Appeared in Pattern Recognition, Vol. 28, pp.823–831, 1995

AUTOMATIC CONTEXT–SENSITIVE KARYOTYPING OF HUMAN CHROMOSOMES BASED ON ELLIPTICALLY SYMMETRIC STATISTICAL DISTRIBUTIONS

GUNTER RITTER,* MARIA TERESA GALLEGOS and KARL GAGGERMEIER

Fakultät für Mathematik und Informatik, Universität Passau, D-94030 Passau, Germany

Abstract – We introduce a statistical model of a metaphase cell consisting of independent chromosomes with elliptically symmetric feature vectors. From this model we derive the ML–classifier for classification in the 24 chromosomal classes, taking into account the correct number of chromosomes in each class. Experimental results show that error rates of the best of these classifiers are less than 2% with respect to chromosomes if applied to the large Copenhagen data set Cpr. Simulation studies suggest that there should be even more information contained in the features of this data set.

Context–sensitive chromosome classification, Karyotyping, Elliptically symmetric (elliptically contoured) distribution, Statistical pattern recognition, Discriminant analysis

1. INTRODUCTION

In a normal, nucleated human cell there are 44 autosomal chomosomes and two sex chromosomes, X, X in females and X, Y in males. The 44 autosomal chromosomes can be decomposed into 22 classes 1..22 each of which consists of a matching pair of two *homologous* chromosomes; also the two X chromosomes in female cells are homologous. A representation of the chromosomal complement showing this class structure is called a *karyotype*. Producing a karyotype of a cell is of practical importance since it much facilitates the detection of abnormalities in chromosome structure. Readers interested in the biological and cytogenetical background of karyotyping are referred to the survey article of Piper et al. [15] and Habbema [7].

Karyotyping is a well–specified problem the solution of which by hand is well understood. It can be carried out by a human expert on normal cells at a speed of 15 min/cell and, depending on the quality of the image of the cell, at a low error rate between 1 and 3 misclassifications among 1000 chromosomes which amounts roughly 2-7% of misclassified cells (cf. Lundsteen et al.[11], Granum [5], cf. also Zimmerman et al. [19]). It was one of the first problems to be tackled by the methods of *automatic* pattern recognition in the early 1960s. This automation problem is interesting in at least four respects:

(1) It is of clinical relevance since a satisfactory solution would relieve cytogeneticists from the routine of karyotyping in most cases normal cells by hand and allow them to concentrate on detecting abnormalities;

(2) it is a challenging task of artificial intelligence and automatic pattern recognition;

(3) it bears interesting aspects of high–dimensional multivariate statistics; and of

(4) combinatorial optimization.

Automated pattern recognition usually consists of four phases; *viz.*: digitization of analogue images, segmentation, feature extraction and classification. In the case of karyotyping the analogue image is a microscopic image of the suitably stained chromosomes of a metaphase cell. By staining the chromosomal band structure becomes visible, thus enabling classification of chromosomes in 24 classes. Figure 1 shows a schematic representation of this band structure and Fig. 2 a photographic image of a stained metaphase cell. The task of classification, whis is the subject we are interested in here, consists in arranging the 46 chromosomes in such a way that their classes are

$$\begin{array}{ccc} 1\,2\ldots22\,X & \textit{and} & 1\,2\ldots22\,X \\ 1\,2\ldots22\,X & & 1\,2\ldots22\,Y \end{array}$$

*Author to whom correspondence should be addressed.

in the female and male cases, respectively. The upper and lower arrays in a karyotype are called *haploids*.

Classification can be carried out by the methods of Bayesian *discriminance analysis*, the individuals to be classified being chromosomes or cells. Accordingly, there exist two techniques of assigning chromosomes to classes:

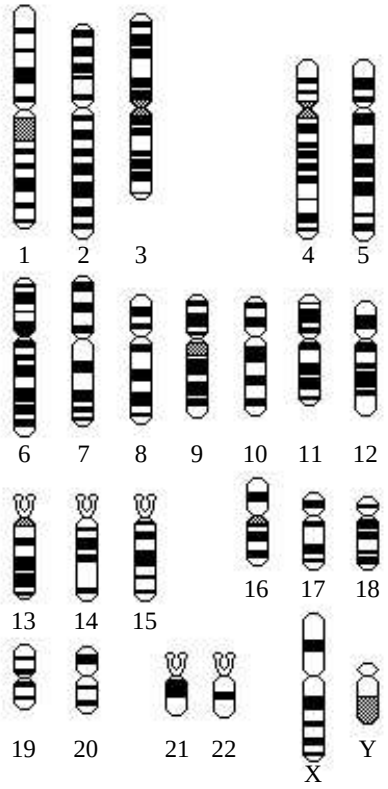

Figure 1: Schematic representation of chromosome classes and band structures

(1) *Chromosome–by–chromosome* or *context–free* allocation. In this, earlier, approach the individuals classified are chromosomes in isolation. This method can be modified by a subsequent (context–sensitive) amendment so that the numbers of chromosomes in the various classes (one or two) are correct. Assessments of methods based upon this approach are contained in Piper [13] and Kirby et al.[8] Error rates for the most accurate of these methods and with the features used so far amount to about 5% relative to chromosomes.

(2) *Context–sensitive* allocation. Here, the individuals classified are cells. By this we mean that the optimization contained in the Bayesian classifier is constrained by the correct numbers in the classes $1, \ldots, 22, X$ and Y. Using the constraints in this straight–forward way was suggested already by Habbema [6, 7] and Slot [16]. However, the first efficient algorithm was described by Tso and Graham [18] who considered classification into the 10 Denver groups and pointed out that the optimization problem involved, in the case of their method of classification, is a Hitchcock transportation problem. Later, Tso et al.[17] applied this method of classification to karyotyping and designed an algorithm applicable to incomplete cells.

