Published in *Computational Statistics & Data Analysis* 35 (2001) 351–372

Polarity-free automatic classification of chromosomes[1]

Gunter Ritter[2] and Christoph Pesch

Universität Passau
Fakultät für Mathematik und Informatik
D-94 030 Passau

Abstract: Automatic classification of the chromosomes of a metaphase eukaryotic cell under a light microscope into their biological classes is usually done in three steps: First, their centromeres are estimated in order to find their polarities, next a number of features are extracted from profiles of the oriented chromosomes, and finally the feature sets are assigned to classes. The first step is prone to errors since it is often not easy to detect the centromere. If it is determined on the wrong half of the chromosome then polarity is false leading to erroneous features in the second step and often to a misclassification. We reduce the error rate by applying the recently developed Bayesian method of *variants* to the profiles; this method uses two feature sets for each chromosome, one for each polarity.

We also take another look at feature extraction from profiles further reducing the error rate. Applied to the profiles of the Edinburgh MRC chromosome analysis system the most accurate methods reported here achieve cross-validation error rates below 1%.

Key words: Automatic chromosome classification, karyotyping, diagnostic classification, Bayesian classification, method of variants.

1 Introduction

This communication deals with the problem of *automatic classification of metaphase chromosomes* under a light microscope. As in classification of handwritten numerals or characters, the goal is recognition of *non-standardized* objects in the plane. This means that the objects contain randomness and automation calls for application of methods from data and discriminant analysis.

A normal, nucleated human cell contains 46 chromosomes consisting of 22 matching pairs of homologous autosomal chromosomes and two sex chromosomes. The autosomal classes are numbered 1..22 and the sex chromosomes are X X in female and X Y in male cells, respectively. During the metaphase of the cycle of cell division the chromosomes appear as separated objects and can, after suitable preparation, be observed under a light microscope, cf. Figure 1. This allows the detection of numerical and structural anomalies of the cell.

Optical distinction of the 24 classes $1, \ldots, 22, X$ and Y is possible since the introduction of suitable staining techniques in the 1970's. Staining displays a band pattern along the chromosome axis characteristic of its class, cf. the schematic representation of chromosome classes shown in Figure 2.

[1] Supported by Deutsche Forschungsgemeinschaft, Ri477/4

[2] Corresponding author. E-mail address: ritter@fim.uni-passau.de

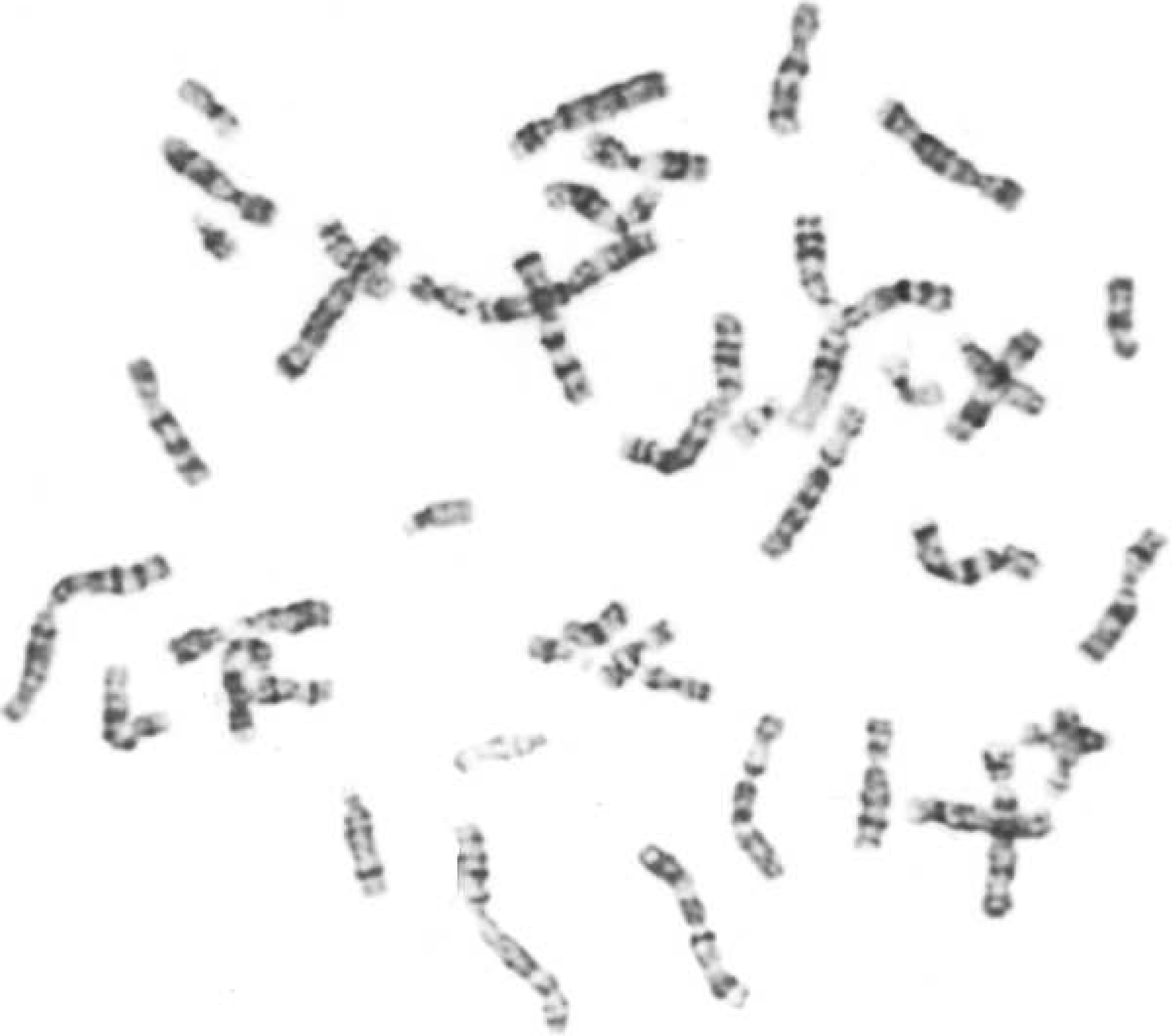

Figure 1: Metaphase human cell under a light microscope

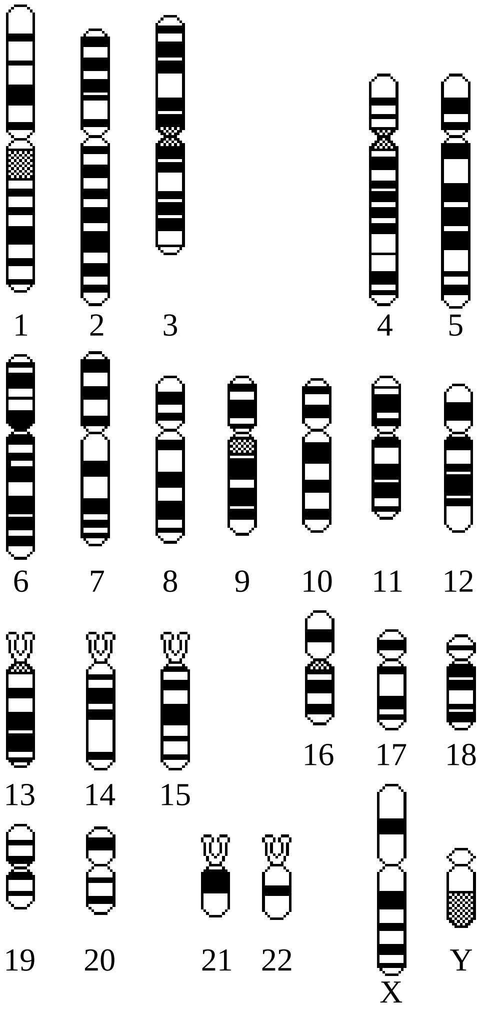

Figure 2: Schematic representation of chromosome classes and band structures

The *centromere* of a chromosome is the region where the two longitudinal halves (chromatids) of the double-stranded chromosome are joined together. It is usually visible as a constriction. The *polarity* of a chromosome representative of its class is defined so that, if put in vertical direction, its centromere is located on its upper half. The parts above and below the centromere are then called the *short* and the *long arms*, respectively. A representation of the chromosomal complement of a cell showing class structure and polarity is called a *karyotype*, cf. Figure 3.

