and define $\eta = \mathbf{Q^t y}$.
(6) Propose the regression-line through $(x_i, y_i)$ as the initial solution ($\lambda = 0$). Denote the coefficients of this line with respect to the basis of $B$-splines defined on the knots $t_1, \ldots, t_p$ by $\mathbf{c_0}$. If $E(\mathbf{c}_0) = \|\mathbf{Tc_0} - \mathbf{y}\|^2 \leq \delta$, terminate the program.
(7) Otherwise, update $\lambda$. In each iteration-step, first solve the quadratic optimization problem (27) for fixed $\lambda$ to get the coefficients $\mathbf{c}$. (Recall that we use an approximation by imposing a stronger monotonicity constraint than is actually needed.) Then compute $E(\mathbf{c})$ and stop if it's close to $\delta$. Then the spline at the knots $t_i$ with coefficients $\mathbf{c}$ is the solution of the problem.

**Data-compatibility based on Kolmogorov-Smirnov**

(1) Collect the data $x_1, \ldots, x_n$.
(2) Fix $\alpha$ and determine the corresponding $\delta_1$. E.g. $\alpha = 0.5$, yields $\delta_1 = 0.828/\sqrt{n}$.
(3) Define a grid $t_1, \ldots, t_p$ $(p = 50)$ on which the $B$-splines $B_1, \ldots, B_{p+2}$ are defined.
(4) Compute the empirical distribution of the data at the gridpoints $F_n(t_1), \ldots, F_n(t_p)$.
(5) Compute the matrices $\mathbf{\Sigma} \in I\!R^{(p+2)\times(p+2)}$ and $\mathbf{T}^* \in I\!R^{p\times(p+2)}$ as defined in equation (26). These matrices are used in step 8.
(6) Compute the empirical distribution of the data: $F_n(x_1), \ldots, F_n(x_n)$. Compute the original bounds $\mathbf{v}(\delta_1)$ and $\mathbf{w}(\delta_1)$ as defined in equation (12) and the matrix $\mathbf{T} \in I\!R^{n\times(p+2)}$ with $T_{ij} = B_j(x_i)$. These computations are used to check the constraint in step 9.
(7) Denote by $\delta_0$ an initial value to compute the bounds $v_i^* = F_n(t_i) - \delta_0$ and $w_i^* = F_n(t_i) + \delta_0 - 1/n$.
(8) Solve the quadratic optimization problem (28).
(9) If the constraint $\mathbf{v}(\delta_1) \leq \mathbf{Tc} \leq \mathbf{w}(\delta_1)$ is satisfied, terminate the program. Then the spline defined on the knots $t_i$ with coefficients $\mathbf{c}$ is the solution of the problem.
(10) Otherwise decrease $\delta$, recompute the bounds $v_i^*$ and $w_i^*$ and find the solution of (28) with these bounds.

## 6 Experimental Results

### 6.1 *Experimental evidence for choice of $\alpha$ threshold*

In section 2 we argued that the $\alpha$-parameter should be set equal to 0.5 since we want a typical rather than an exceptional density-model to fit the given data. This way, we intend to strike a reasonable balance between the covering probability of the confidence interval on the one hand, and the power against alternatives on the other. In this section we report on some experiments that confirm the appropriateness of this choice.

**Example 1:** In the first example, we generated 100 samples of size 1000 from a Gaussian mixture distribution $\pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2)$ where $\pi_1 = \pi_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$ and $\mu_1 = 0$ and $\mu_2 = d$; hence $d$ is the separation between the two clusters which in our experiments varies from 2 to 4. Figure 1 shows some typical histograms of datasets drawn from the given distribution for various values of $d$. We applied our histogram-segmentation method using different values for $\alpha$ on each of the generated datasets and stored the number of extracted clusters. The results are summarized in table 1. From Figure 1 it is clear that the histogram begins to show bi-modality at around $d = 2.5$. This is picked up nicely when $\alpha = 0.5$. Setting $\alpha = 0.1$ however, results in a wide confidence band that accommodates an overly smooth (and hence unimodal) solution for $F$. The decision in favour of bi-modality is therefore postponed until $d \geq 3$. Conversely, fixing $\alpha = 0.9$ increases data-fidelity and the transition from 1 to 2 clusters occurs earlier (somewhere around 2.3).

