Appeared in Pattern Recognition Letters **18** (1997) 525–539

Outliers in Statistical Pattern Recognition and an Application to Automatic Chromosome Classification

Gunter Ritter and María Teresa Gallegos
Universität Passau,
Fakultät für Mathematik und Informatik,
D-94 030 Passau

Abstract: We propose a heuristic method of parameter estimation in mixture models for data with outliers and design a Bayesian classifier for assignment of m objects to $n \geq m$ classes under constraints.

This method of outlier handling combined with the classifier is applied to the well–known problem of automatic, constrained classification of chromosomes into their biological classes. We show that it decreases the error rate relative to the classical, normal, model by more than 50%. When applied to the Edinburgh feature data of the large Copenhagen image data set Cpr our best classifier yields an error rate close to 1.3% relative to chromosomes; 4 out of 5 cells are correctly classified.

Key words: Outlier estimation, mixture distributions, trimming method, Bayesian classification, statistical pattern recognition, automatic chromosome classification, karyotyping, diagnostic classification, biomedical data model.

1 Introduction

The contribution of this paper is application of an idea from modern statistics to classification problems: *outliers*. The method is applicable to general classification problems with numerical features; it is mainly motivated by our goal to reduce the error rate of automatic chromosome classification.

Experience shows that almost all real data contain outliers in the sense that (at least one of) their features differ dramatically from what may be expected. There does not seem to be a rigorous and unified concept of outlier to date; there are two vague notions instead. The first one is (a) that of a 'spurious' observation obeying *no statistical law*. There might even exist data consisting solely of such outliers which would render these data intractible for statistics. If an experiment containing such outliers is repeated then anything can happen: fewer or more outliers and they may look similar or very different.

The other (b) refers to a distributional model of the data, in most cases a member of some parametric family like the normal, the elliptically-contoured (elliptically-symmetric), or the exponential family, and is the subject of our interest here. This notion means observations that obey *some statistical law* but appear more often than the assumed distribution allows. If there are (b_1) only one or very few outliers in the data ('few' relative to the total number of observations) then these outliers might be rare events with respect to the distribution and do not necessarily indicate an inappropriate distributional assumption on the data. If there are (b_2) many outliers, 10% say, then they must be attributed to an inaccurate choice of the distribution. This notion

is often associated with 'heavy tails'. When an experiment in a situation involving such outliers is repeated then the outliers will look similar in number and characteristic.

In cases (a) and (b_1) the best thing a statistician can do is

- identifying the outliers and rejecting them in order to restore the characteristics of the data, or
- applying robust methods of estimation.

These methods try to eliminate or at least reduce the influence of outliers and to model the regular observations, only. Therefore, if one applied these methods in case (b_2) one would loose genuine and useful statistical information on the population.

The usual parametric families are not flexible enough to serve as proper statistical models of populations with outliers of type (b_2) in the sense that *one* of their members fits the 'regular' objects and the outliers at the same time. Besides the regular objects one has to model the outliers as well. This can be done by

- modelling the population by means of a mixture of two distributions, one fitting the regular observations and the other the outliers.

The mixture distribution should be chosen in such a way as to minimize the number of observations among the given data that appear as outliers. In this sense, appropriate model selection is synonymous with absence of outliers in the data. It seems that the existing literature has concentrated mainly on estimating parameters of the *regular* populations, giving up the outliers for the purpose of inference. In this paper we try to accommodate the outliers by carefully adapting the distribution. In this way we consider outliers to be, albeit inconvenient, members of their populations. This method is subsequently applied to *discriminant analysis*.