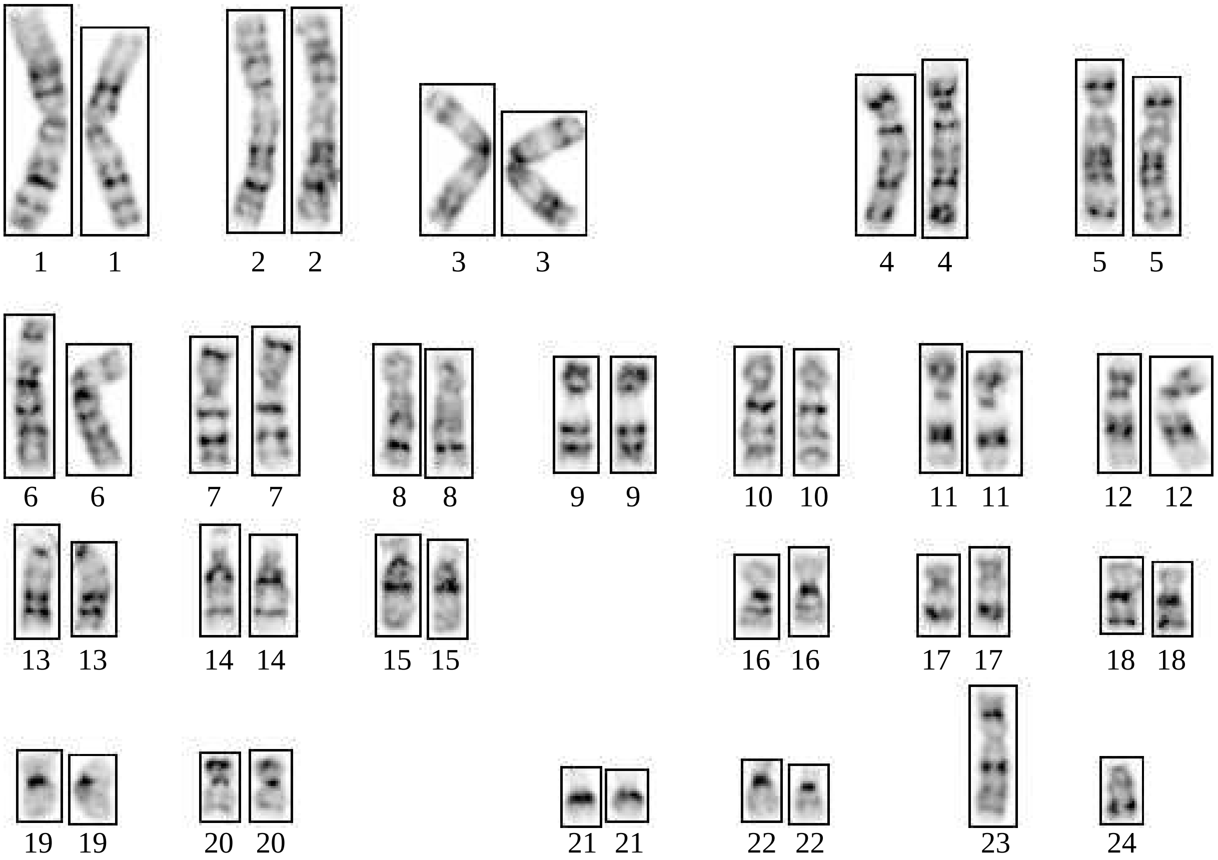

Figure 3: The karyotype of a male human cell

Producing a karyotype from an image of a metaphase cell is a complex, clear-cut, and highly specific image recognition task which is mainly based on chromosome size, the position of the centromere, and the band pattern. It is performed by a human expert at the speed of 15 min/cell and, depending on the quality of the cell image, at the low error rate between one and three misclassifications among 1000 chromosomes, cf. Lundsteen et al., [10], p. 88, Granum [5], p. 380, and Kleinschmidt et. al. [8], p. 314; this amounts to roughly 2–7% of misclassified cells. First attempts to automate chromosome classification date back to the early 1960's, cf. Ledley and Ruddle [9]. The problem was later taken up by many working groups. One objective of our research is the design of statistical algorithms for this classification task; in particular, we study to what extent algorithms can approach human performance in this field. This is a question of artificial intelligence. As such, the problem bears interesting aspects of high-dimensional multivariate statistics and of combinatorial optimization.

A reduction of the error rate of automatic chromosome classification has also significance for cytogenetics. In order to facilitate the detection of structural anomalies in a eukaryotic cell, the operator usually starts an analysis by arranging its metaphase chromosomes in a karyotype.

Automation of this process at an error rate close to the expert's would eliminate much repetitive work in chromosome analysis. For more information on the biological and cytogenetical background of karyotyping we refer the interested reader to the monographs [24] and [6], Chapter 8, to the survey article [14], and to [7] and [16], Section 3.1, for brief overviews.

Automatic classification of chromosomes on the basis of microscopic grey-level images of stained metaphase cells usually follows a number of consecutive steps.

(a) Removal of background noise, stains, and other objects in the image different from chromosomes.

(b) Segmentation of the image into its individual chromosomes and the background.

(c) Extraction of primitive numeric features like area, total "mass", and mean grey value from each chromosome.

(d) Recognition of the oblong shape. This is done by constructing an axis along each chromosome.

(e) Estimation of the location of the centromere for determining polarity.

(f) Representation of band pattern and shape by so-called *profiles*, i.e., univariate functions along the axis. Piper and Granum [13] describe three kinds of profiles: the *density* profile, a representation of the grey-value distribution along the axis, the modulus of its "derivative," called the *gradient* profile, and the so-called *shape* profile. Figure 4 shows a chromosome, its axis, and the associated *density* profile.

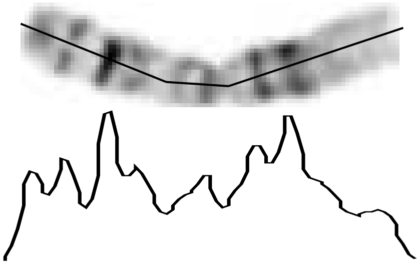

Figure 4: A chromosome with its axis and its density profile

(g) Extraction of numerical features from the profiles by integration against *weight functions.* Piper and Granum [13] use, e.g., piecewise linear functions.

(h) Normalization of features, cf. Section 2.

(i) Selection and identification of a suitable statistical model of feature sets by methods of statistical data analysis.

(j) Classification of chromosomes in their biological classes on the basis of their feature sets, cf. Section 3.

The steps (a) and (b) are often carried out interactively. Details on the steps (c), (d), (e), and (f) are found in [13]. Step (i) was the subject of close investigation in [19], [17], and [16]. The present communication contributes to the remaining steps (g), (h), and (j).

A number of paradigms are nowadays in use for supervised classification. They are the *maximum likelihood* (ML) discriminant rule, the *Bayesian* paradigm, *discriminant functions* such as *linear* *forms*, *polynomials* or *neural networks*, *support vector machines*, *k-nearest-neighbor* classifiers, *partitioning classification*, *decision trees*, and *ensemble classifiers* (*mixtures of experts*). The special appeal of the Bayesian paradigm is its mathematically proven optimality. As in the earlier communications [17], and [16] we prefer to use Bayesian discriminant analysis in the form of the maximum-a-posteriori (MAP) classifier here, too.

For assigning chromosomes to biological classes, discriminant analysis can be done in the *context-free* or the *context-sensitive* (*constrained*) way. In the former, earlier, approach the individuals classified are chromosomes in isolation. Here, the possible allocations of one cell are in a one-to-one correspondence with the more than $3 \cdot 10^{63}$ mappings $1..46 \rightarrow 1..24$. The latter approach takes into account the known number of chromosomes in each class reducing this number to $46!/2^{23} + 46^2 \cdot 44!/2^{22} \approx 2 \cdot 10^{51}$ assignments and yielding, therefore, better results. Using this prior information in this straightforward way was suggested already by Habbema [7] and Slot [21]. Tso and Graham [22] describe the first efficient algorithm for constrained ML-classification into 10 "Denver" groups[3] pointing out that the optimization problem involved is a Hitchcock transportation problem. The authors note that it can also be applied to cells with missing or extra chromosomes such as trisomies.

Context-sensitive classification needs the *joint* distribution of the feature vectors of all chromosomes in a cell. We adopt the assumption of *independence* of feature vectors made by many research groups so that the joint distribution is the product of the single distributions. The optimality of Bayes' method can only be exploited if accurate distributional models are employed. The classical "normal" model makes the assumption that feature vectors are *normally* distributed. However, the challenge in automatic chromosome classification today is proper outlier handling. Therefore, besides the classical model, a number of well-known and new parametric distributional families were employed [19], [17], and [16]. These are the *elliptically contoured* (*elliptically symmetric*) [3] and the *quadratically asymmetric* [15] families and mixture models with outliers based upon them. Quadratic asymmetry is an extension of elliptical symmetry. Whereas the latter reflects location, scatter, and heavy or light tailedness of data, quadratic asymmetry also models asymmetries. Moreover, a robust method of covariance estimation proposed by Piper, Poole, and Carothers [14], called *covariance weighting*, was applied. This led to a series of robust statistical classifiers for automatic chromosome classification, named IEC, IECO, IQA, and IQAO. In all acronyms, the letter "I" stands for the assumption of statistical **i**ndependence of chromosomes, "EC" means **e**lliptically **c**ontoured and "QA" **q**uadratically **a**symmetric, and "O" stands for a mixture model with **o**utliers. The most accurate of these classifiers is IQAO; all the others are special cases of it.

Applied to the Edinburgh features [13] of the large Copenhagen image data set Cpr [12], [17] consisting of 2804 metaphase human cells the classifier IQAO achieves a test-set (cross-validation) error rate of 1.22% with respect to all chromosomes in the data set, [16]. To our knowledge, this is the lowest previously achieved error rate. It means a reduction by a factor of 2.4 compared with the classifier based on the normal model; the improvement is mainly due to the more precise data model. The error rate includes in particular trisomy cells and cells with missing chromosomes. These classification results are based solely on *feature* data and do not presuppose any knowledge about their origin.