| $\alpha = 0.1$ | $d$=2 | $d$=2.5 | $d$=2.8 | $d$=3 | $d$=3.5 | $d$=4 |
|---|---|---|---|---|---|---|
| 1 cluster | 100 | 100 | 93 | 44 | 0 | 0 |
| 2 clusters | 0 | 0 | 7 | 56 | 100 | 100 |

| $\alpha = 0.5$ | $d$=2 | $d$=2.5 | $d$=2.8 | $d$=3 | $d$=3.5 | $d$=4 |
|---|---|---|---|---|---|---|
| 1 cluster | 100 | 76 | 16 | 0 | 0 | 0 |
| 2 clusters | 0 | 24 | 84 | 100 | 100 | 100 |

| $\alpha = 0.9$ | $d$=2 | $d$=2.5 | $d$=2.8 | $d$=3 | $d$=3.5 | $d$=4 |
|---|---|---|---|---|---|---|
| 1 cluster | 95 | 23 | 0 | 0 | 0 | 0 |
| 2 clusters | 5 | 77 | 100 | 100 | 100 | 100 |

Table 1
*The table displays for each distance d the number of samples for which our method found 1 or 2 clusters. To illustrate the appropriateness of setting $\alpha = 0.5$ we also list the results for more extreme choices $\alpha = 0.1$ (over-smoothing) and $\alpha = 0.9$ (under-smoothing).*