Bayesian discriminant analysis compares certain linear combinations of posterior densities of a feature vector with respect to the classes considered. Statistical decision theory shows that this classification method has minimal loss. Since posterior densities are products of likelihood functions, i.e., class-conditional densities, and prior probabilities this optimality can only be exploited if accurate models of likelihood functions are used. Deviation from the class-conditional densities means transition to a suboptimal classifier. The more accurate the model of the likelihood functions the better the classification. In the context of classification one must classify *all* members of the population, including outliers. If these are not adequately modelled then there is a high risk of their misclassification. This is the reason why we advocate modelling the populations to precision in discriminant analysis, adapting class–conditional distributions also to outliers of type (b_2). Stated in other terms, we must find class–conditional distributions that represent *all* observations including the potential outliers. The low error rate that we achieve in our application to chromosome classification supports this rationale.

The outlines of the paper are as follows. In Section 2, we propose a trimming method for estimating, besides the parameters of the regular population, the parameters of the outliers, too, thus identifying the mixture model of data with outliers. This, however, presupposes a large data set so that sufficiently many outliers are present to allow identification of their model.

In Section 3 we discuss a Bayesian method for assigning m objects to n $(\geq m)$ classes where classes may consist of several *categories*, e.g., regular and outlier objects. In Sections 4 and 5, this method is applied to the well-known problem of automatic classification of human chromosomes into their 24 biological classes.

Classification of the suitably-stained chromosomes of a eukariotic cell under a light microscope in their classes is a well-defined problem which is carried out by a human expert in about 10 minutes and at an error rate of 0.3% with respect to chromosomes (clinical applications). Attempts to automate this task go back to the late fifties, cf. Ledley and Ruddle [1966]. For a survey we refer the reader to Granum [1982] and Piper et al. [1980]. The task consists of

(i) selecting a metaphase cell under the microscope,

(ii) segmenting the cell in its chromosomes,

(iii) finding the medial axes of chromosomes,

(iv) computing shape and density profiles along the axes,

(v) extracting a number of numerical features, such as *area*, *length*, *density*, *centromeric index*, *number of profile maxima*, *shape*, and *band pattern* from each chromosome, and

(vi) classifying the feature sets obtained by applying methods of discriminant analysis.

Often the phases (i) and (ii) are carried out interactively. We deal here with the last one of these six successive phases, more specifically with statistical classification methods.

In our earlier paper Ritter et al. [1995] we designed a series of Bayesian classifiers, called IEC_{normal}, $IEC_{exponential}$, IEC_{Pareto}, and $IEC_{empirical}$. They are all based on the statistical assumptions of *independence* of chromosomes in a cell and of *elliptical symmetry* (cf. Fang et al. [1990]) of the feature set of each chromosome. The first of these assumptions is classical while the second one is an extension of the classical assumption of normality and is new in the context of chromosome classification. The subscript stands for the type of the radial function. In the present paper we proceed a step further applying the outlier method of Section 2 to chromosome analysis.

We use the Edinburgh feature data of the large Copenhagen image data set consisting of 2 804 human cells. These data contain outliers. There are chromosomes which are severely bent, cf. Fig. 1(b), or overlapped by other chromosomes, cf. Fig. 1(c). In the first case, the image processing involved in phase (iii) is likely to miss the correct medial axis

Fig. 1: (a) Regular, straight chromosome; (b) U-shaped chromosome; (c) overlapped chromosomes

Fig. 2: (a) Histogram of a feature exposing outliers. (b) Histogram of feature 'area' with only few outliers.

of the chromosome so that profiles cannot be properly extracted, a confusion resulting in erroneous mesurements of most features of this chromosome. Nevertheless, the features 'area' and 'density', not depending on the medial axis, are still reliable. In the second case at least the

band pattern, providing usually a set of very characteristic and discriminating features, is contaminated; yet, the features 'area' and 'length' are still reliable. Such chromosomes, while being perfectly normal from a biological and medical point of view, will appear as outliers from their classes with respect to any member of the known parametric models although they contain some information on their classes. Since there are a large number of them they can be considered as outliers of type (b_2). We, therefore, use the trimming method described in Section 2, decomposing each class into two categories, a *regular* and an *outlier* category. Assuming again that the feature sets of both the regular and the outlier populations are elliptically contoured, we obtain mixture models of the classes which we can identify. Feeding the mixture densities into the constrained classifiers of Section 3 we obtain a series of new classifiers $IECO_{\varphi\psi}$, the letter I standing for *independent* chromosomes, EC for *elliptically contoured*, and O standing for *outlier*; the subscripts φ and ψ indicate types of radial functions of the regular and outlier densities, respectively. In order to take account of numerical abberations we design a classifier for cells containing 46 chromosomes and two classifiers for cells with one missing and one extra chromosome (trisomy), respectively.