All profile-dependent features based on *asymmetric* weight functions, such as odd cosines, are sensitive to polarity of chromosomes since profiles are asymmetric. The degree of asymmetry depends on chromosome class, cf. Figure 5. Therefore, the customary approach to automatic chromosome classification needs the position of the centromere at an early stage. However,

[3] The Denver groups [2] are subgroups of the 24 classes distinguishable by size and position of centromere; they were in use before introduction of staining techniques.

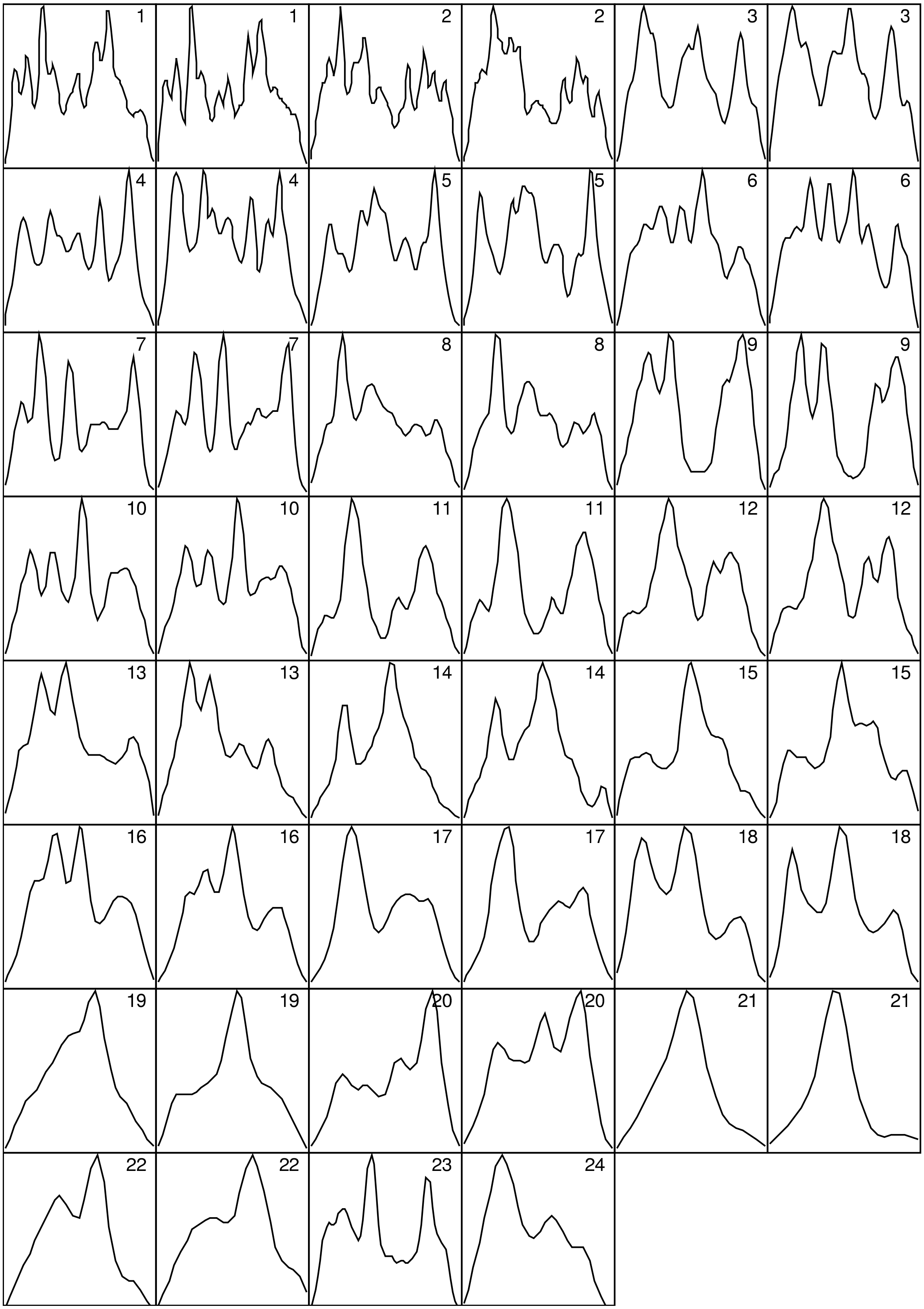

Figure 5: Density profiles of the chromosomes shown in Figure 3

sometimes the centromeric constriction is not clearly visible, cf. [13]. The reason may be other constrictions along the chromosome; moreover, in some classes such as the first, the centromere is almost in the middle of the chromosome making it hard to decide from the shape of the chromosome alone, on which side it lies. One weakness of the approach proposed in [13] and of the classifiers IEC, IECO, IQA, and IQAO mentioned above is the fact that they rely on polarity for feature extraction. We will call such classifiers *polarity dependent*. In fact, one source of classification errors is estimation of centromeres on the wrong side of the chromosomes thus giving rise to erroneous features. According to [13], polarity errors amount to between 5 and 10% in the mean and rise as high as 30% in class 1.

Instead of trying to guess the position of the centromere prematurely from the chromosome's shape, thereby estimating its polarity, we use here a general concept for handling *ambiguities* of objects set forth in Ritter and Gallegos [18]: *variants*, cf. Section 3. In the present case, ambiguous interpretations of a chromosome result from its unknown polarity, the two possible polarities of the profile giving rise to two competing feature sets which are the variants here. In Gallegos and Ritter [4] it is investigated under what conditions optimal *constrained* classification is possible despite the presence of variants. This theory can be applied to polarities. In Section 3, we develop two such algorithms; they employ both variants in a symmetric way. For each statistical model above they give rise to Bayesian classifiers in the presence of **v**ariants. We call the new classifiers *polarity free* and name them VIEC, VIECO, VIQA, and VIQAO, respectively.

We take also another look at feature extraction from profiles using cosines and at normalization, cf. Section 2. Test-set (cross-validation) and training-set error rates based on the new features demonstrate the superior performance of the new classifiers and features, Section 4. In fact, to the best of our knowledge, this paper presents the first automatic classifiers achieving cross-validation error rates below 1% for classification of human chromosomes into their 24 classes on the basis of a *correctly segmented*, everyday, clinical image data set without manual orientation of chromosomes, cf. Table 3(a).

2 Features and their normalization

The results obtained in [19], [17], and [16] were based on 24 of the 30 features of the Edinburgh MRC chromosome analysis system; 28 of them are described in [13], two more were added later on. These 24 features are size (a mixture of area and length), density = mass/area, area centromeric index = the area of the long arm divided by the total area of the chromosome, coefficient of variation of the density distribution (cvdd), normalized square root of squared density differences (nssd), number of bands = number of density profile maxima, and six so-called wdd-features (weighted density distribution) from each of the three profiles. The remaining 6 features are either redundant or contain too many outliers to be useful.

Whereas we use again basically the features described in [13] we propose some modifications of which we show in Table 3(b) that they reduce classification error rates. First, we truncate each density profile at both ends at 27.5% of the mean value of all density profiles in the data set and subtract the minimum of the resulting profile. This shortening of the tips of profiles results in a better centering.

Let us describe next how profiles are converted to features. For the sake of clarity we assume first that a profile is a function $[0,1] \to \mathbb{R}$ defined on the unit interval. Let $p_1, \ldots, p_m$ be the raw profiles of one kind (density, gradient, or shape) of all chromosomes $1\,..\,m$ in a cell. (In a normal human cell $m = 46$.) Effects of culture, staining, and microscopy cause the images to be darker or lighter. Therefore, the profiles contain a (random) factor independent of the chromosomes in the cell (but depending on the cell). This random "darkness index" does not contain information on the chromosomes. Since it increases the scatter and correlates the chromosomes, it has to be removed. This is done by normalizing the sum of the integrals of all profiles p_1, …, p_m to 1. Unfortunately, this normalization introduces new correlation between chromosomes. By the strong law of large numbers, which is already applicable to the number $m = 46$ of chromosomes, the normalizing sum depends only weakly on the chromosomes (however, it depends on the cell). Therefore, normalization renders the chromosomes only weakly correlated and the assumption of independence is barely hurt. Let $\bar{p}_1, \dots, \bar{p}_m \colon [0,1] \to \mathbb{R}$ be the so normalized profiles.

Now consider an arbitrary profile p among p_1, …, p_m and its normalization $\bar{p}$. A characteristic quantity leading to a first feature is the integral of $\bar{p}$,

$$\int_0^1 \bar{p}(t)\,\mathrm{d}t = \frac{\int_0^1 p(t)\,\mathrm{d}t}{\sum_{j=1}^n \int_0^1 p_j(t)\,\mathrm{d}t}. \tag{2.1}$$

Turning to the extraction of band-pattern information from $\bar{p}$ we normalize $\bar{p}$ across the chromosome by putting

$$\overline{\overline{p}} = \frac{\bar{p}}{\int_0^1 \bar{p}(t)\,\mathrm{d}t} = \frac{p}{\int_0^1 p(t)\,\mathrm{d}t}.$$

Attaching a copy of $\overline{\overline{p}}$ to $\overline{\overline{p}}$ in reverse direction yields a symmetric profile $\overline{\overline{\overline{p}}} \colon [-1,1] \to \mathbb{R}$

$$\overline{\overline{\overline{p}}}(t) = \begin{cases} \overline{\overline{p}}(t), & \text{if } t \geq 0, \\ \overline{\overline{p}}(-t), & \text{otherwise,} \end{cases}$$

which may be considered as an absolutely continuous, 2-periodic function on $\mathbb{R}$. The Fourier coefficients $\mathcal{F}_k$ of $\overline{\overline{\overline{p}}}$ are

$$\begin{aligned} \mathcal{F}_k \overline{\overline{\overline{p}}} &= \frac{1}{2} \int_{-1}^1 \overline{\overline{\overline{p}}}(t) \exp(\pi i k t)\,\mathrm{d}t = \int_0^1 \overline{\overline{p}}(t) \cos(\pi k t)\,\mathrm{d}t \\ &= \frac{\int_0^1 p(t) \cos(\pi k t)\,\mathrm{d}t}{\int_0^1 p(t)\,\mathrm{d}t}, \qquad k = 1, 2, \dots. \end{aligned} \tag{2.2}$$

Like p, $\bar{p}$, $\overline{\overline{p}}$, and $\overline{\overline{\overline{p}}}$ these Fourier coefficients, too, contain the whole information on the band pattern. By absolute continuity, they are of order $o(k^{-1})$ as $k \longrightarrow \infty$.