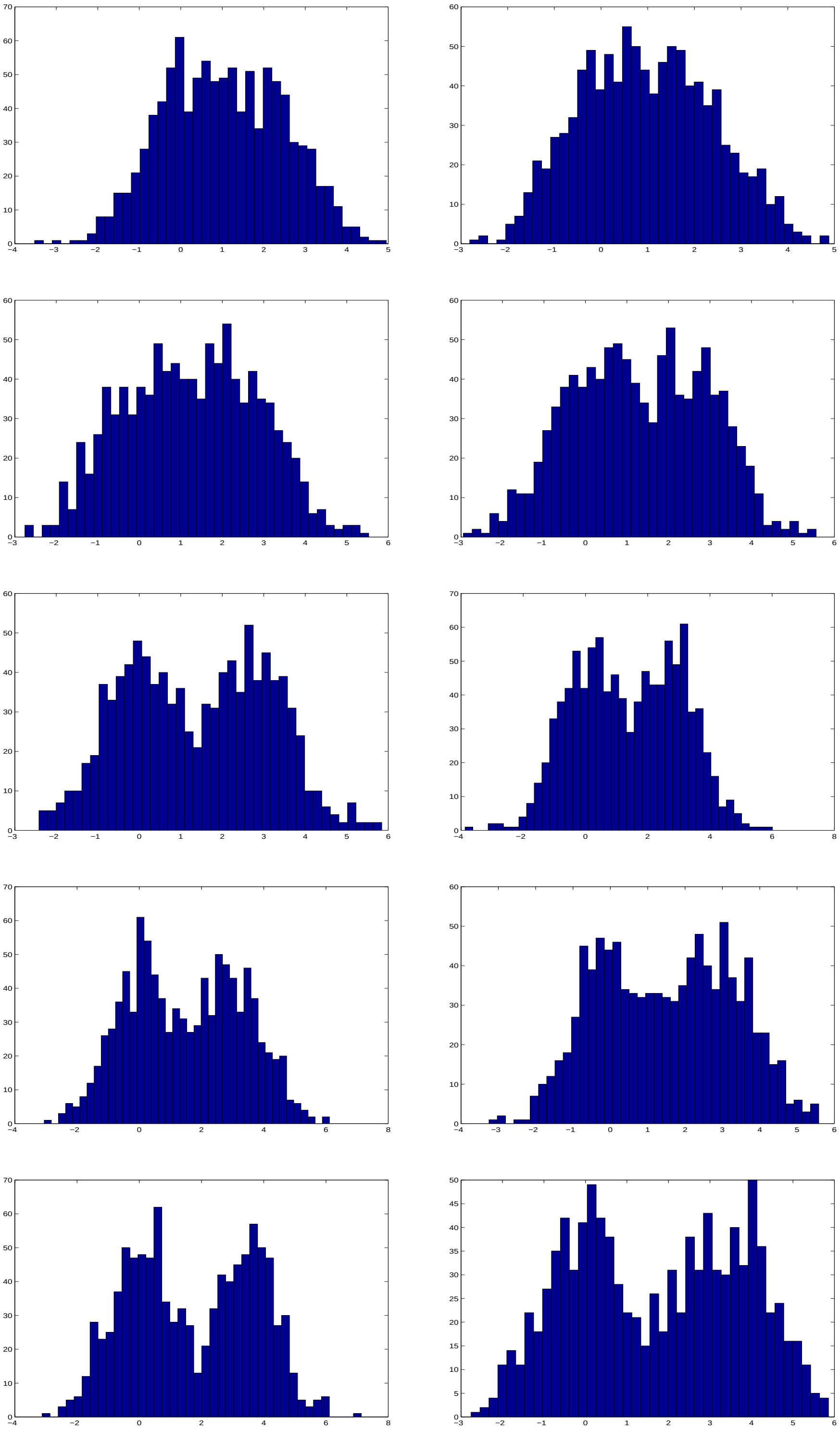

Fig. 1. Typical histograms of datasets in Section 6.1, Example 1. The samples of size 1000 taken from Gaussian mixture distributions $0.5N(0,1) + 0.5N(d,1)$, i.e. an equal mixture of two unit-variance Gaussians separated by a distance $d$. Each row shows two typical realisations for increasing distance: $d$ grows from $d = 2$ (first row), over $d = 2.5$, $d = 2.8$ and $d = 3$ to $d = 3.5$ in the last row (i.e. Example 1).

**Example 2:** In a second experiment, we generated 100 samples of size 1000 from a Gaussian mixture distribution with three super-imposed components $\sum_{i=1}^{3} \pi_i N(\mu_i, \sigma_i^2)$ where $\mu_1 = 0$, $\sigma_1 = 1$, $\pi_1 = 0.37$; $\mu_2 = d$, $\sigma_2 = 1$, $\pi_2 = 0.26$ and $\mu_3 = 2d$, $\sigma_3 = 1$, $\pi_3 = 0.37$.

The typical histograms shown in figure 2 reveal 1 cluster for $d = 1.5$, 1 or 2 clusters for $d = 2$, 2 clusters for $d = 2.5$, $d = 3$ results into 2 or 3 clusters and for $d = 3.5$ the 3 clusters are clearly isolated. Table 2 displays the results for three choices of $\alpha$. As in the previous example, $\alpha = 0.5$ behaves as expected while $\alpha = 0.1$ appears to be biased towards overly smooth solutions and therefore under-estimates the number of clusters. Notice that the solution for the other extreme choice $\alpha = 0.9$ seems to behave satisfactorily. However, this choice results in very narrow confidence bands, forcing the solution to adhere closely to the empirical distribution function. As a result it will tend to over-estimate the number of clusters as is demonstrated in the next example.

| $\alpha = 0.1$ | $d = 1$ | $d = 1.5$ | $d = 2$ | $d = 2.5$ | $d = 3$ | $d = 3.5$ | $d = 4$ |
|---|---|---|---|---|---|---|---|
| 1 cluster | 100 | 100 | 99 | 51 | 1 | 1 | 1 |
| 2 clusters | 0 | 0 | 1 | 49 | 99 | 72 | 1 |
| 3 clusters | 0 | 0 | 0 | 0 | 0 | 27 | 98 |

| $\alpha = 0.5$ | $d = 1$ | $d = 1.5$ | $d = 2$ | $d = 2.5$ | $d = 3$ | $d = 3.5$ | $d = 4$ |
|---|---|---|---|---|---|---|---|
| 1 cluster | 100 | 100 | 52 | 1 | 0 | 0 | 0 |
| 2 clusters | 0 | 0 | 48 | 99 | 80 | 4 | 0 |
| 3 clusters | 0 | 0 | 0 | 0 | 20 | 96 | 100 |

| $\alpha = 0.9$ | $d = 1$ | $d = 1.5$ | $d = 2$ | $d = 2.5$ | $d = 3$ | $d = 3.5$ | $d = 4$ |
|---|---|---|---|---|---|---|---|
| 1 cluster | 99 | 84 | 11 | 0 | 0 | 0 | 0 |
| 2 clusters | 1 | 16 | 89 | 94 | 19 | 0 | 0 |
| 3 clusters | 0 | 0 | 0 | 6 | 81 | 100 | 100 |

Table 2
*The table displays for each $d$ the number of samples for which our method found 1, 2 or 3 clusters. For each $d$ the $\alpha$-parameter is set equal to one of the extreme values* 0.1 *and* 0.9 *or to the suggested (typical) value* 0.5.