The *constrained* classifiers take into account the correct number of chromosomes in each biological class. Thus, the classes estimated are not the 24 biological classes but essentially the permutations of the chromosomes in the cell which means a classification problem with more than 10^{51} classes. Tso and Graham [1983] showed how it can be efficiently solved.

We show experimentally that the new classifiers substantially improve classification results. The best previous classifier known to us is due to Kleinschmidt et al. [1994]; it has been identified in Ritter et al. [1995] as the MAP classifier for the Pareto–tailed elliptically–contoured statistical model of chromosomes and we called it IEC_{Pareto} in our taxonomy. Since such a distribution has a heavy tail (relative to the normal distribution) it tolerates outliers with respect to normal data and can be considered as a robust method of classification. Yet, our currently best classifier $IECO_{norPar}$, applied to the same feature data, reduces its cross–validation error rate by more than 30%; cf. the comparison of MAP classifiers under various distributional assumptions shown in Table 5.1. A reduction by 26% is due to the method of outlier handling which is the main subject of this paper and another 6% to more admissible and robust estimation of covariance matrices, cf. Subsection 2.3. The new classifier attains an error rate of 1.32%, cf. Table 5.1(b). This error rate is relative to *all* cells in the data set.

Our algorithms were implemented in the Programming Language C. Classification of a cell on a SUN workstation SPARC 10 takes about 0.4 seconds.

An abstract of this paper was published as a conference report Gallegos and Ritter [1996].

2 Populations with outliers

2.1 A statistical outlier model

Not much is known about outlier treatment in the *multidimensional* case. As explained in the introduction, we assume that the whole, contaminated, population consists of a mixture of two populations, a *regular* population with distribution μ_{REG} and an *outlier* population with distribution μ_{OUT}:

$$\mu_{POP} = (1-\alpha)\mu_{REG} + \alpha\,\mu_{OUT}.$$

Here, α is the prior probability of appearance of an outlier. Such mixture models of data with outliers go back to Dixon [1953] and Tukey [1960]; cf. also Barnett and Lewis [1994], p. 46

(iii). The point which seems to be new in our approach is that we identify not only μ_{REG}, but also μ_{OUT} for establishing a closer statistical model of μ_{POP} for the purpose of *classification*. Our experiments in Section 5 show that this procedure is superior to applying merely robust methods.

If the distributions μ_{REG} and μ_{OUT} are assumed to be members of given parametric families then there exist Bayesian algorithms for estimating their parameters, cf. Mardia et al. [1979], Ch. 13.2. A popular method is the EM algorithm studied in Dempster et al. [1977]. It is also possible to apply clustering algorithms; for an overview see Bock [1996]. Another method for detecting the regular observations is the minimum–volume– ellipsoid estimator, cf. Rousseeuw [1987], p. 258. One might also assume that there are more outlier classes than just one. These may, however, sometimes become too small for reliable estimation of their parameters. A related algorithm, derived from the EM algorithm, which has the advantage of estimating also the *number* of classes is designed in Cheeseman et. al. [1988]. Some of these algorithms are not efficient enough to allow application to very large data sets and/or high–dimensional feature spaces.

We found that the simple, heuristic trimming method described in the following subsection is sufficient for a substantial reduction of the error rate. We do not claim that this is the best method to search for an appropriate mixture–model structure of class–conditional distributions and comparison of different methods for identifying mixture models is not the purpose of this paper. We rather compare various statistical models of class–conditional distributions like the normal, the elliptically–symmetric, and mixture models for classification and give evidence that application of mixture models can improve classification results to a considerable extent.