In practice we are given the discrete profiles h_i defined by $h_i(\ell) = \int_{\ell/L_i}^{(\ell+1)/L_i} p_i(t)\,\mathrm{d}t$, $0 \leq \ell < L_i$, $1 \leq i \leq m$, where L_i is the length of h_i. Letting h be an arbitrary discrete profile among h_1, …, h_m, letting L be its length, and discretizing (2.1) and (2.2) we obtain the features

$$a = \frac{\frac{1}{L} \sum_{\ell < L} h(\ell)}{\sum_j \frac{1}{L_j} \sum_{\ell < L_j} h_j(\ell)} \tag{2.3}$$

and

$$a_k = \frac{\sum_{\ell < L} h(\ell) \cos\left(\pi k (\ell + \frac{1}{2})/L\right)}{\sum_{\ell < L} h(\ell)}, \qquad k = 1, 2, \dots. \tag{2.4}$$

By uniqueness of the (discrete) Fourier transform, the discrete and normalized profile values $h(k)/\sum_{\ell<L} h(\ell)$ are linear functions of the coefficients a_1, …, a_{L-1}; hence, so are the coefficients a_L, a_{L+1}, … and it suffices to consider the features a_1, …, a_{L-1}. In fact, because of noise contained in the profiles, only the first few of them are useful.

We extract 29 features from each chromosome which we now describe. Let us denote the truncated (discrete) density profile of chromosome i by d_i, by g_i its gradient profile, and by

m_i its shape profile. Let L_i be the length of the density profile, let $D_i = \sum_{\ell<L_i} d_i(\ell)$, $G_i = \sum_{\ell<L_i-1} g_i(\ell)$, $M_i = \sum_{\ell<L_i} m_i(\ell)$, and let $D'_i = D_i/L_i$, $G'_i = G_i/(L_i-1)$, $M'_i = M_i/L_i$. The first four features of the following list are adopted from [13], the remaining ones are modifications of features presented there.

1. The *size* of the chromosome. This is the mean of its area normalized at the 60% fractile across the cell and its length, again normalized at the 60% fractile.
2. The *density* of the chromosome. This is the quotient of the sum of the grey values over all its pixels divided by its area; the sum of these quotients is normalized to 1 across the cell.
3. The *number of density maxima* relative to the cell.
4. $nssd = \sqrt{\sum_{\ell<L_i-1} g_i^2(\ell)}/D_i$.
5. The mean of the density profile normalized across the cell, $D'_i/\sum_j D'_j$, cf. (2.3).
6. The mean of the gradient profile normalized across the cell, $G'_i/\sum_j G'_j$, cf. (2.3).
7. The mean of the shape profile normalized across the cell, $M'_i/\sum_j M'_j$, cf. (2.3).

8. – 13. The first six cosine coefficients of the density profile,
$\frac{1}{D_i}\sum_{\ell<L_i} d_i(\ell)\cos\big(\pi k(\ell+\frac{1}{2})/L_i\big)$, $k \in 1..6$, cf. (2.4).

14. – 21. The first eight cosine coefficients of the gradient profile,
$\frac{1}{G_i}\sum_{\ell<L_i-1} g_i(\ell)\cos\big(\pi k(\ell+\frac{1}{2})/(L_i-1)\big)$, $k \in 1..8$, cf. (2.4), standardized as described below.

22. – 29. The first eight cosine coefficients of the shape profile,
$\frac{1}{M_i}\sum_{\ell<L_i} m_i(\ell)\cos\big(\pi k(\ell+\frac{1}{2})/L_i\big)$, $k \in 1..8$, cf. (2.4).

The numbers of cosines used were determined by calibration. Note that the odd cosine coefficients depend on the polarity of the profile. In the case of polarity-free classification, we consider both polarities of the chromosome. Hence, for each polarity, we obtain a set of 29 features. These two sets are handled by the method of variants, described below in Section 3.

In polarity-dependent estimation, the eight cosine coefficients of the gradient profiles are standardized across the cell, i.e. the mean value of *all* coefficients in the cell is subtracted and the result is divided by the standard deviation. Polarity-free estimators need two sets of features for each chromosome. Which set corresponds to the correct polarity is not known at this stage in the *classification* process. (It is, of course, known and used in the process of parameter estimation.) Therefore, we can standardize only those features which are identical in both sets. These are the (symmetric) even cosine coefficients. The odd coefficients are left unchanged.

Finally, features are rescaled to a handy size for better numerical control. The centromeric index was never explicitly used as a feature, not even in the case of polarity-dependent classifiers, since it generates too many outliers. (The polarity-dependent classifiers use it, of course, for the sake of orientation.) It should, however, be pointed out that the centromeric index is contained implicitly in the shape profile and, hence, in the eight features extracted from it. This concludes the description of features.

3 Variants

In Ritter and Gallegos [18], variants are introduced as a Bayesian model for handling ambiguities in recognition and identification problems. Formally, variants are different observations of the same object leading to different interpretations. In general, each object may have its own number of variants, one of which is the *regular* variant yielding the correct interpretation. The other variants can be understood as perturbed observations of the object. In the present application, the variants are the two feature sets resulting from the two possible polarities of the chromosome. The regular variant corresponds to the correct, biologically defined polarity. Unfortunately, as in the present case, it is often unknown which of the variants is the regular one. Then, selection of the regular variant and classification of the object in the presence of variants both are interesting problems.

3.1 The Simple Constrained Classifiers

We establish two algorithms for optimal polarity-free, constrained classification of chromosomes using the methods of [4]. In order to ease the exposition we restrict matters to classification of normal human cells, i.e., human cells without missing or extra chromosomes; other species and cells with numerical anomalies can be handled in a similar way, cf. also [17] and [16]. Because of the pairwise homologies stated at the beginning of the introduction we follow [17] and [16] introducing *virtual* classes. These are defined by doubling homologous classes. Normal female and male human cells can be handled by introducing the virtual classes $1..47$; the virtual classes j and $j+23$, $j \in 1..22$, are identified with the biological class j, the virtual classes 23 and 46 with the biological class X, and the virtual class 47 with the biological class Y.

In the present case, we observe 46 chromosomes (objects) $1..46$ belonging to 47 (virtual) *classes* $1..47$ with the identifications stated above. No class is covered more often than once. Our state space E is the Cartesian product $\mathbb{R}^d$, $d = 29$ being the number of features. Let $Z_{j,1}$, $Z_{j,2}\colon (\Omega, P) \to E$, $j \in 1..47$, be the two random *variants* (polarities) of an object of (virtual) class j, $Z_{j,1}$ being the regular variant. Let $\mathcal{I}_{47} \subseteq \mathcal{S}_{47}$ denote the set of permutations σ of $1..47$ such that $\sigma(47) \in \{46, 47\}$. We represent an assignment of chromosomes to (virtual) classes by an element $\sigma \in \mathcal{I}_{47}$. An object $i \in 1..46$ is assigned to class $\sigma(i)$ and $\sigma(47)$ is the indicator of sex: $\sigma(47) = 47$ if the cell is female ("σ is female") and $\sigma(47) = 46$, otherwise ("σ is male"). We observe a realization $\mathbf{x} = (x_{1,1}, x_{1,2};\ x_{2,1}, x_{2,2};\ \ldots;\ x_{46,1}, x_{46,2}) \in E^{2\cdot 46}$ of the random array

$$\begin{aligned} X &= (X_{1,1}, X_{1,2};\ X_{2,1}, X_{2,2};\ \ldots;\ X_{46,1}, X_{46,2}) \\ &= (Z_{\Phi(1),V_1}, Z_{\Phi(1),3-V_1};\ \ldots;\ Z_{\Phi(46),V_{46}}, Z_{\Phi(46),3-V_{46}}) \\ &=: Z_{\Phi,V}. \end{aligned}$$

Here, $x_{i,h}$ and $X_{i,h}$ stand for the observation at *site* $h \in 1..2$ of the object $i \in 1..46$, $\Phi\colon (\Omega, P) \to \mathcal{I}_{47}$ is a random assignment, and $V_i\colon (\Omega, P) \to 1..2$ stands for the unknown variant at the first site of object i. We also let $H_i\colon (\Omega, P) \to 1..2$ denote the (unknown) site of the regular variant (correct polarity) of object i. We have plainly $H_i = V_i$ here. We assume that the pairs $(Z_{1,1}, Z_{1,2})$, …, $(Z_{47,1}, Z_{47,2})$, the random permutation Φ, and the sites H_1, …, H_{46} are all statistically independent.