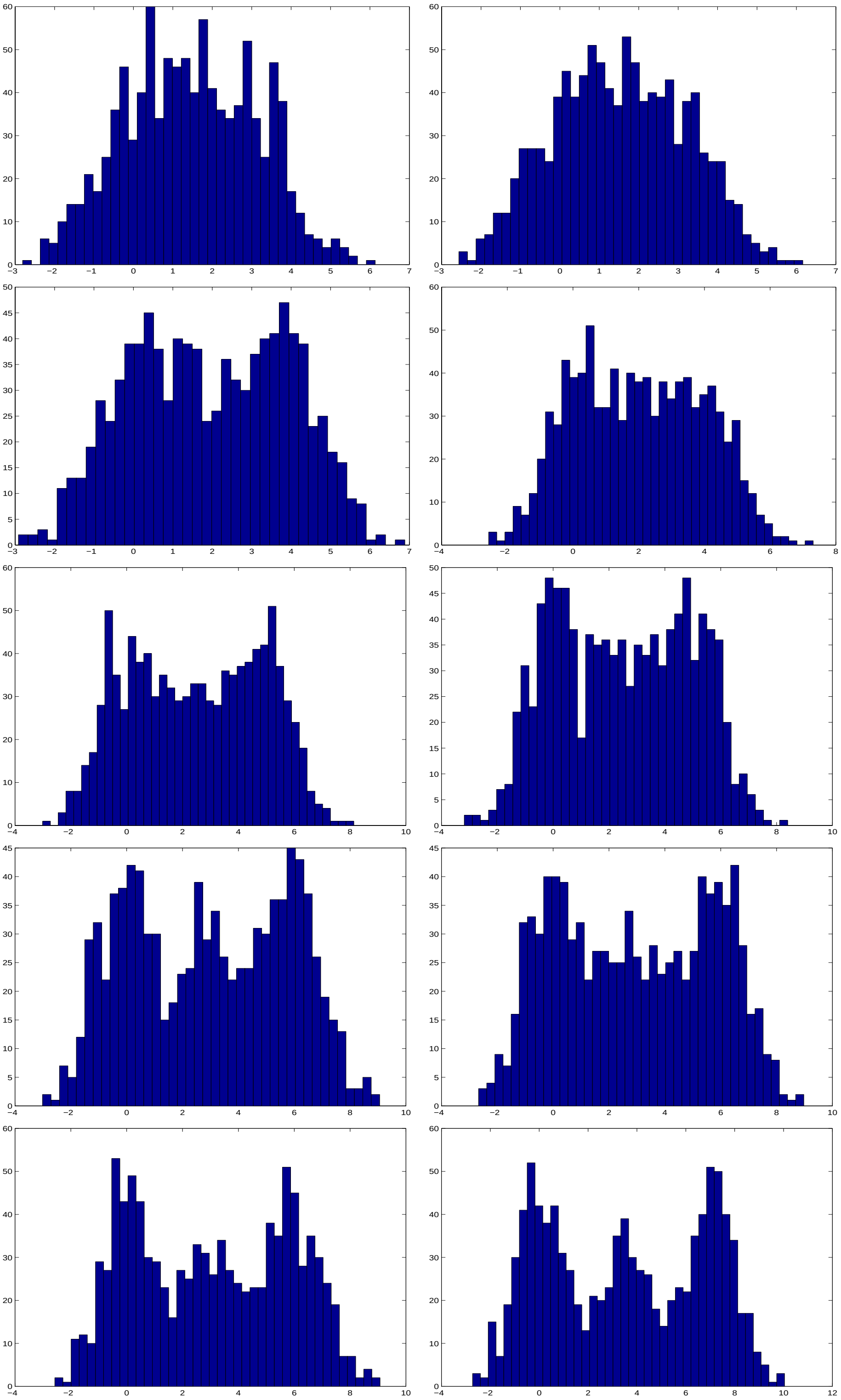

Fig. 2. Typical histograms of datasets in Section 6.1, Example 2. Samples of size 1000 are taken from a 3-component Gaussian mixture $0.37N(0,1) + 0.26N(d,1) + 0.37N(2d,1)$, i.e. a mixture of two unit-variance Gaussians separated by a distance $2d$, with a slightly smaller unit-Gaussian cluster in between. Each row shows two typical realisations for increasing separation: $d$ grows from $d = 1.5$ (first row), over $d = 2$, $d = 2.5$ and $d = 3$ to $d = 3.5$ in the last row.