2.2 Parameter estimation in mixture models with outliers

Let the feature space be denoted by S and let $\mathcal{M}$ be a family of distributions on S. Let there be given a scalar indicator $\chi : \mathcal{M} \times \mathcal{S} \to \mathbf{R}$ which discriminates sufficiently well between regular objects and outliers with respect to $\mathcal{M}$: If $\chi(\mu, \mathbf{x})$ exceeds some specific cutoff then an object with features $\mathbf{x} \in \mathbf{S}$ is classified as an outlier with respect to the distribution $\mu \in \mathcal{M}$ and as regular, otherwise. In the case of an elliptically-contoured family $\mathcal{M}$ a natural indicator is the Mahalanobis distance. More generally, moments could be used. Our method of parameter estimation then proceeds in two steps.

For estimating μ_{OUT} we need a method for identifying outliers, cf. Davies and Gather [1993] or Hawkins [1980]. This, in turn, needs at least an approximation of μ_{REG}. In a first step we provide this by using a modified, multivariate version of Anscombe's [1960] *premium protection rule*; it is related to the *ellipsoidal trimming method*, cf. Barnett and Lewis [1994], p. 277. Let us denote the whole, contaminated, population by POP and its estimated distribution by $\hat{\mu}_{POP} \in \mathcal{M}$. An observation $\mathbf{x} \in POP$ is classified as *basic* if $\chi(\hat{\mu}_{POP}, \mathbf{x})$ does not exceed a cutoff cut_{BAS} to be specified in advance. This value should be chosen in such a way as to qualify the largest part of POP, say 90%, as basic. Denote the resulting basic population by BAS and its estimated distribution by $\hat{\mu}_{BAS} \in \mathcal{M}$. This is the required approximation of μ_{REG}. The distributions $\hat{\mu}_{POP}$ and $\hat{\mu}_{BAS}$ are assumed to belong to the same parametric family $\mathcal{M}$. Of course, its parameter set must be small enough to make reliable parameter estimation possible on the basis of the observations.

The second step needs outlier identification. An observation $\mathbf{x} \in POP$ is classified as an *outlier* with respect to $\hat{\mu}_{BAS}$ (or μ_{REG}) if $\chi(\hat{\mu}_{BAS}, \mathbf{x})$ exceeds another cutoff cut_{OUT} to be specified in advance. If μ_{REG} is assumed to be normal and $\chi(\hat{\mu}_{BAS}, \cdot)$ is the Mahalanobis distance with respect to expectation and variance of $\hat{\mu}_{BAS}$ then a reasonable choice of cut_{OUT} is about two

standard deviations. In practice, cut_{OUT} (and cut_{REG}) will have to be determined by calibration. Let the resulting outlier population be denoted by the symbol OUT. The populations BAS and OUT will usually overlap to a certain degree since BAS is determined by $\hat{\mu}_{POP}$ and OUT by $\hat{\mu}_{BAS}$. Let the estimated distribution of the population OUT be denoted by $\hat{\mu}_{OUT}$. This is our estimation of μ_{OUT}. As mentioned above, we use the distribution $\hat{\mu}_{BAS}$ as an approximation of the distribution μ_{REG}. This is justified by definition of the outlier population OUT. In short, we have the following diagram of dependencies:

$$POP \longrightarrow \hat{\mu}_{POP} \stackrel{cut_{BAS}}{\longrightarrow} BAS \longrightarrow \hat{\mu}_{BAS} \stackrel{cut_{OUT}}{\longrightarrow} OUT \longrightarrow \hat{\mu}_{OUT}.$$

The process could be iterated. This would not substantially improve the quality of the mixture model and we prefer to keep the algorithm simple.