The following equivalence relation will take care of homologous classes. Let us call two permutations φ, $\sigma \in \mathcal{I}_{47}$ *equivalent* or *homologous*, $\varphi \sim \sigma$, if they assign all chromosomes to the same *biological* classes, i.e., if either both permutations are female and $\varphi(i) = \sigma(i) \mod 23$ for all i *or* if they are both male and $\varphi(i) = \sigma(i) \mod 23$ for all i such that $\varphi(i) \in 1..22 \cup 24..45$ and $\varphi(i) = \sigma(i)$ for $\varphi(i) \in \{23, 47\}$.

The following statistical model is designed to optimally estimate the correct classes *without* explicitly estimating the polarities, too. Its parameter set is $\Theta = \mathcal{I}_{47} \times (1\,.\,.2)^{46}$ and its decision set is $\mathcal{I}_{47}$. It is defined as the quadruplet

$$((X_{1,1}, X_{1,2};\ \ldots;\ X_{46,1}, X_{46,2}), (P_\vartheta)_{\vartheta\in\Theta}, \mathcal{I}_{47}, G),$$

where $P_{(\varphi,\mathbf{v})}$, $(\varphi, \mathbf{v}) \in \Theta$, is the probability law of

$$(Z_{\varphi(1),v_1}, Z_{\varphi(1),3-v_1};\ \ldots;\ Z_{\varphi(46),v_{46}}, Z_{\varphi(46),3-v_{46}})$$

and where the gain G is defined by

$$G((\varphi, \mathbf{v}), \sigma) := \begin{cases} 1, & \text{if } \varphi \sim \sigma, \\ 0, & \text{otherwise.} \end{cases}$$

That is, there is a gain of one unit for a biologically correct assignment and no gain otherwise. The Bayesian estimator for the statistical decision model above is the *maximum-a-posteriori* (MAP) classifier

$$\begin{aligned} \mathrm{MAP}(\mathbf{x}) &= \operatorname*{argmax}_{\sigma\in\mathcal{I}_{47}} E[G((\Phi, V), \sigma)/X = \mathbf{x}] \\ &= \operatorname*{argmax}_{\sigma\in\mathcal{I}_{47}} P[\Phi \sim \sigma/Z_{\Phi,V} = \mathbf{x}] \\ &= \operatorname*{argmax}_{\sigma\in\mathcal{I}_{47}} P[Z_{\Phi,V} \in \mathrm{d}\mathbf{x}/\Phi \sim \sigma]P[\Phi \sim \sigma]. \end{aligned} \tag{3.5}$$

The distributions of $Z_{\varphi,V}$ and $Z_{\sigma,V}$ are equal if $\varphi \sim \sigma$. Hence, by independence, (3.5) reads also

$$\begin{aligned} \mathrm{MAP}(\mathbf{x}) &= \operatorname*{argmax}_{\sigma\in\mathcal{I}_{47}} P[Z_{\sigma,V} \in \mathrm{d}\mathbf{x}]P[\Phi \sim \sigma] \\ &= \operatorname*{argmax}_{\sigma\in\mathcal{I}_{47}} \left(\prod_{i=1}^{46} P[(Z_{\sigma(i),V_i}, Z_{\sigma(i),3-V_i}) \in (\mathrm{d}x_{i,1}, \mathrm{d}x_{i,2})]\right) P[\Phi \sim \sigma]. \end{aligned} \tag{3.6}$$

Again by independence, the generic factor in the last product is

$$\begin{aligned} &P\left[Z_{\sigma(i),V_i} \in \mathrm{d}x_{i,1}, Z_{\sigma(i),3-V_i} \in \mathrm{d}x_{i,2}\right] \\ =&P\left[Z_{\sigma(i),1} \in \mathrm{d}x_{i,H_i}, Z_{\sigma(i),2} \in \mathrm{d}x_{i,3-H_i}\right] \\ =&\sum_{h=1}^{2} P\left[Z_{\sigma(i),1} \in \mathrm{d}x_{i,h}, Z_{\sigma(i),2} \in \mathrm{d}x_{i,3-h}\right] P[H_i = h] \\ =&\sum_{h=1}^{2} P\left[Z_{\sigma(i),1} \in \mathrm{d}x_{i,h}\right] P[H_i = h] \end{aligned}$$

since $x_{i,1}$ and $x_{i,2}$ come from the two polarities of the same chromosome. Let $q_{i,h} = P[H_i = h]$, the prior probability of $h \in 1\,.\,.2$ to be the regular polarity of object i, let $f_j(x) = P[Z_{j,1} \in \mathrm{d}x]/\mathrm{d}x$, $x \in E$, the density function of the regular polarity of class j, and let p_f (p_m) be the prior probability of a cell to be female (male). Assuming that Φ is uniformly distributed conditional on being female or male, we have

$$P[\Phi \sim \sigma] = \begin{cases} P[\Phi \sim \sigma/\Phi \text{ is female}]\, p_f = \frac{2^{23}}{46!} p_f\ , & \text{if } \sigma \text{ is female,} \\ P[\Phi \sim \sigma/\Phi \text{ is male}]\, p_m = \frac{2^{22}}{46!} p_m\ , & \text{if } \sigma \text{ is male.} \end{cases}$$

In particular, $P[\Phi \sim \sigma]$ depends on the indicator of sex, $\sigma(47)$, only, i.e.,

$$P[\Phi \sim \sigma] = \frac{2^{23}}{46!}\alpha_{\sigma(47)} \tag{3.7}$$

with $\alpha_{46} = p_m/2$ and $\alpha_{47} = p_f$. With these notations, the classifier (3.6) reads

$$\begin{aligned} \mathrm{MAP}(\mathbf{x}) &= \operatorname*{argmax}_{\sigma\in\mathcal{J}_{47}} \alpha_{\sigma(47)} \prod_{i=1}^{46}\sum_{h=1}^{2} f_{\sigma(i)}(x_{i,h})q_{i,h} \\ &= \operatorname*{argmax}_{\sigma\in\mathcal{S}_{47}} \prod_{i=1}^{47} d_{i,\sigma(i)} \end{aligned} \tag{3.8}$$

with the coefficients

$$d_{i,j} = \begin{cases} \sum_{h=1}^{2} f_j(x_{i,h})q_{i,h}\,, & i \in 1\,..46,\ j \in 1\,..47, \\ \alpha_{46} = p_m/2\,, & i = 47,\ j = 46, \\ \alpha_{47} = p_f\,, & i = j = 47, \\ 0\,, & i = 47,\ j < 46. \end{cases}$$

It has been known for some time that constrained, context-sensitive ML-classification of independent chromosomes of a cell leads to a *linear assignment* problem; cf. [22], [23]. Ritter and Gallegos [17] and Ritter and Gaggermeier [16] extended this to MAP-classification. (Their approach differs somewhat from the present one since they disregarded homologies in the gain function.) Taking negative logarithms in the expression (3.8) we obtain a linear assignment problem also in the presence of variants. Its weights are shown in Table 1.

Table 1: Weight matrix of size 47 × 47 for polarity-free assignment of 46 chromosomes to their classes *without* estimation of polarities. The density function f_j and the probabilities $q_{i,h}$, p_m, and p_f are explained in the paragraph before Eqn. (3.7); $x_{i,h}$ denotes variant h of chromosome i.

chromosomes $i =$			
1 . . 46	$-\ln\sum_{h=1}^{2} f_j(x_{i,h})q_{i,h}$		
sex ind.	∞	$-\ln\frac{p_m}{2}$	$-\ln p_f$
classes $j =$	1 . . 22 X 1 . . 22	X	Y

We have thus derived algorithm SCC (*Simple Constrained Classifier*) which realizes the MAP-classifier for assigning each object to its class *without* estimating its polarity. It uses the density function of the *regular* variant, only; hence the adjective "simple". The linear assignment problems contained in this and the following algorithm can be efficiently solved by the Hungarian method, cf. Papadimitriou and Steiglitz [11], or by Balinski's algorithm [1].

Algorithm SCC

Input: Density functions f_j of the regular variants of all (virtual) classes $j \in 1\,.\,.47$;
prior probabilities p_f, p_m, and $q_{i,h}$ for all objects i, and all sites h;
the observation $\mathbf{x} = (x_{i,h})_{\substack{i\in 1\,.\,.46\\ h\in 1\,.\,.2}}$.