**Example 3:** From the experiments reported above, one might be tempted to choose a large value for $\alpha$, e.g. $\alpha = 0.9$. This, however, would be a mistake as the covering probability $1-\alpha$ would then become too small. As a consequence, the probability that the real underlying distribution is within the computed bounds is small (e.g. only 10% if we pick $\alpha = 0.9$) and the proposed solution will be largely determined by the random fluctuations in the sample. This transpires from the next experiment in which we compare, for different values of $\alpha$, the number of clusters found in a sample of size 1000 from a uniform $U(0, 1)$ distribution (i.e. there is one real underlying cluster). We see that for $\alpha = 0.9$ the number of clusters is over-estimated in about one third of the experiments.

| | $\alpha = 0.1$ | $\alpha = 0.5$ | $\alpha = 0.9$ |
|---|---|---|---|
| 1 cluster | 100 | 96 | 66 |
| 2 clusters | 0 | 4 | 28 |
| 3 clusters | 0 | 0 | 6 |

Table 3

*Number of clusters found by the proposed algorithm for different values of $\alpha$. Each sample of size 1000 is drawn from a uniform $U(0, 1)$ sample (i.e. there is only one real underlying cluster). In total, 100 experiments were performed. It transpires that for $\alpha = 0.9$ the number of clusters is significantly overestimated, whereas $\alpha = 0.5$ has a very acceptable error-rate.*

### *6.2 Application to Image Segmentation*

We have applied the proposed 1-dimensional clustering method to the problem of image segmentation, and we illustrate the results for both natural and synthetic images (decoration designs). In all the experiments reported below we used constraints based on the CvM distance function in combination with $\alpha = 0.5$. For more information we refer the interested reader to *http://www.cwi.nl/∼pauwels/research/*.

**Natural images:**

To accomplish segmentation, the pixels of an image are mapped into a number of colour spaces such as RGB, LAB and opponent-colours. An interesting alternative are the purely data-driven axes that are extracted from a PCA-analysis of RGB-space. One then gets different 1-dimensional histograms by looking at the projection on different coordinate axes in these spaces. The clusterings of the resulting histograms can easily be assigned a saliency score by checking whether or not there is more than one cluster and if so, how well-separated and pronounced these clusters are (e.g. by comparing the distance between their means to their variance). In the experiments reported below (see Figs. 3 and 4) we display for each image one or two of the most salient histograms and the corresponding clustering.

**Decoration designs:** These are synthetic and manually designed images that need to be decomposed in foreground and background (to extract design motifs). To avoid being misled by variations in the background, each image is smoothed at a number of increasingly coarser scales. The smoothed images are transformed to the LAB-colour space and each projected dataset is clustered. This procedure is stopped as soon as our clustering-method presents two clusters. Some results are shown in figure 5.

## 7 Conclusion

In this paper we have introduced a non-parametric clustering algorithm for *1-dimensional* data. The procedure looks for the *simplest (i.e. smoothest) density that is still compatible with the data.* Compatibility is given a precise meaning in terms of distribution-free statistics based on the empirical distribution function, such as the *Kolmogorov-Smirnov* or the *Cramér-von Mises* statistic. This approach is therefore genuinely non-parametric and does not involve fixing arbitrary cost- or fudge-factors. The only parameter that needs to be specified (and is fixed in advance for once and for all) is the statistical risk factor $\alpha$. In a follow-up paper we will elaborate how this strategy can be extended to more dimensions. For more information we refer the reader to *http://www.cwi.nl/~pauwels/research/*.

**Acknowledgements:** This research was partially supported by the European Commission's 5th Framework IST Project FOUNDIT (Project nr. IST-2000-28427).