2.3 Estimation of covariances: covariance weighting

The quality of the model identified in 2.2 depends on the quality of the estimates of covariances (and expectations) which are needed for identifying $\hat{\mu}_{BAS}$ and $\hat{\mu}_{OUT}$. The obvious estimators enjoy only asymptotic optimality properties and it is beneficial to apply more admissible and robust methods also in the present situation. Several heuristic methods exist for improving estimates of covariances; cf. Friedman [1989]. These mostly come down to either squashing the eigenvalues, or enlarging the diagonal, or diminishing the off–diagonal entries. Theoretically well–founded methods can be found in the literature cited in Friedman [1989], beginning of Section 3. Some of them are derived from the *James–Stein estimator*; cf. Stein [1956], James and Stein [1961],

Consider a random vector $Y : (\Omega, P) \to \mathbf{R}^a$. It has been known for a long time that the obvious estimator $\varphi(Y) = Y$ of EY is (unbiased and) admissible within the classes of all *unbiased* or *translation–invariant* estimators of EY; i.e., this estimator has minimum square risk within these classes uniformly for all possible vectors $EY \in \mathbf{R}^a$. Stein, loc. cit., showed that, if Y is i.i.d. and not deterministic, it is inadmissible within *all* estimators if $b \geq 3$. Moreover, James and Stein [1961], Formula (6), proved that the estimator

$$(1) \quad \varphi_{JS}(Y) = (1 - \frac{a-2}{Y^T\, V(Y)^{-1}\, Y})\, Y$$

has uniformly smaller risk with respect to the quadratic form associated with $V(Y)^{-1}$ and is optimal among all estimators of the form $(1 - \frac{b}{Y^T\, V(Y)^{-1}\, Y})\, Y$, $b \in \mathbf{R}$. It is a multiple of Y by a factor $c < 1$ depending on Y.

Piper et al. [1994] apply the James–Stein estimator to the off-diagonal entries of covariance matrices deriving from it a method of improved covariance estimation which they call *covariance weighting*. Covariance weighting means multiplication of the off-diagonal elements of a covariance matrix by a factor $c < 1$. This factor will be called *Stein factor*. In the context of chromosome classification, Piper [1987] noticed that, if the data set is small, even the factor $c = 0$ can sometimes be optimal. In our case, the error rate decreases, even with the large data set used, by about 6% if the factor c = 0.9 is applied; cf. Table 5.1.

It is interesting to discuss the effect of covariance weighting on the eigenvalues. First, since the diagonal is not affected, the sum of the eigenvalues remains invariant. For the next argument we restrict matters to the case of a positive-definite matrix V with a constant diagonal with entries v since a more general investigation would lead us too far afield; the eigenvalues of V are scattered around v. In this case, $V^c = (1 - c)vI + cV$ is the matrix obtained from V by

covariance weighting. From this representation it is clear that the eigenvalues of V^c are the convex combinations $(1-c)v + c\lambda$ of v and the eigenvalues λ of V. Hence, the smaller the factor c the smaller the scatter of the eigenvalues of V^c around v.

3 Bayesian estimation of assignments of objects to mixed classes

3.1 Statistical decision problem and decision set

We now design a Bayesian classifier for the optimal assignment of m objects $i \in 1..m$ to n classes $j \in 1..n$, $m \leq n$, where each class j consists of several *categories* in C_j. (In our application to chromosome classification there are the two categories 'regular' and 'outlier', and the two categories of sex. Eight additional categories are needed for the extra class in a trisomy cell.) Each class may appear only once and categories within classes may be chosen according to given prior probabilities. We also consider the case where one or more classes are not represented by objects, i.e., the object classes make up some unknown subset $M \subset 1..n, \#M = m < n$.

To make things precise assume that the model of an object of class $j \in 1..n$ and of category $k \in C_j$ is a random variable $Z_j^k : (\Omega, P) \to S$ in a state space S. Introducing the random categories $K_j : (\Omega, P) \to C_j$ we observe a random variable $X = (X_1, \ldots, X_m) : (\Omega, P) \to S^m$, a permutation of $(Z_j^{K_j})_{j \in M}$ for the unknown subset M of classes. The task consists in finding the class of the object i with observation X_i. To this end, let $I_m^{(n)}$ be the set of all assignments (injective mappings) $1..m \to 1..n$. We ask for the assignment $\varphi \in I_m^{(n)}$ such that

$$(X_i)_{i=1}^m \sim (Z_{\varphi(i)}^{K_{\varphi(i)}})_{i=1}^m.$$

Thus, a possible decision set of the present statistical decision problem is $I_m^{(n)}$.