Output: Classification of objects into their classes.

```
begin
  foreach $(i,j) \in 1\,.\,.46 \times 1\,.\,.47$ do
    $c_{i,j} := \begin{cases} -\ln \sum_{h=1}^{2} f_j(x_{i,h}) q_{i,h}\,, & i \in 1\,.\,.46,\ j \in 1\,.\,.47, \\ -\ln(p_m/2)\,, & i = 47,\ j = 46, \\ -\ln p_f\,, & i = j = 47, \\ \infty\,, & i = 47,\ j < 46. \end{cases}$
  od
  $\hat\sigma := \operatorname*{argmin}_{\sigma\in\mathcal{S}_{47}} \sum_{i=1}^{47} c_{i,\sigma(i)}$;    (*A linear assignment problem *)
  return $\hat\sigma_{|1\,.\,.46}$;
end
```

If a *karyotype* is to be produced then the Algorithm SCC is not sufficient since, in this case, the polarities, too, are required. Therefore, we also include Algorithm SCCS (*Simple Constrained Classifier-Selector*) which realizes the MAP-estimator for estimating class *and* polarity. This time, one uses the decision set $\mathcal{I}_{47} \times (1\,.\,.2)^{46}$, the second factor standing for the sites of the regular variants. The gain function G is defined by

$$G((\varphi, \mathbf{v}), (\sigma, \mathbf{h})) := \begin{cases} 1, & \text{if } \varphi \sim \sigma \text{ and } \mathbf{v} = \mathbf{h}, \\ 0, & \text{otherwise.} \end{cases}$$

That is, there is a gain of one unit for a biologically correct assignment with all polarities correct and no gain otherwise. A similar reasoning as above shows that the MAP-estimator is again based on the solution of a linear assignment problem. Indeed, by independence

$$\begin{aligned} &E[G((\Phi, V), (\sigma, \mathbf{h}));\ X \in \mathrm{d}\mathbf{x}] \\ =&P[\Phi \sim \sigma, V = \mathbf{h}, Z_{\Phi,V} \in \mathrm{d}\mathbf{x}] \\ =&P[Z_{\Phi,V} \in \mathrm{d}\mathbf{x}/\Phi \sim \sigma, V = \mathbf{h}]P[\Phi \sim \sigma]P[V = \mathbf{h}] \\ =&P[Z_{\sigma,\mathbf{h}} \in \mathrm{d}\mathbf{x}]P[\Phi \sim \sigma]P[H = \mathbf{h}] \\ =&\prod_{i=1}^{46} P[Z_{\sigma(i),1} \in \mathrm{d}x_{i,h_i}]P[\Phi \sim \sigma] \prod_{i=1}^{46} P[H_i = h_i], \end{aligned}$$

Together with (3.7) it follows that the Bayesian estimator for the statistical decision model above is the MAP-classifier

$$\mathrm{MAP}(\mathbf{x}) = \operatorname*{argmax}_{\sigma,\mathbf{h}} \alpha_{\sigma(47)} \prod_{i=1}^{46} f_{\sigma(i)}(x_{i,h_i}) q_{i,h_i}.$$

The optimal permutation $\hat{\sigma}$ is the maximizer of the expression

$$\alpha_{\sigma(47)} \prod_{i=1}^{46} \max\{f_{\sigma(i)}(x_{i,1})q_{i,1}, f_{\sigma(i)}(x_{i,2})q_{i,2}\}.$$

By taking negative logarithms, this becomes again a linear assignment problem. The resulting weights are shown in Table 2. The site $\hat{h}_i$ of the regular variant of object i is the maximizer of $f_{\hat{\sigma}(i)}(x_{i,h})q_{i,h}$, $h \in 1..2$.

Table 2: Weight matrix of size 47×47 for polarity-free assignment of 46 chromosomes to their classes *with* estimation of polarities. The density function f_j and the probabilities $q_{i,h}$, p_m, and p_f are explained in the paragraph before Eqn. (3.7); $x_{i,h}$ denotes variant h of chromosome i.

$$\begin{array}{l r|c|c|c|}
\cline{3-5}
 & 1 & & & \\
\text{chromosomes } i = & \cdot & & & \\
 & & \multicolumn{3}{c|}{-\ln\max_{h\in 1..2} f_j(x_{i,h})q_{i,h}} \\
 & \cdot & & & \\
 & 46 & & & \\
\hline
\text{sex ind.} & & \infty & -\ln\frac{p_m}{2} & -\ln p_f \\
\cline{3-5}
\text{classes } j = & & 1\ \ .\ \ .\ \ 22\ \ \mathrm{X}\ \ 1\ \ .\ \ .\ \ 22 & \mathrm{X} & \mathrm{Y}
\end{array}$$

The following algorithm represents the MAP-estimator of the class assignment and the sites of the regular variants given the observation $\mathbf{x} = (x_{i,h})_{\substack{i\in 1..46 \\ h\in 1..2}}$.

Algorithm SCCS

Input: Density functions f_j of the regular variants of all classes $j \in 1..47$;
prior probabilities p_f, p_m and $q_{i,h}$ of all objects i, and all sites h;
the observation $\mathbf{x} = (x_{i,h})_{\substack{i\in 1..46 \\ h\in 1..2}}$.

Output: Classification of objects into their classes and the sites of their regular variants.

```
begin
  foreach $(i, j) \in 1..46 \times 1..47$ do
        $c_{i,j} := \begin{cases} -\ln\max_{h\in 1..2} f_j(x_{i,h})q_{i,h}\,, & i \in 1..46,\ j \in 1..47, \\ -\ln(p_m/2)\,, & i = 47,\ j = 46, \\ -\ln p_f\,, & i = j = 47, \\ \infty\,, & i = 47,\ j < 46. \end{cases}$
  od
```

```
    $\hat\sigma := \operatorname*{argmin}_{\sigma\in\mathcal{S}_{47}} \sum_{i=1}^{47} c_{i,\sigma(i)}$;     (*A linear assignment problem *)
    foreach $i \in 1..46$ do
        $h_i := \operatorname*{argmin}_{h\in 1..2} -\ln\left(f_{\hat\sigma(i)}(x_{i,h})q_{i,h}\right)$;
    od
    return $\hat\sigma_{|1..46}$, $h_1, \dots, h_{46}$;
end
```

The algorithmic difference between polarity-free and polarity-dependent classification is best seen in Tables 1 and 2: instead of likelihoods as in [16] the tables contain the weighted sums and weighted maxima over the likelihoods of the two polarities, respectively. The polarity-dependent algorithms designed in [17] and [16] for handling cells with missing or extra chromosomes can be modified in the same way for polarity-free classification. They need in addition prior probabilities of missing chromosomes and of various trisomies.

3.2 Distributional models

The algorithms above may be combined with the statistical models f_j mentioned in the introduction:

(i) *Elliptical symmetry.* Here, f_j is the density of the elliptically symmetric distribution with expectation e_j, variance V_j, and radial function $\varphi\colon [0,\infty[\to [0,\infty[$. We will use the radial functions $\varphi_{\text{normal}}(r) = \beta_{\text{nor}}e^{-r^2/2}$ and $\varphi_{\text{Pareto}}(r) \sim \beta_\lambda r^{-\lambda}$ for some λ exceeding the number of features + 2; for more details cf. [19].

(ii) *Quadratic asymmetry.* In this model, f_j is the density of the quadratically asymmetric distribution with expectation e_j, variance V_j, quadratic asymmetry Q_j, and a radial function as in (i); for more details cf. [15], and [16].

(iii) *Mixture models* with elliptically symmetric or quadratically asymmetric outliers. The density f_j of class j is a mixture of two densities, the density of the regular observations of class j, $f_{j,\text{REG}}$, and the density of the outliers of class j, $f_{j,\text{OUT}}$; i.e., $f_j = (1-\gamma)f_{j,\text{REG}} + \gamma f_{j,\text{OUT}}$ for some (usually small) number γ, $0 < \gamma < 1$. As $f_{j,\text{REG}}$ we choose the normal density with expectation $e_{j,\text{REG}}$ and covariance matrix $V_{j,\text{REG}}$ and $f_{j,\text{OUT}}$ is either an elliptically symmetric density with expectation $e_{j,\text{OUT}}$, covariance matrix $V_{j,\text{OUT}}$, and a radial function $\varphi\colon [0,\infty[\to [0,\infty[$ or a quadratically asymmetric density with the quadratic asymmetry $Q_{j,\text{OUT}}$ as an additional parameter. More refined models could be used for the regular distribution but, in both cases, best results are achieved if $f_{j,\text{REG}}$ is normal and φ is of Pareto's type; for more details cf. [17] and [16].

We name the new polarity-free estimators derived from Algorithm SCC VIEC_φ, VIQA_φ, VIECO_φ, and VIQAO_φ, $\varphi \in \{\text{normal}, \text{Pareto}, \dots\}$, according to the statistical model of feature sets used.

4 Experimental Results

In this section, we offer test results for the polarity-free classifiers $\text{VIEC}_{\text{normal}}$, $\text{VIEC}_{\text{Pareto}}$, $\text{VIECO}_{\text{Pareto}}$, and $\text{VIQAO}_{\text{Pareto}}$ and compare them with the previous polarity-dependent estimators $\text{IEC}_{\text{normal}}$, $\text{IEC}_{\text{Pareto}}$, $\text{IECO}_{\text{Pareto}}$, and $\text{IQAO}_{\text{Pareto}}$. We use again the large Copenhagen image data set Cpr [12], [17] and the profiles computed by the Edinburgh MRC chromosome analysis system [13] as a benchmark. The results pertain to all cells in the data set, also those with a missing chromosome or containing a trisomy.