The statistical model applied by many research groups postulates *independence* and *normality* of the feature vectors of chromosomes* which necessitates estimating expectations and covariances of features within chromosomes. Classification is then carried out by minimizing the sum of the squared Mahalanobis distance of feature vectors of chromosomes to class centers over all permutations. Some simplifications of this model are also considered in the literature. One of these simplifications is the assumption of *independent features* (cf. Piper [12]) supported by the fact that chromosomes are mixtures of the (independent) ancestors of the individual; another one is that of *equal correlation* matrices of feature vectors of chromosomes (cf. Kirby et al. [8]). It was even found that, besides the effect of speedup, simplification of the first–mentioned kind decreases the error rate if the training set is small so that postulating uncorrelatedness of features proves superior to relying on unsafe estimates of correlation coefficients. For large training sets, however, the inclusion of covariance coefficients within chromosomes reduces the error rate (cf. Piper [13]). The error rate of the best of these methods is at about 3% relative to chromosomes and 38% relative to cells.

The starting point of the present communication is the observation that the joint distribution of the features of chromosomes has a tail as compared with the normal distribution and outliers resulting from this tail are responsible of a large amount of misclassified chromosomes. We show that dropping the assumption of normality while retaining that of independence substantially reduces the error rate. Instead of normality we assume that feature vectors of chromosomes are *elliptically symmetric* (*elliptically contoured*). Thus, the joint distribution of the feature set of a random chromosome is characterized by a mean value, a variance matrix, and a radial function (cf. Subsection 2.1). We denote the resulting (optimal) ML–estimator in the context–sensitive setting by IEC_γ. The acronym IEC means "independent and elliptically contoured chromosomes" and the subscript γ stands for the radial function or its shape; a formal definition appears in Subsection 3.1. Many of the estimators found in the literature are special cases for various choices of γ (cf. 3.2). If the *empirical* radial function is used for γ the resulting

*When talking about the distribution of the feature vectore of a chromosome we actually mean the distribution of the feature vector *conditioned on the chromosome class*.

Figure 2: Microscopic view of stained metaphase cell

estimator has an error rate of about 1.9% relative to chromosomes and 26% relative to cells if applied to the large Cophenhagen data set Cpr consisting (to date) of 2804 cells.

The outlines of the paper are as follows. In Section 2 we describe the statistical assumptions on the feature vectors of chromosomes used here, in particular the assumptions of independence and elliptical symmetry. Section 3 contains the ML–estimators derived from these statistical models and algorithms for computing them. In Section 4 we present experimental results using the holdout and resubstitution methods for classification. Possible sources of misclassification, exceeding the Bayes error for the present feature set, are missing independence, missing elliptical symmetry of feature vectors, and errors in the estimated parameters. In order to assess the statistical model used this section also contains simulation studies. Here we generate artificial cell data with the use of several elliptically symmetric distributions (differing in their radial distributions) and classify these cells with respect to their population parameters and the holdout and resubstitution methods. In this way we obtain some information about the error rates to be expected if one of these distributions were the true one and thereby about the quality of the statistical assumptions made.

We will adhere to the following notation. The symbol $\mathbb{R}$ denotes the real line and $\mathbb{R}_+$ the set of real numbers ≥ 0. We will write $\| \, \|$ for the Euclidean norm on a Euclidean space $\mathbb{R}^m$ and a superscript T means transposition of a vector or matrix. For a random vector $X : \Omega \to \mathbb{R}^m$ possessing the necessary moments, EX and $VX = E(X-EX)(X-EX)^T$ denote its expectation and variance matrix, respectively, and f_X denotes its density function. The permutation group of m elements will be denoted $\mathcal{S}_m$.

2. THE STATISTICAL MODEL OF A CELL

We now introduce our statistical model of a cell. This will be used in Section 3 in order to derive various optimal estimators in different situations. For definiteness, we deal with *human* cells, mutatis mutandis, the methods may, however, be applied to any other species and, in fact, to any classification problem of independent, random objects subject to suitable constraints.

2.1. *Elliptical symmetry*

A random vector $S_0 : \Omega \to \mathbb{R}^d$ possessing a density function f_{S_0} is called *spherically symmetric* (cf. Dempster [2] and Fang et al. [3]), if f_{S_0} is of the form

$$f_{S_0}(x) = \varphi(\|x\|) \qquad (x \in \mathbb{R}^d) \tag{1}$$

with a *radial function* $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ specifying the distribution of S_0. The radial function models a specific *tail behaviour* of S_0. Moreover, the norm $\|S_0\|$ has a density which, by spherical symmetry, equals

$$f_{\|S_0\|}(r) = \omega_{d-1} r^{d-1} \varphi(r), \qquad r \geq 0. \tag{2}$$

The number $\omega_{d-1} = 2\pi^{d/2}/\Gamma(d/2)$ is the surface of the $(d-1)$–dimensional unit sphere. We will call $f_{\|S_0\|}$ the *radial density* of S_0. From (2) it follows that a measurable function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is the radial function of a spherically symmetric random vector S_0 if and only if

$$\omega_{d-1} \int_0^\infty r^{d-1} \varphi(r) \mathrm{d}r = 1. \tag{3}$$

If $E\|S_0\| = \omega_{d-1} \int_0^\infty r^d \varphi(r)\, \mathrm{d}r < \infty$ then the expectation ES_0 of S_0 exists and vanishes. If $E\|S_0\|^2 = \omega_{d-1} \int_0^\infty r^{d+1} \varphi(r)\, \mathrm{d}r < \infty$ then its variance matrix $VS_0 = E[S_0 S_0^T]$ exists, is finite, and must be a multiple αI of the identity matrix I. The factor α is determined by the equality

$$d \cdot \alpha = \text{trace } VS_0 = E\|S_0\|^2 = \omega_{d-1} \int_0^\infty r^{d+1} \varphi(r)\, \mathrm{d}r. \tag{4}$$