3.2 Likelihood function

For determining the likelihood function $L_{\mathbf{x}}(\varphi)$, $\varphi \in I_m^{(n)}$, of the above statistical classification problem posed in Subsection 3.1 we introduce the random injective mapping $\Phi_m : (\Omega, P) \to I_m^{(n)}$. It is natural to assume that the random variables $Z_j^k, j \in 1..n, k \in C_j, \Phi_m$, and $K_l, l \in 1..n$, are all independent. If the density of Z_j^k with respect to some given reference measure ρ on S is $f_{Z_j^k}$ then the likelihood function of the classification problem is given by

$$\begin{aligned}
& L_{\mathbf{x}}(\varphi) \\
= \; & P[X_1 \in dx_1, \ldots, X_m \in dx_m / \Phi = \varphi] / \rho^{\otimes m}(d\mathbf{x}) \\
= \; & P[Z_{\Phi(1)}^{K_{\Phi(1)}} \in dx_1, \ldots, Z_{\Phi(m)}^{K_{\Phi(m)}} \in dx_m / \Phi = \varphi] / \rho^{\otimes m}(d\mathbf{x}) \\
= \; & P[Z_{\varphi(1)}^{K_{\varphi(1)}} \in dx_1, \ldots, Z_{\varphi(m)}^{K_{\varphi(m)}} \in dx_m] / \rho^{\otimes m}(d\mathbf{x}) \\
= \; & \prod_{i=1}^m f_{Z_{\varphi(i)}^{K_{\varphi(i)}}}(x_i), \qquad (2)
\end{aligned}$$

$\mathbf{x} = (x_1, \ldots, x_m) \in S^m, \varphi \in I_m^{(n)}$. The prior probability of membership of an object of class j to category $k \in C_j$ will be denoted by $p_{j,k} = P[K_j = k]$; of course, $\sum_{k \in C_j} p_{j,k} = 1$ for all j. By the

formula of total probabilities the ρ-density of $Z_j^{K_j}$ is the mixture

$$(3) \quad f_{Z_j^{K_j}}(x) = \sum_{k \in C_j} p_{j,k} f_{Z_j^k}(x),$$

$j \in 1..n, x \in S$, of the density functions $f_{Z_j^k}$, $k \in C_j$.

3.3 Prior probabilities of missing objects

We next establish a statistical model of the distribution of Φ_m. As usual, the symmetric group of m elements is denoted by $\mathcal{S}_m$. Note that any injective mapping $\varphi \in I_m^{(n)}$ is specified by its image $Im\,\varphi$ and its permutation $Per\,\varphi := \iota^{-1} \circ \varphi \in \mathcal{S}_m$, where $\iota : 1..m \to Im\,\varphi$ is the order isomorphism. We make the following assumptions on $Im\,\Phi_m$ and $Per\,\Phi_m$:

$$(4) \quad \begin{cases} (i) & \text{there are numbers } \alpha_j \geq 0, 1 \leq j \leq n, \text{ such that, for all } M \subseteq 1..n, \\ & \#M = m, \text{ we have } P[Im\,\Phi_m = M] = c_m \prod_{j \notin M} \alpha_j \text{ and} \\ (ii) & Per\,\Phi_m \text{ is uniformly distributed on } \mathcal{S}_m. \end{cases}$$