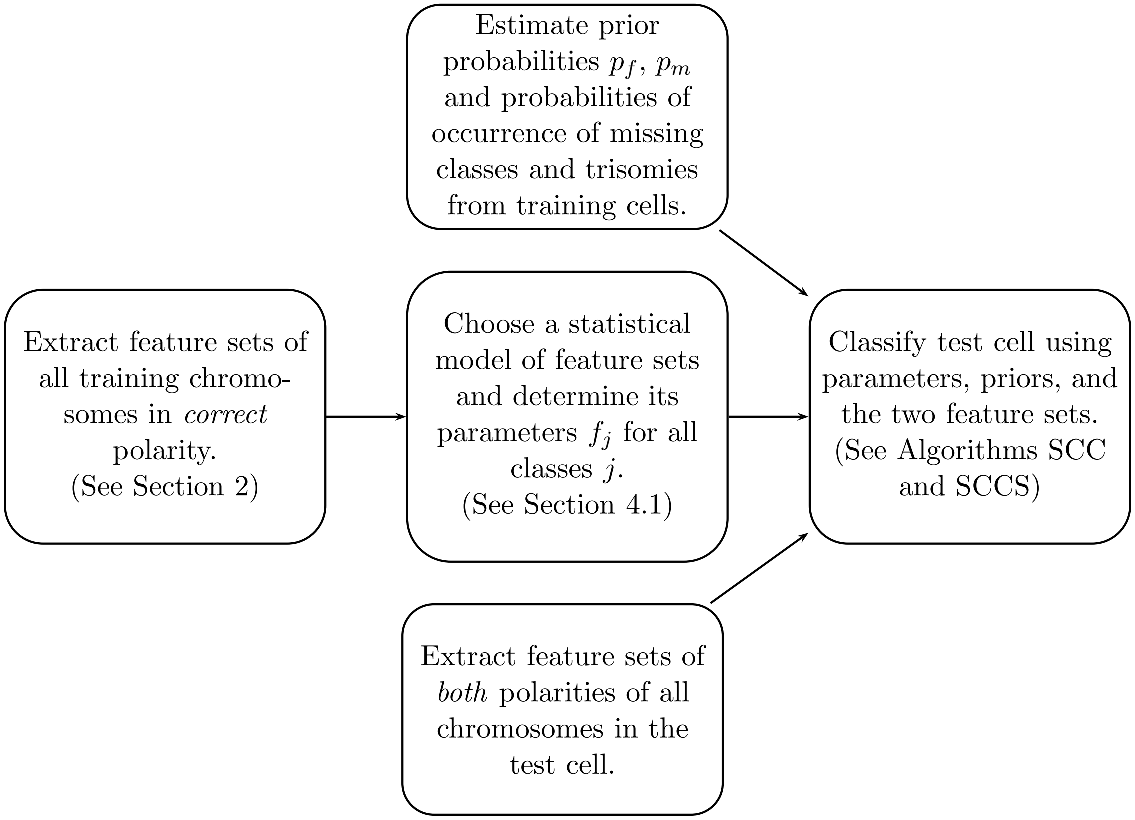
Estimate prior probabilities p_f, p_m and probabilities of occurrence of missing classes and trisomies from training cells.
Extract feature sets of all training chromosomes in *correct* polarity. (See Section 2)
Choose a statistical model of feature sets and determine its parameters f_j for all classes j. (See Section 4.1)
Classify test cell using parameters, priors, and the two feature sets. (See Algorithms SCC and SCCS)
Extract feature sets of *both* polarities of all chromosomes in the test cell.

Figure 6: Overview of polarity-free classification

4.1 Parameter estimation

The model parameters (expectation, covariance matrices, quadratic asymmetries, exponents of radial functions) of polarity-free classifiers are estimated as described in [17] and [16] except that features are extracted from profiles in *correct* (i. e., manually determined) polarity. Parameters γ of mixture models with outliers were determined by the trimming method described in [17]. This method requires two cutoffs, $\mathrm{cut}_{\mathrm{BAS}}$ and $\mathrm{cut}_{\mathrm{OUT}}$, for the generation of the basic and outlier populations. The basic population BAS of each class consists of those observations whose Mahalanobis distance, with respect to the entire population POP, does not exceed $\mathrm{cut}_{\mathrm{BAS}}$. The outlier population OUT consists of those observations whose Mahalanobis distance, with respect to the basic population, exceeds $\mathrm{cut}_{\mathrm{OUT}}$. The connections are illustrated by the following diagram.

$$\begin{tikzcd} \mathrm{POP} \arrow[r] & \begin{array}{c} e_{\mathrm{POP}} \\ V_{\mathrm{POP}} \end{array} \arrow[r, "\mathrm{cut}_{\mathrm{BAS}}"] & \mathrm{BAS} \arrow[r] & \begin{array}{c} e_{\mathrm{BAS}} \\ V_{\mathrm{BAS}} \end{array} \arrow[r, "\mathrm{cut}_{\mathrm{OUT}}"] & \mathrm{OUT} \arrow[r] & \begin{array}{l} e_{\mathrm{OUT}} \\ V_{\mathrm{OUT}} \\ (Q_{\mathrm{OUT}}) \end{array} \end{tikzcd}$$

For the estimation of parameters, covariance matrices of populations containing outliers were robustly estimated, i.e., a shrinking factor of 0.9 was applied to their off-diagonal elements, cf. [14], [17]. These populations are the entire population in all cases and the outlier populations needed for estimating the quadratic asymmetries in the cases IQAO and VIQAO but not the basic populations.

The prior probabilities $q_{i,h}$ are chosen as 0.5. The prior probabilities of a cell to be female or male, of missing chromosomes, and of various trisomies are those listed in [17], Section 5, and [16], Section 4.3.

4.2 Classification and test results

In the classification process, the shrinking factor above was applied to *all* covariance matrices. Moreover, all quadratic asymmetries were multiplied by another shrinking factor, the Q-factor ≤ 1, cf. [16]. Note that, for classifying a cell, two feature sets are extracted from each of its chromosomes, one for each polarity.

Let us finally compare the estimated probabilities of misclassification of the new polarity-free classifiers and new features extracted from the profiles mentioned at the beginning of this section (Table 3(a)) with those obtained from previous polarity-dependent classifiers applied to the same feature set (Table 3(b)) and to the feature set of the Edinburgh MRC chromosome analysis system (Table 3(c)).

In all tables, test-set results refer to the cross-validation (holdout, jackknifing) method described in [17] whereas training-set results use all cells for training and classification. Roughly speaking, in Table 3, S.D. is 1 with respect to cells and 0.1 with respect to chromosomes. This comparison shows that the new feature set has an advantage of 10 to 15% over the previous one; a similar improvement is due to the polarity-free classifiers.

Our implementation in the Programming Language C on a workstation SUN Ultra 1, Model 140, takes for classification of one cell about 0.08 s in the case IEC, 0.14 s in the case IECO, and 1.1 s in the case IQAO, respectively. Execution times of the polarity-free V-classifiers, needing two feature sets, are about twice as large as those of the polarity-dependent ones.

5 Discussion

Variants possess a large number of applications. We use them in this paper in order to handle the ambiguity stemming from the two polarities of chromosomes. Other applications to automatic chromosome analysis are possible, for instance to the problems of shape recognition, of segmentation, and of centromere handling. The polarity-free classifiers remove errors due to wrong polarities caused by erroneously estimated centromeres. Compared with the previous features and polarity-dependent classifiers the new polarity-free V-classifiers together with the modified features remove between 20% and 30% of classification errors depending on their statistical types. The best error rates attained in this way for the data set Cpr are just below 1%.

The manual error rate contained in this data set was estimated as 0.3% with respect to chromosomes [8], p. 314. There are mainly two causes for the difference between the error rate of 0.3% of the expert and that of our currently best classifiers. The expert uses probably additional features and, more importantly, handles outliers in a more flexible way than an automatic system can do. We believe that most of the remaining errors of our best classifiers are due to remaining outliers in the data of which there are several causes. These are

(i) wrong medial axes mainly in bent and badly-shaped chromosomes,

Table 3: Overall test- and training-set error rates for various classifiers and featuers. The notation p/q means that p% of chromosomes were misclassified in q% of cells. The exponents λ and λ_{OUT} of the Pareto-type radial functions (see 3.2), the parameters $\mathrm{cut}_{\mathrm{BAS}}$ and $\mathrm{cut}_{\mathrm{OUT}}$ (see 4.1) controlling the trimming method, and the Q-factors (see 4.2) for robust estimation of quadratic asymmetries are indicated where they apply.

(a) Polarity-free classifiers, applied to the new 29 features. The best classifier reported in this communication is $\mathrm{VIQAO}_{\mathrm{Pareto}}$ applied with robust estimation of variances; cf. the rightmost test-set result.

(b) Classifiers with polarities determined by estimated centromeres, applied to the new 29 features.

(c) Classifiers with polarities determined by estimated centromeres, applied to 24 features from the Edinburgh MRC chromosome analysis system. Results taken from [17] and [16].