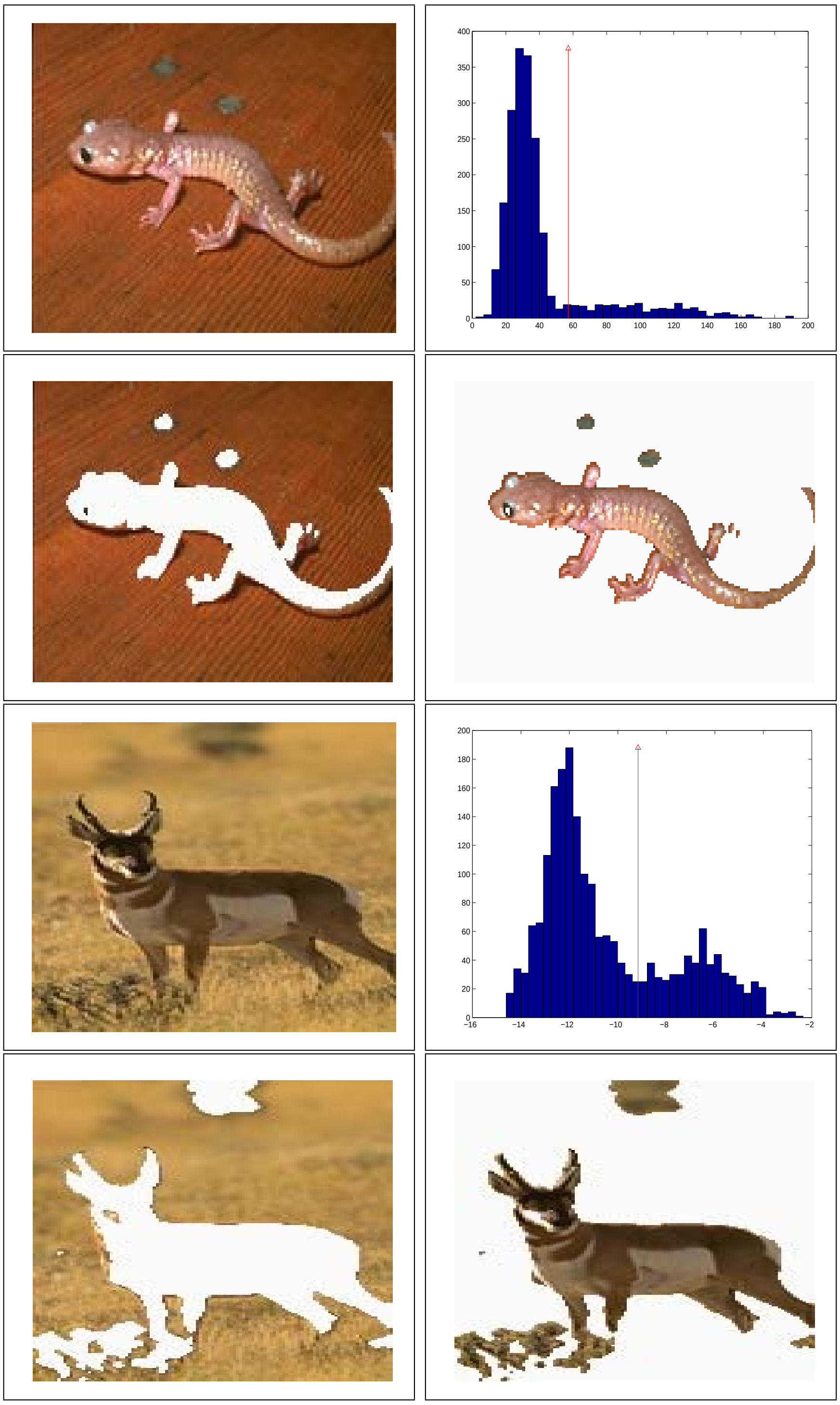

Fig. 3. Histogram-based colour segmentation of natural images. The histograms are based on the principal components (PCA) of the pixels in RGB-space. Using PCA components amounts to a data-driven way of maximizing colour-contrast. *First (last resp.) two rows:* The original image (top left), the histogram of the first PCA component (top right), and the image decomposition based on the histogram segmentation (next row).