A random vector $S : \Omega \to \mathbb{R}^d$ is called *elliptically symmetric* (*elliptically contoured*) if it is the affine image of some spherically symmetric random variable S_0; i.e., there exists a vector $c \in \mathbb{R}^d$ and a matrix $A \in \mathbb{R}^{d \times d}$ such that $S = c + AS_0$. By spherical symmetry of S_0, the matrix A may be assumed to be symmetric and positive. If S possesses a density function f_S (which we will assume) then A may be assumed to be positive definite. In this case, the elliptically symmetric random vector is specified by the vector $c \in \mathbb{R}^d$, the positive definite matrix $A \in \mathbb{R}^{d \times d}$, and the radial function φ of S_0. If $E\|S\|^2 < \infty$, or equivalently $E\|S_0\|^2 < \infty$, we may and do suppose that S_0 is normalized, i.e.,

$$VS_0 = I \text{ and } \alpha = 1.$$

We next show how to compute these three parameters from the distribution of S in the latter case. As a consequence of symmetry of S_0 we first have

$$c = ES; \tag{5}$$

moreover, $VS = A(VS_0)A = A^2$, i.e.,

$$A = \sqrt{VS}. \tag{6}$$

It follows that

$$S_0 = (VS)^{-1/2}(S - ES), \tag{7}$$

the *Mahalanobis transform* of S. Concerning the radial function φ it is useful to introduce the *Mahalanobis distance* of $x, y \in \mathbb{R}^d$ with respect to a positive definite matrix V. This is the number

$$m_V(x, y) := \sqrt{(x-y)^T V^{-1} (x-y)}. \tag{8}$$

Since $\|S_0\| = m_{VS}(S, ES)$ by (7) and (8), from (2) it follows that

$$\varphi(r) = r^{1-d} f_{m_{VS}(S,ES)}(r)/\omega_{d-1}. \tag{9}$$

Formula (9) is the basis for the radial function φ via an estimate of the radial density $f_{m_{VS}(S,ES)}$, i.e., the distribution of the random Mahalanobis distance $m_{VS}(S, ES) = \|S_0\|$.

Using an integral transformation, (7) and (1), the density f_S may be written in terms of ES, VS, φ, and m_{VS} as

$$\begin{aligned} f_S(x) &= (\det(VS)^{-1/2}) f_{S_0}((VS)^{-1/2}(x - ES)) \\ &= (\det(VS)^{-1/2}) \varphi(\|(VS)^{-1/2}(x - ES)\|) \\ &= (\det(VS)^{-1/2}) \varphi(m_{VS}(x, ES)). \end{aligned} \tag{10}$$

This relation will be needed for representing the joint distribution of the random karyotype in 2.4. For more details on elliptically symmetric distributions the interested reader is referred to Dempster [2] and Fang et al. [3].

2.2. *The configuration set*

For ease of exposition we first deal with female cells; in Subsection 3.3 male cells will be treated in a similar way. Neglecting incomplete cells with fewer than 46 chromosomes as well as cells with extra chromosomes a female cell consists of 23 matching pairs of two homologous chromosomes belonging to the classes 1..23. Given an arbitrary numbering of the chromosomes by the set 1..46 there are 2^{23} permutations $\sigma \in \mathcal{S}_{46}$ ($\mathcal{S}_{46}$ represented as the group of bijective mappings $1..46 \to 1..46$) such that chromosome i is of class $\sigma(i)$ mod 23.* We will call the (uniquely defined) induced mapping

$$\begin{array}{rcl} \tilde{\sigma} : 1..46 & \to & 1..23 \\ i & \to & \sigma(i) \mod 23 \end{array}$$

the *configuration* of the cell. Each of these permutations σ represents the same configuration $\tilde{\sigma}$. The classification problem consists in estimating this configuration $\tilde{\sigma}$, i.e. one of the correctly classifying permutations σ. Hence, the decision set of the statistical decision problem under consideration is the configuration set $\{\tilde{\sigma}/\sigma \in \mathcal{S}_{46}\}$ consisting of $46!/2^{23} \approx 6.56 \cdot 10^{50}$ mappings $1..46 \to 1..23$, where each value is taken on exactly twice. A particular configuration is the *karyotypic configuration* $\tilde{\kappa}$ induced by the identity permutation κ. A cell in configuration $\tilde{\kappa}$ is present in its *karyotype*.

2.3. *Feature vector and parameter set*

Let 1..46 be the numbers of the chromosomes of a cell in karyotypic configuration, more specifically, let the two haploids be numbered $1, \ldots, 23$ and $24, \ldots, 46$. To each chromosomes $i \in 1..46$ there is associated a random, real, d–dimensional *feature vector* $Z_i : (\Omega, P) \to \mathbb{R}^d$. Therefore, the measurements of the two haploids are modeled by random, real $23 \cdot d$–dimensional feature vectors

$$(Z_1, \ldots, Z_{23}), (Z_{24}, \ldots, Z_{46}) : (\Omega, P) \to \mathbb{R}^{23 \cdot d}.$$

What we actually observe with an unclassified cell is not the "karyotype" $Z = (Z_1, \ldots, Z_{46})$ but this joint vector Z in disorder: A vector $x \in \mathbb{R}^{46 \cdot d}$ will be considered as being composed of 46 vectors $x_1, \ldots, x_{46} \in \mathbb{R}^d$, i.e., x_i is the ith block of length d in x. Each permutation $\sigma \in \mathcal{S}_{46}$ induces in a natural way an isomorphism $\pi_\sigma \in GL(46 \cdot d)$ defined by $\pi_\sigma x := (x_{\sigma(i)})_{1 \le i \le 46}$. This isomorphism maps the subspace of $\mathbb{R}^{46 \cdot d}$ belonging to the ith block of length d onto the subspace belonging to the $\sigma^{-1}(i)$th block of length d. What we observe with an unclassified cell is the vector $\pi_\sigma Z$ for some $\sigma \in \mathcal{S}_{46}$; then $\tilde{\sigma}$ is the configuration of the cell.