(a)

	$\mathrm{VIEC}_{\mathrm{normal}}$	$\mathrm{VIEC}_{\mathrm{Pareto}}$	$\mathrm{VIECO}_{\mathrm{Pareto}}$	$\mathrm{VIQAO}_{\mathrm{Pareto}}$
		$\lambda = 33.0$	$\lambda_{\mathrm{OUT}} = 33.0$	$\lambda_{\mathrm{OUT}} = 34.0$
test set (%)	1.94/27.2	1.27/18.3	0.99/14.9 $\mathrm{cut}_{\mathrm{BAS}} = 10.0$ $\mathrm{cut}_{\mathrm{OUT}} = 7.3$	0.92/14.2 $\mathrm{cut}_{\mathrm{BAS}} = 6.5$ $\mathrm{cut}_{\mathrm{OUT}} = 7.1$ Q-factor = 0.4
training set (%)	1.82/26.0	1.20/17.6	0.76/12.2 $\mathrm{cut}_{\mathrm{BAS}} = 13.5$ $\mathrm{cut}_{\mathrm{OUT}} = 8.5$	0.58/10.1 $\mathrm{cut}_{\mathrm{BAS}} = 7.9$ $\mathrm{cut}_{\mathrm{OUT}} = 7.5$ Q-factor = 1.0

(b)

	$\mathrm{IEC}_{\mathrm{normal}}$	$\mathrm{IEC}_{\mathrm{Pareto}}$	$\mathrm{IECO}_{\mathrm{Pareto}}$	$\mathrm{IQAO}_{\mathrm{Pareto}}$
		$\lambda = 33.0$	$\lambda_{\mathrm{OUT}} = 33.0$	$\lambda_{\mathrm{OUT}} = 33.0$
test set (%)	2.16/28.9	1.49/20.8	1.15/16.5 $\mathrm{cut}_{\mathrm{BAS}} = 8.5$ $\mathrm{cut}_{\mathrm{OUT}} = 7.0$	1.10/15.9 $\mathrm{cut}_{\mathrm{BAS}} = 6.4$ $\mathrm{cut}_{\mathrm{OUT}} = 7.5$ Q-factor = 0.5
training set (%)	2.02/27.6	1.41/19.9	0.85/13.6 $\mathrm{cut}_{\mathrm{BAS}} = 9.5$ $\mathrm{cut}_{\mathrm{OUT}} = 8.0$	0.64/11.1 $\mathrm{cut}_{\mathrm{BAS}} = 8.7$ $\mathrm{cut}_{\mathrm{OUT}} = 7.7$ Q-factor = 1.0

(c)

	$\mathrm{IEC}_{\mathrm{normal}}$	$\mathrm{IEC}_{\mathrm{Pareto}}$	$\mathrm{IECO}_{\mathrm{Pareto}}$	$\mathrm{IQAO}_{\mathrm{Pareto}}$
		$\lambda = 28.0$	$\lambda_{\mathrm{OUT}} = 30.5$	$\lambda_{\mathrm{OUT}} = 29.0$
test set (%)	2.68/34.5	1.84/25.2	1.32/19.3 $\mathrm{cut}_{\mathrm{BAS}} = 8.5$ $\mathrm{cut}_{\mathrm{OUT}} = 7.0$	1.22/17.5 $\mathrm{cut}_{\mathrm{BAS}} = 6.1$ $\mathrm{cut}_{\mathrm{OUT}} = 6.7$ Q-factor = 0.6
training set (%)	2.55/33.3	1.72/24.0	1.09/16.5 $\mathrm{cut}_{\mathrm{BAS}} = 8.5$ $\mathrm{cut}_{\mathrm{OUT}} = 7.0$	0.78/13.3 $\mathrm{cut}_{\mathrm{BAS}} = 7.5$ $\mathrm{cut}_{\mathrm{OUT}} = 7.2$ Q-factor = 1.0

(ii) Giemsa stains on or at the chromosome,

(iii) X- or Y-shapedness because of late metaphases,

(iv) overlappings of other chromosomes, and

(v) large structural anomalies such as translocations, deletions, duplications, inversions, or huge satellites.

A comparison of the classifier VIEC$_{\text{normal}}$ (which has no outlier handling potential) with the classifiers VIEC$_{\text{Pareto}}$, VIECO$_{\text{Pareto}}$, and VIQAO$_{\text{Pareto}}$ (which have high outlier handling potentials) shows that they can to a certain degree be made up for by methods of statistical data analysis.

One cause of asymmetries in the feature data previously used were wrong polarities. Table 3(a) shows that the polarity-free *asymmetric* classifier VIQAO is still noticeably superior to the *symmetric* classifiers VIEC and VIECO. This means that, despite the removal of outliers caused by wrong polarities, there remains asymmetry in the data.

All three Tables 3(a) – (c) show a large gap between the test-set and the training-set error rates of the quadratically asymmetric classifiers IQAO and VIQAO. The reason is that the expectations e_{OUT} and the matrices V_{OUT} and Q_{OUT} together make up 1276 (900) real parameters for a feature set of dimension 29 (24). They have to be estimated on the basis of the outliers which, in Cpr, amount to about $\gamma \approx 12\%$ of the whole data set. These fewer than 1000 chromosomes in each class are not sufficient for reliably estimating such a large number of real parameters for test-set results. This is also plain from the small Q-factors allowed in the test-set classifiers. Moreover, the large number of parameters leads to substantial overfitting in the training-set classifiers. The genuine Bayesian error rate of VIQAO$_{\text{Pareto}}$, e.g., will lie somewhere between the test-set error rate of 0.92% and the training-set error rate of 0.58%. A data set for parameter estimation even larger than Cpr would diminish this gap.

Approaching the error rate of automatic chromosome classification to that of the expert is not an easy task and it is not just *one* idea that will do it. It seems that, with constrained classification, elliptical symmetry, quadratic asymmetry, mixture models, and variants, the potentiality of statistical methods is to a high degree exhausted. On the other hand, there is still the possibility to improve the image processing required for profile extraction. This is the subject of the communication [20].

Acknowledgements. We thank Dr. Jim Piper for providing us access to the Edinburgh MRC chromosome analysis system. We also thank the Associate Editor and the referee for their comments which helped to make this a more complete paper.

References

[1] Balinski, M.L., Signature methods for the assignment problem. Operations Res. 33 (1985) 527–536.

[2] Denver Conference, A proposed standard system of nomenclature of human mitotic chromosomes. Lancet 1 (1960) 1063–1065.

[3] Fang, K.-T., S. Kotz, and K.W. Ng, Symmetric Multivariate and Related Distributions. Chapman and Hall, London, 1990.

[4] Gallegos, M. T. and G. Ritter, Bayesian classification and selection. In preparation

[5] Granum, E., Application of statistical and syntactical methods of analysis and classification to chromosome data. In: J. Kittler, K.S. Fu and L.F. Pau (eds.), Pattern Recognition Theory and Applications, Reidel Publishing Company 1982, 373–398.

[6] Griffiths, A.J.F., J.H. Miller, D.T. Suzuki, R.C. Lewontin, and W.M. Gelbart, An Introduction to Genetic Analysis. Freeman and Company, New York, 1993.

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

[8] Kleinschmidt, P., I. Mitterreiter and J. Piper, Improved chromosome classification using monotonic functions of mahalanobis distance and the transportation method. Mathematical Methods of Operations Research 40 (1994) 305–323.

[9] Ledley, R.S. and F.H. Ruddle, Chromosome analysis by computer. Scientific American 214 (1966), 40–46.

[10] Lundsteen, C., A.-M. Lind and E. Granum, Visual classification of banded human chromosomes. Ann. Hum. Genet., Lond. 40 (1976) 87–97.

[11] Papadimitriou, C.H. and K. Steiglitz, Combinatorial Optimization. Prentice–Hall, Englewood Cliffs, New Jersey, 1982.

[12] Piper, J., Variability and bias in experimentally measured classifier error rates. Pattern Recognition Letters 13 (1992) 685–692.

[13] Piper, J. and E. Granum, On fully automatic feature measurement for banded chromosome classification, Cytometry 10 (1989) 242–255.

[14] Piper, J., I. Poole, and A. Carothers, Stein's paradox and improved quadratic discrimination of real and simulated data by covariance weighting. Proceedings of the 12 th IAPR International Conference on Pattern Recognition, Jerusalem, Isreal, 1994 (Los Alamitos, CA, IEEE Comput. Soc. Press, 1994) 529–532.

[15] Ritter, G., Quadratically asymmetric distributions and their application to chromosome classification. In: A. Prat and E. Ripoll (eds.), Proceedings in Computational Statistics, Short Communications, COMPSTAT 1996 (1996) 99–100.

[16] Ritter, G. and K. Gaggermeier, Automatic classification of chromosomes by means of quadratically asymmetric statistical distributions. Pattern Recognition 32 (1999) 997–1008.

[17] Ritter, G. and M.T. Gallegos, Outliers in statistical pattern recognition and an application to automatic chromosome classification. Pattern Recognition Letters 18 (1997) 525–539.

[18] Ritter, G. and M.T. Gallegos, Bayesian object identification: variants. Submitted.

[19] Ritter, G., M.T. Gallegos, and K. Gaggermeier, Automatic context-sensitive karyotyping of human chromosomes based on elliptically symmetric statistical distributions. Pattern Recognition, 28 (1995) 823–831.

[20] Ritter, G. and G. Schreib, Using dominant points and variants for profile extraction from chromosomes. To appear in Pattern Recognition.

[21] Slot, R.E., On the profit of taken into account the known number of objects per class in classification methods, IEEE Trans. Inform. Theor. 25 (1979) 484–488.

[22] Tso, M.K.S. and J. Graham, The transportation algorithm is an aid to chromosome classification. Pattern Recognition Letters 1 (1983) 489–496.

[23] Tso, M., P. Kleinschmidt, I. Mitterreiter, and J. Graham, An efficient transportation algorithm for automatic chromosome karyotyping. Pattern Recognition Letters 12 (1991) 117–126.

[24] Turpin, R. and J. Lejeune, Les Chromosomes Humains. Gauthier-Villars, Paris, 1965.