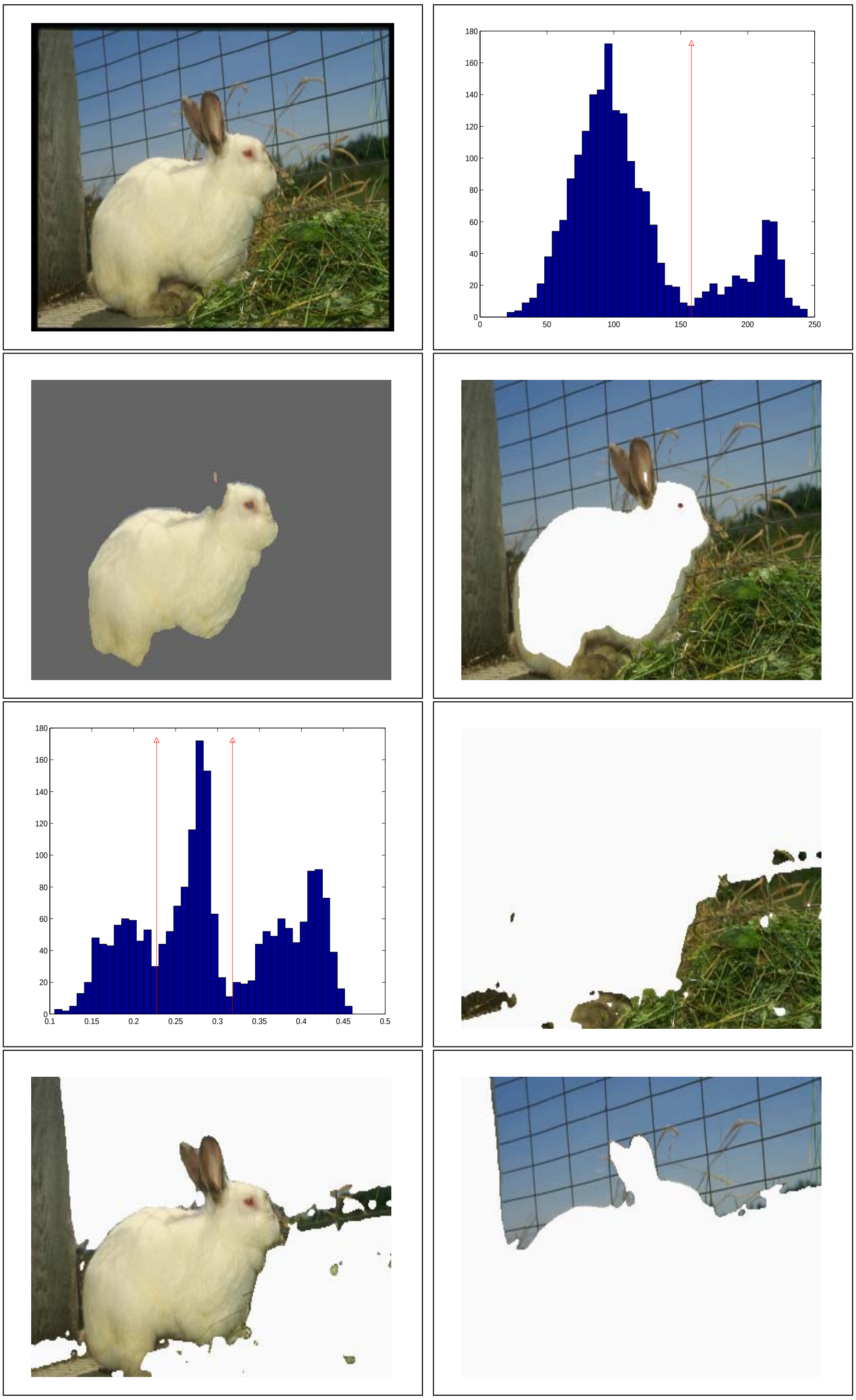

Fig. 4. Histogram-based colour segmentation of natural image (continued). *First two rows:* The original image (top left), the histogram of the first PCA component (top right), and the image decomposition based on the histogram segmentation (2nd row). *Third and fourth row:* Histogram for 2nd PCA component can be segmented in three groups. The corresponding image regions are shown on the 3rd and 4th row.