2.4. *The joint distribution of the random karyotype*

We assume that the random variables $Z_1, \ldots, Z_{46}$ are *independent* and *elliptically symmetric* with parameters $e_i = EZ_i$, $V_i = VZ_i$, and φ_i, the radial function of the Mahalanobis transform $V_i^{-1/2}(Z_i - e_i)$ of Z_i [cf. 2.1(7)]. Moreover, we assume that feature vectors of homologous chromosomes are statistically identical, i.e., in a *female* cell, we have

$$e_{i+23} = e_i, \ V_{i+23} = V_i, \ \varphi_{i+23} = \varphi_i$$

for all i, $1 \le i \le 23$. In other words, the parameters e_i, V_i and φ_i depend on classes only.

The assumption of independence is questionable. Habbema [7], reasoned that correlation between chromosomes in one cell may be caused by a "common history of culture, preparation, staining, and photography" and suggested that using the whole cell as an entity should be beneficial to error rates. Another source of dependence is normalization of features across the chromosomes of a cell. Nevertheless, this assumption turned out to be useful in the past and we shall stick to it here.

*Contrary to usual convention in mathematics and computer science we represent integers mod m as elements of the set $1..m$ here. In particular, 23 mod 23 and 46 mod 23 are represented by the number 23. In this way, class numbers of chromosomes are those used in cytogenetics.

The joint distribution of the random karyotype $Z = (Z_1, \ldots, Z_{46})$ is now completely specified by the parameters e_i, V_i, φ_i $(1 \le i \le 23)$ of chromosome classes and, by 2.1(10), we have for $x = (x_1, \ldots, x_{46}) \in \mathbb{R}^{46 \cdot d}$

$$f_Z(x) = \prod_{i=1}^{46} f_{Z_i}(x_i) = \prod_{i=1}^{46} (\det V_i^{-1/2}) \varphi_i(m_{V_i}(x_i, e_i)) = \text{const} \prod_{i=1}^{46} \varphi_i(m_i(x_i)); \tag{1}$$

here, we used the notation

$$m_i(y) = m_{V_i}(y, e_i) \quad 1 \le i \le 46, y \in \mathbb{R}^d. \tag{2}$$

3. THE CLASSIFIERS

3.1. *The ML–estimator*

We now derive the optimal estimate of the configuration $\tilde{\sigma}$ of a cell from the observation $X(\omega) = \pi_\sigma Z(\omega) = x \in \mathbb{R}^{46 \cdot d}$ (cf. 2.2 and 2.3). The statistical decision model $(X, (\mu_\sigma)_{\sigma \in \mathfrak{S}_{46}})$ of this classification problem consists of the random observation $X = \pi_\sigma Z : \Omega \to \mathbb{R}^{46 \cdot d}$ of the (disordered) cell, the parameter set $\mathfrak{S}_{46}$, and the probability distributions $\mu_\sigma (\sigma \in \mathfrak{S}_{46})$,

$$\frac{\mu_\sigma(\mathrm{d}x)}{\mathrm{d}x} = f_{\pi_\sigma Z}(x) = f_Z(\pi_{\sigma^{-1}} x) = \prod_{i=1}^{46} f_{Z_i}(x_{\sigma^{-1}(i)}) = \text{const} \prod_{i=1}^{46} \varphi_i(m_i(x_{\sigma^{-1}(i)})) \tag{1}$$

[cf. 2.4(1)]. Since μ_σ depends on $\tilde{\sigma}$ only, it is not important to distinguish between σ and $\tilde{\sigma}$. From (1) we infer that the negative log–likelihood function $l_x : \mathfrak{S}_{46} \to \mathbb{R}$ of the statistical decision model is (up to an additive constant) given by

$$l_x(\sigma) = \sum_{i=1}^{46} - \ln \varphi_i(m_i(x_{\sigma^{-1}(i)})).$$

Supposing that $\varphi_i = \varphi$ is independent of i we obtain

$$l_x(\sigma) = \sum_{i=1}^{46} - \ln \varphi(m_i(x_{\sigma^{-1}(i)})), \quad x \in \mathbb{R}^{46 \cdot d}, \sigma \in \mathfrak{S}_{46}. \tag{2}$$

It is reasonable to assume that all permutations occur equally likely; thus the (optimal) MAP–estimator is the ML–estimator defined by

$$IEC_\varphi(x) = \operatorname*{argmin}_{\sigma \in \mathfrak{S}_{46}} l_x(\sigma), \qquad x \in \mathbb{R}^{46 \cdot d}. \tag{3}$$

The formulae (2), (3) and 2.4(2) combined define the ML–estimators employed here. We illustrate this with a few examples.

3.2. *Examples*

(a) (*Normal case*) If the radial function φ is of the form

$$\varphi(r) = (2\pi)^{-d/2} e^{-r^2/2} \qquad (r \ge 0) \tag{1}$$

then feature vectors of chromosomes are normal and, by their independence, the whole cell is normal. The ML–estimator is

$$IEC_{\text{normal}}(x) = \operatorname*{argmin}_{\sigma \in \mathfrak{S}_{46}} \sum_{i=1}^{46} m_i^2(x_{\sigma^{-1}(i)}).$$

This estimator was proposed by Habbema [7] and Slot [16].

In Piper [12], there appears a variant of the Estimator IEC_{normal}: if the *features* within each class are assumed to be independent then the variances V_i reduce to diagonal matrices. If the training set is small (up to 200 cells) then this assumption seems to be superior to using estimates of the covariances which are too unsafe under these circumstances.

(b) (*Exponential case*) Let now S_0 possess an exponential tail, i.e.,

$$\varphi(r) = \beta e^{-\lambda r} \qquad (r \geq 0). \tag{2}$$

In order to ensure 2.1(3) and $\alpha = 1$ we have to choose $\lambda = \sqrt{d+1}$ and $\beta = \lambda^d/\Gamma(d)\omega_{d-1}$. According to Subsection 3.1, the ML–estimator becomes

$$IEC_{\exp}(x) = \operatorname*{argmin}_{\sigma \in \mathcal{S}_{46}} \sum_{i=1}^{46} m_i(x_{\sigma^{-1}(i)}).$$

(c) (*Pareto case*) Finally, let S_0 possess a Pareto tail, i.e.,

$$\varphi(r) = \begin{cases} \beta r^{-\lambda}, & r \geq r_0, \\ 0, & 0 \leq r < r_0, \end{cases}$$

for some real number $\lambda > d+2$. In order to ensure 2.1(3) and $\alpha = 1$ we have to put $r_0 = \sqrt{d(\lambda - d - 2)/(\lambda - d)}$ and $\beta = (\lambda - d) r_0^{\lambda - d}/\omega_{d-1}$. From 3.1 it follows that the ML–estimator is

$$IEC_{\text{Pareto}}(x) = \operatorname*{argmin}_{\sigma \in \mathcal{S}_{46}} \sum_{i=1}^{46} \ln m_i(x_{\sigma^{-1}(i)}) \qquad (x \in \mathbb{R}^{46 \cdot 4}).$$

This estimator and the estimator $IEC_{\exp}$ were introduced by Kleinschmidt et al. [10], who pointed out their robustness when applied to real data.

A radial function with similar tail behavior is

$$\varphi(r) = \frac{\beta}{\eta + r^\lambda} \qquad (r \geq 0)$$

for some real number $\lambda > d + 2$. In order to ensure 2.1(3) and $\alpha = 1$ one has to put

$$\eta = \left[d \left(\sin \frac{d+2}{\lambda} \pi \right) / \left(\sin \frac{d}{\lambda} \pi \right) \right]^{\lambda/2}$$

and

$$\beta = \left(\eta^{(\lambda - d)/\lambda} \lambda \sin \frac{d}{\lambda} \pi \right) / (\pi \omega_{d-1}).$$

Again similar is Pearson's type VII distribution with

$$\varphi(r) = \frac{\beta}{(1 + r^2/\eta)^{\lambda/2}}$$

for some real number $\lambda > d + 2$ and suitable constants β and η (cf. Fang et al. [3]).

3.3. *Male cells*

The statistical model in the male case is similar to that in the female case, however, we have to redefine the notions of *configuration* and *karyotype*. A male cell consists of 22 classes of homologous chromosomes and the nonhomologous chromosomes X and Y to which we assign classes 23 and 24, respectively. We assume that the first "haploid" $(Z_1, \ldots, Z_{23})$ consists of the features of chromosomes of classses 1 to 22 and the X chromosome and the second $(Z_{24}, \ldots, Z_{46})$ consists of those of chromosomes of the classes 1 to 22 and the Y chromosome.

Given an arbitrary numbering of the chromosomes of a male cell by the set 1..46 there are 2^{22} permutations $\sigma \in \mathcal{S}_{46}$ such that chromosome i is of class $\sigma(i) \mod 23$ if $\sigma(i) < 46$. We call the mapping

$$\begin{array}{rccl} \tilde{\sigma}: & 1..46 & \to & 1..24 \\ & i & \mapsto & \begin{cases} \sigma(i) \mod 23, & \text{if } \sigma(i) < 46, \\ 24, & \text{if } \sigma(i) = 46, \end{cases} \end{array}$$

the *configuration* of the male cell. Each of these permutations σ represents the configuration $\tilde{\sigma}$. The *karyotypic* configuration $\tilde{\kappa}$ is induced by the identity permutation κ. The classification problem consists in estimating $\tilde{\sigma}$, the decision set containing this time $46!/2^{22} \approx 1.31 \cdot 10^{51}$ configurations.

Again, we assume that feature vectors of chromosomes are independent and elliptically symmetric, their parameters depending on their classes 1..24. Except for the constant, formula 2.4(1) for the joint density of features of chromosomes and the estimators in Subsection 3.1 remain unchanged. If the sex of a cell is unknown then both estimators are applied and the sex is that of the estimator with the smaller negative log–likelihood value. For another way of dealing with unknown sex cf. the following subsection.

3.4. *Optimization*

The minimization task in computing ML–classifiers for statistical models with independent feature vectors of chromosomes reduces to a Hitchcock (transportation) problem as we pointed out by Tso and Graham [18]. Indeed, for IEC, we have the female case [cf. 3.1(2),(3)]

$$IEC_\varphi(x) = \operatorname*{argmin}_\sigma l_x(\sigma) = \operatorname*{argmin}_\sigma \sum_{i=1}^{46} -\ln(\varphi(m_i(x_{\sigma^{-1}(i)})))$$

$$= \operatorname*{argmin}_\sigma \sum_{i=1}^{46}\sum_{j=1}^{46} -\ln(\varphi(m_i(x_j)))\delta_{i,\sigma(j)}$$

$$= \operatorname*{argmin}_\sigma \sum_{i=1}^{23}\sum_{j=1}^{46} -\ln(\varphi(m_i(x_j)))(\delta_{i,\sigma(j)} + \delta_{i+23,\sigma(j)})$$

(δ_{ij} is Kronecker's delta). Putting $c_{ij} := -\ln(\varphi(m_i(x_j)))$ and $s_{ij} := \delta_{i,\sigma(j)} + \delta_{i+23,\sigma(j)}$ $(1 \le i \le 23, 1 \le j \le 46)$ the problem reduces to finding natural numbers s_{ij} such that

$$\sum_{i=1}^{23}\sum_{j=1}^{46} c_{ij}s_{ij} \text{ is minimal subject to the restrictions } \sum_i s_{ij} = 1, \sum_j s_{ij} = 2.$$

There exist very efficient solution algorithms for the Hitchcock problem, e.g., the extension of the Balinski algorithm [1] due to Kleinschmidt et al.[9] employed for karyotyping in Tso et al. [17]. The latter paper also contains a related algorithm for classification into 24 classes independent of the sex of a cell and an algorithm for karyotyping cells with missing chromosomes. These algorithms are applicable also to the present statistical model.