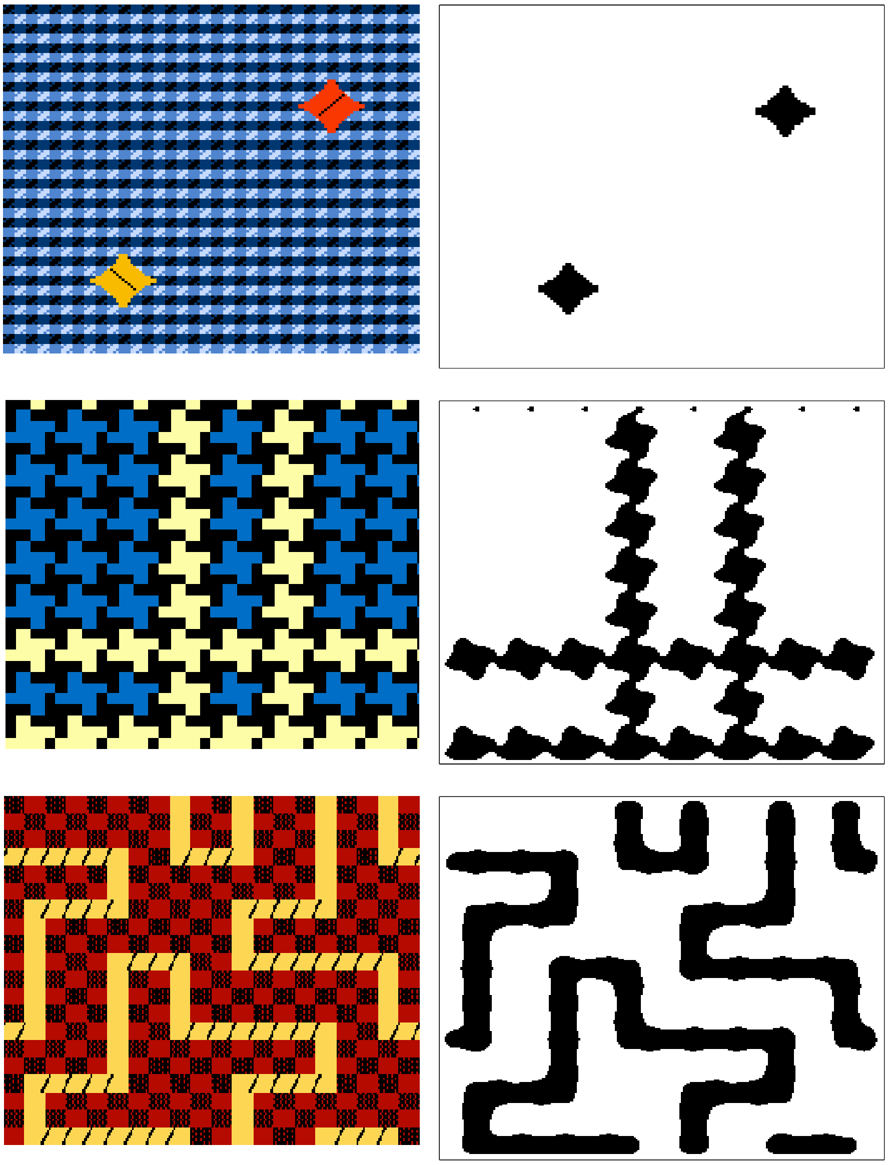

Fig. 5. Background-foreground separation for textile decoration patterns. *Left column:* Original patterns. *Right column:* Background-foreground separation based on colour clustering of first PCA-components. Prior to clustering, the images are first smoothed using a Gaussian filter, to highlight the salient regions.

## References

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover, 1970

[2] T.W. Anderson and D.A. Darling, *Asymptotic theory of certain goodness of fit criteria based on stochastic processes*, Ann. Math. Statist. Vol. 23, 1952 pp. 193-212

[3] C. de Boor, *A Practical Guide to Splines.* Applied Mathematical Sciences Vol. 27, Springer-Verlag 1978

[4] A.P. Dempster, N.M. Laird and D.R. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Soc. Ser. B.* 39:1-38, 1977.

[5] R.O. Duda, P.E. Hart and D.G. Stork. *Pattern Classification.* WileyInterscience. Wiley and Sons, 2001.

[6] J. Durbin, *Distribution Theory for Tests Based on the Sample Distribution Function.* SIAM Regional Conf. Series in Applied Mathematics, Society for industrial and applied mathematics Philadelphia 1973

[7] R.L. Eubank: *Spline Smoothing and Nonparametric Regression.* Statistics: textbooks and monographs Vol. 90, Dekker New York 1988

[8] T. Hastie, R. Tibshirani, J. Friedman: *The Elements of Statistical Learning.* Springer, 2001.

[9] H. Schwetlick and V. Kunert, *Spline smoothing under constraints on derivatives.* Bit, Vol. 33, 1993 pp. 512-528

[10] G. Wahba: *Spline Models for Observational Data.* CBMS-NSF Regional Conf. Series in Applied Math. No 59, Society for Industrial and Applied Math., Philadelphia 1990.

[11] `http://www.cwi.nl/∼pauwels/research/`