4. **EXPERIMENTAL RESULTS**

We have implemented the algorithms described in Section 3 on a SUN IPX workstation. Our implementation works at a speed of about 1 cell/s.

4.1. *Data set*

For our experiments we used the large Copenhagen data set Cpr; cf. Piper [13]. It consists to date of 2804 metaphase amnion cells 1305 of which are complete female cells (and 1428 are complete male cells). Each chromosome is described by 30 features $0, \ldots, 29$ extracted by the MRC chromosome analysis system described in Piper and Granum [14]. These features contain information about size, density, convex hull perimeter, centromeric index, shape, and band pattern (cf. also Granum [5] and Piper et al. [15] for a description of features). Since features 0, 7, 27 and 4, 26 are highly correlated with features 1 and 3, respectively, and since the last feature 29 shows a tendency to increase error rates, we have worked with the remaining 24 features. As the purpose of this paper is comparison of different statistical models (and their related ML–estimators) we restricted matters to classifying female cells only and, in addition, we left aside all cells with missing or extra chromosomes.

4.2. *Estimated probabilities of misclassification*

It is customary and appropriate to use the leave–out–one cross validation for assessing error rates: The parameters are computed on the basis of all cells but one and this residual cell is classified; this procedure is repeated for every single cell. This method would be very time consuming since the matrix

of covariances would have to be computed and inverted each time anew. We therefore resorted to a jackknifing method: We subdivided the 1305 complete female cells randomly in ten different ways in training sets of 1100 cells and the complementary test set of 205 cells. Table 1 shows the error rates measured with these test sets $0, \ldots, 9$ using the parameters of the relative training sets for the estimators IEC_{normal}, IEC_{exp}, IEC_{Pareto} [cf. 3.2(a)–(c)], and IEC_{emp}. In all cases the sample expectations $\widehat{e}_i$ and variances $\widehat{V}_i$ were used in the estimators. The subscript "emp" refers to a smoothed version of the sample radial density: We first computed the histograms h_i (absolute frequency) of the random variables $m_{\widehat{V}_i}(S, \widehat{e}_i)$, $1 \leq i \leq 23$, subdividing the positive real axis in intervals of length 10^{-2}. After adding these 23 histograms, intervals between $r = 0$ and the mode with no observations were assigned the value 0.1 in order to avoid zeroes (the algorithms are only weakly sensitive to the radial function at low values of r). We then approximated the resulting histogram h roughly by a $\mathcal{C}^1$–reference function g on the interval $[0, 15]$.* The quotient h/g, which fluctuates around 1, was smoothed using a triangle function with the varying breadth $(h(r) + 1)^{-0.3}$ at r resulting in a function $q : \mathbb{R}_+ \to \mathbb{R}_+$. The smoothed version of the empirical radial density is qg and the sample radial function $\widehat{\varphi}$ required in 3.1(2),(3) is computed via 2.1(9) as $\widehat{\varphi} = r^{1-d} qg/\omega_{d-1}$.

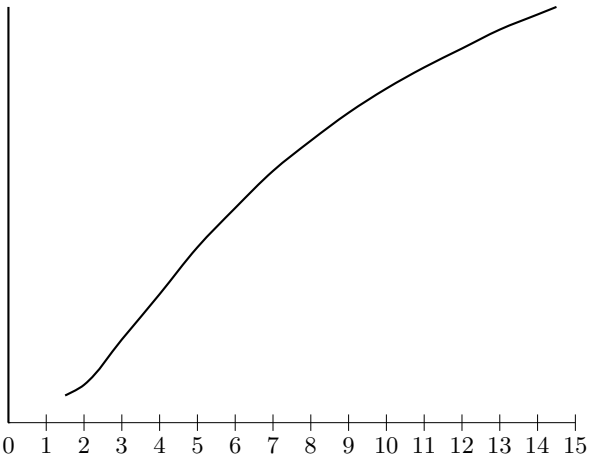

Figure 3: Smoothed version of $-\ln \widehat{\varphi}$

Table 2 shows the mean holdout error rates computed from Table 1 and the "optimistic" resubstitution error rates. The resubstitution method consists in using the whole data set of 1305 cells as training and test sets. The Bayes error can be expected to lie between the holdout and the resubstitution error rates (cf. Fukunaga [4]). Since the results in different lines of Table 1 are only weakly correlated we can conclude that a S.D. in Table 2 (cells) is roughly 1 (using the formula for the uncorrelated case). In the mean, there are about three to four chromosomes misclassified in a misclassified cell. It follows from the error rates of Table 2 (cells) that a S.D. in Table 2 (chromosomes) is roughly 0.1.

4.3. *Simulations*

In order to validate the statistical model described in Section 2 we have generated artificial cells by means of the three models Z^{normal}, Z^{exp} and Z^{Pareto}. Here, $Z^\gamma : \Omega \to \mathbb{R}^{46 \cdot 24}$ consists of 46 independent, elliptically symmetric components $Z_l^\gamma : \Omega \to \mathbb{R}^{24}$ where expectations and variances are those estimated from the 1305 complete female cells in Cpr and the radial functions are independent of classes and as described in Examples 3.2(a)–(c). Specifically, for Z^{exp}, from 3.2(b) it follows $\lambda = 5$. For Z^{Pareto} we choose φ in such a way that the tail of $\|S_0\|$ approximately coincides with that estimated from Cpr, i.e., $\lambda = 28$ (cf. also the reference function g in the footnote of Subsection 4.2). From 3.2(c) it follows that one has to choose $r_0 = \sqrt{12}$ and $\beta = 576/\omega_{d-1}$.

We carried out three different test procedures and the results obtained are shown in Table 3.

(α) (*Bayes error rate*) We tested 50,000 random realizations of Z^γ using the correct parameters of Z^γ in the classifier defined by 3.1(2),(3) and 2.4(2).

(β) (*Holdout error rate*) We generated 1100 random realizations from Z^γ for estimating the empirical expectations $\widehat{e}_i$, variances $\widehat{V}_i$, and the empirical radial function $\widehat{\varphi}$. In order to estimate $\widehat{\varphi}$ we computed histograms of the random variables $m_{\widehat{V}_i}(S, \widehat{e}_i)$ [cf. 2.1(9)] subdividing the positive real axis in intervals of length 10^{-2}. After adding the histograms, intervals between $r = 0$ and the mode with no observations were assigned the value 0.1 and the resulting histogram was smoothed out using triangle functions of varying breadth so that in particular all further zeroes lying in the tail of the histogram disappeared. These parameters were used in the classifier IEC_γ in order to test 50,000 random realizations sampled from Z^γ different from the 1100 above.

*The reference function g is defined by $g(r) = \begin{cases} 0.1, & 0 \leq r < 1.5, \\ 10.753(r - 1.5)^5(4 - r) + 0.1, & 1.5 \leq r < 3.4, \\ 0.0052826 r^{12} e^{-0.3778 r^2}, & 3.4 \leq r < 4.743, \\ 334503 r^{-5}, & r \geq 4.743. \end{cases}$

Classifier subdivision	IEC_{normal}	IEC_{exp}	IEC_{Pareto}	IEC_{emp}
0	272/69	219/60	182/54	186/54
1	279/77	270/73	198/53	194/52
2	255/68	235/61	165/51	159/49
3	271/82	266/77	187/65	183/63
4	286/80	233/64	159/49	161/50
5	321/85	303/77	215/53	218/54
6	287/78	250/67	172/51	172/51
7	296/87	238/73	185/60	183/59
8	284/86	225/75	168/60	166/60
9	237/67	205/56	152/44	150/43

Table 1: Number of errors with respect to chromosomes and cells for 205 classified cells, four different classifiers, and 10 different subdivisions of data set. The figures m/n mean that a total of m chromosomes were misclassified in n cells. S.D. with respect to cells is 7 for IEC_{normal} and IEC_{exp} and 6 for IEC_{Pareto} and IEC_{emp}.

Classifier method	IEC_{normal}	IEC_{exp}	IEC_{Pareto}	IEC_{emp}
Chromosomes				
Holdout(%)	2.96	2.59	1.89	1.88
Resubstitution(%)	2.55	2.29	1.63	1.65
Cells				
Holdout(%)	38.0	33.3	26.3	26.1
Resubstitution(%)	33.8	30.2	23.6	23.8

Table 2: Holdout and resubstitution error rates with respect to chromosomes and cells for four different classifiers

(γ) (*Resubstitution error rate*) 1300 random realizations from Z^γ were used as training and test sets. No smoothing of the radial function was carried out and removal of zeroes as in (β) is not necessary here. The whole procedure was repeated three times and Table 3 shows the resulting averages.

A S.D. in the first two lines of Table 3 (cells) is roughly 0.07, and in the last line 0.2. Since, on the average, there are only slightly more than two chromosomes misclassified in a misclassified cell it follows that a S.D. in the first two lines of Table 3 (chromosomes) is roughly 0.03, and in the last line 0.1.

5. **DISCUSSION**

The results presented in Tables 1 and 2 show that error rates are reduced by about 35% if an optimal elliptically symmetric model instead of a normal model of feature vectors of chromosomes is employed. The new issue arising from this model is the radial function of the associated spherically symmetric distribution. Basically, one has to distinguish between radial functions whose negative logarithm is convex, affine or concave. In the case of the Cpr data set the negative logarithm is essentially concave (Fig. 3) explaining the superior performance of the classifier IEC_{Pareto} besides IEC_{emp}. The advantage of IEC_{emp} over IEC_{Pareto} is its independence of the particular feature set since the empirical radial function is used. On the other hand, IEC_{Pareto} is robust since this estimator does not depend on the Parameter λ of the Pareto–type radial density [cf. Example 3.2(c)]. Hence it is optimal for a whole class of distributions.

It is important to hit the correct behavior of the radial function in the domain right after the larger of the two inflection points of the radial density (which is roughly shaped like a χ–distribution). The Ma-

Model method	Z^{normal}	Z^{exp}	Z^{Pareto}
Chromosomes			
Bayes(‰)	0.72	0.89	0.73
Holdout(‰)	0.77	0.99	0.87
Resubstitution(‰)	0.61	0.91	0.55
Cells			
Bayes(%)	1.62	1.99	1.63
Holdout(%)	1.74	2.22	1.96
Resubstitution(%)	1.38	1.97	1.23

Table 3: Simulated error rates with respect to chromosomes and cells for sampling from the statistical model shown in the top line using the parameters indicated in the first column

halanobis distances of many of the feature vectors of chromosomes fall into this domain and membership to their classes is already doubtful, at least to an automatic classifier.

The simulation results presented in Table 3 show that holdout error rates lie around 0.1% of chromosomes and 2% of cells if the correct statistical model is used for classification and with rare exceptions there are only two classes mixed up in a misclassified cell. Astonishingly, this describes just the performance of an experienced human operator classifying chromosomal images of good quality; cf. the references cited in the introduction. This suggests two conclusions:

(1) Although IEC_{Pareto} and IEC_{emp} yield the best results when applied to real data their related statistical models seem far from being adequate. We conjecture that there is more information contained in the feature set of the MRC chromosome analysis system than we can presently extract, although it may take a much larger data set in order to exploit this information.

(2) It seems that a human expert is not only capable of extracting all the relevant features from the (analogue or digital) images of chromosomes but he or she also possesses some close statistical model of them and disposes of the means for determining a good approximation to its ML–estimator. Smallness of simulated holdout error rates implies that error rates observed with the real data cannot be caused by insufficient estimation of parameters alone. It follows that future work has to concentrate on detecting closer statistical models of feature vectors of chromosomes in order to further reduce error rates.

Acknowledgements — We thank Dr Jim Piper, Edinburgh, for kindly supplying the Cpr data set and for a stimulating discussion. We thank Professor Peter Kleinschmidt for initiating us to this field of research and for having given us access to an early version of Kleinschmidt, Mitterreiter and Piper [10] which motivated this communication. We also thank Frau Verena Lauren and Herr Erwin Silbereisen for their implementations of several useful algorithms.

References

[1] M.L. Balinski. Signature methods for the assignment problem. *Operations Research*, 33:527–536, 1985.

[2] A.P. Dempster. *Elements of Continuous Multivariate Analysis.* Addison-Wesley, Reading Massachusetts, 1969.

[3] Kai-Tai Fang, Samuel Kotz, and Kai-Wang Ng. *Symmetric Multivariate and Related Distributions.* Chapman and Hall, London, New York, 1990.

[4] K. Fukunaga. *Introduction to statistical pattern recognition.* Academic Press, Boston, San Diego, New York, London, Sydney, Tokyo, Toronto, 1990.

[5] E. Granum. Application of statistical and syntactical methods of analysis and classification to chromosome data. In J. Kittler, K.S. Fu, and L.F. Pau, editors, *Pattern Recognition Theory and Applications (Proc. NATO ASI Series)*, pages 373–398, Dordrecht, 1982. Reidel.

[6] J.D.F. Habbema. A discriminant analysis approach to the identification of human chromosomes. *Biometrics*, 32:919–928, 1976.

[7] J.D.F. Habbema. Statistical methods for classification of human chromosomes. *Biometrics*, 35:103–118, 1979.

[8] S.P.J. Kirby, C.M. Theobald, J. Piper, and A.D. Carothers. Some methods of combining class information in multivariate normal discrimination for the classification of human chromosomes. *Statistics Med.*, 10:141–149, 1991.

[9] P. Kleinschmidt, C.W. Lee, and H. Schannath. Transportation problems which can be solved by use of Hirsch-paths for the dual problem. *Math. Prog.*, 37:153–168, 1987.

[10] Peter Kleinschmidt, Ilse Mitterreiter, and Jim Piper. Improved chromosome classification using monotonic functions of Mahalanobis distance and the transportation method. *ZOR–Math. Meth. Oper. Res.*, 40:305–323, 1994.

[11] C. Lundsteen, A.-M. Lind, and E. Granum. Visual classification of banded human chromosomes. *Ann. Hum. Genet., London*, 40:87–97, 1976.

[12] Jim Piper. The effect of zero feature correlation assumption on maximum likelihood based classification of chromosomes. *Signal Processing*, 12:49–57, 1987.

[13] Jim Piper. Variability and bias in experimentally measured classifier error rates. *Patt. Rec. Lett.*, 13:685–692, 1992.

[14] Jim Piper and Erik Granum. On fully automatic feature measurement for banded chromosome classification. *Cytometry*, 10:242–255, 1989.

[15] Jim Piper, Erik Granum, D. Rutovitz, and H. Ruttledge. Automation of chromosome analysis. *Signal Processing*, 2:203–221, 1980.

[16] R. E. Slot. On the profit of taking into account the known number of objects per class in classification methods. *IEEE Trans. Inf. Theory*, 25:484–488, 1979.

[17] M. Tso, P. Kleinschmidt, I. Mitterreiter, and J. Graham. An efficient transportation algorithm for automatic chromosome karyotyping. *Patt. Rec. Lett.*, 12:117–126, 1991.

[18] M.K.S. Tso and J. Graham. The transportation algorithm as an aid to chromosome classification. *Patt. Rec. Lett.*, 1:489–496, 1983.

[19] S.O. Zimmerman, D.A. Johnston, F.E. Arrighi, and M.E. Rupp. Automated homologue matching of human G-banded chromosomes. *Comput. Biol. Med.*, 16:223–233, 1986.

About the Author — GUNTER RITTER graduated in electrical engineering from Ohm–Polytechnikum in Nuremberg, Germany, in 1962 and in mathematics from the University of Erlangen in 1970. He received the degrees of Dr.rer.nat. and Dr.rer.nat.habil. in 1974 and 1978, respectively, both from the latter university. He was awarded the prize of the faculty of sciences for outstanding doctoral dissertation in 1974. Since 1983, he has been a professor of mathematics at the Department of Mathematics and Computer Science of the University of Passau, Germany, where he holds a chair of mathematics. He has repeatedly been a visiting scientist at the Departments of Mathematics of the University of Washington in Seattle, Washington, and of the University of New South Wales in Sydney, Australia. He has initiated a working group in mathematical knowledge representation and image processing at the University of Passau. He has authored and coauthored a number of research papers in scientific journals in the areas of measure theory, probability theory, statistics, Fourier analysis, and potential theory. His current research and teaching interests are in the areas of measure theory, theoretical and applied probability, statistical methods of image processing, mathematical knowledge representation and learning theory, stochastic optimization, and neural networks.

About the Author — MARÍA TERESA GALLEGOS graduated in statistics in 1988 from Pontificia Universidad Católica in Lima, Peru, and in mathematics in 1992 from the University of Passau, Germany. She won a prize of the German Mathematical Society at the German Student Conference 1992 in Berlin for an outstanding diploma thesis. She is currently an assistant scientist at the Department of Mathematics and Computer Science of the University of Passau. Her research interests are in the areas of statistics and mathematical learning theory, in particular statistical pattern recognition and learning automata.

About the Author — KARL GAGGERMEIER graduated in computer science in 1990 from the University of Passau, Germany. He is currently an assistant scientist at the Department of Mathematics and Computer Science of this university. His research interests are in the areas of game and learning theory, in particular learning automata, and pattern recognition.