Assuming now that $Im\,\Phi_m$ and $Per\,\Phi_m$ are independent we obtain for the prior probabilities

$$(5) \quad \mu_m(\varphi) = P[\Phi_m = \varphi] = c'_m \prod_{j \notin Im\,\varphi} \alpha_j, \quad \varphi \in I_m^{(n)},$$

where $\alpha_j \geq 0$ for $j \in 1..n$. In Subsection 4.3, we will need the case $m = n - 1$. Here, (4) reads

$$(6) \quad \begin{cases} (i) & \text{there are numbers } \alpha_j \geq 0, 1 \leq j \leq n, \text{ such that, for all } j \in 1..n, \text{ we} \\ & \text{have } P[j \notin Im\,\Phi_{n-1}] = \alpha_j \text{ and} \\ (ii) & Per\,\Phi_{n-1} \text{ is uniformly distributed on } \mathcal{S}_{n-1}. \end{cases}$$

The distribution (5) arises in a natural way: First toss a coin for each $j \in 1..n$, coin j with probability of failure $g_j < 1$. Let the random set $N : (\Omega, P) \to 2^{1..n}$ consist of those elements in $1..n$ for which the toss was a success. Then perform a random permutation of N with uniform distribution and independent of the coin tossings in order to derive a random injective mapping $\Phi : (\Omega, P) \to \bigcup_{m \geq 0} I_m^{(n)}$ from N. It is clear that Φ satisfies (4)(ii) conditional on $\#N = m$. Moreover, for $M \subseteq 1..n$, $\#M = m$, we have by construction

$$\begin{aligned} & P[Im\,\Phi = M / \#Im\,\Phi = m] \\ = \; & \frac{P[Im\,\Phi = M]}{P[\#Im\,\Phi = m]} \\ (7) \quad = \; & c''_m \prod_{j \in M} (1 - g_j) \prod_{j \notin M} g_j \\ = \; & c'''_m \prod_{l=1}^{n} (1 - g_l) \prod_{j \notin M} \frac{g_j}{1 - g_j}. \end{aligned}$$

This expression is of the form (4)(i) with $\alpha_j = \frac{g_j}{1-g_j}$. Thus, the distribution of Φ_m is the distribution of Φ conditional on $\#Im\,\Phi = m$.

3.4 MAP estimator

By (2) and (5) the MAP estimator of classes is of the form

$$
\begin{aligned}
& MAP(\mathbf{x}) \\
(8) \quad = \; & argmax_{\varphi} L_{\mathbf{x}}(\varphi)\mu(\varphi) \\
= \; & argmax_{\varphi} \prod_{i=1}^{m} f_{Z^{K_{\varphi(i)}}_{\varphi(i)}}(x_i) \prod_{j \notin Im\,\varphi} \alpha_j .
\end{aligned}
$$

Let us define a permutation σ of $1..n$ by putting $\sigma(i) = \varphi(i)$ for $i \in 1..m$ and by denoting the elements of $1..n \setminus Im\,\varphi$ in their natural order, say, by $\sigma(m+1)..\sigma(n)$. It follows

$$
(9) \quad MAP(\mathbf{x}) = argmax_{\sigma} \prod_{i=1}^{m} f_{Z^{K_{\sigma(i)}}_{\sigma(i)}}(x_i) \prod_{i=m+1}^{n} \alpha_{\sigma(i)} .
$$

Putting

$$
(10) \quad c_{i,j} := \begin{cases} -ln \sum_{k \in C_j} p_{j,k} f_{Z^k_j}(x_i), & i \le m, \\ -ln\, \alpha_j, & i > m, \end{cases}
$$

we obtain from (9), (3), and (10)

$$
(11) \quad MAP(\mathbf{x}) = argmin_{\sigma} \sum_{i=1}^{n} c_{i,\sigma(i)},
$$

a standard linear assignment problem. Although the size of the solution space is of the order of $n!$ this problem can be efficiently solved by the Hungarian method, cf. Papadimitriou and Steiglitz [1982], or by Balinski's [1985] algorithm. The connection between linear assignment and constrained Bayesian classification was discovered by Tso and Graham [1983] and developped to an instrument of karyotyping by Tso et al. [1991].

If $S = \mathbf{R}^d$ and if the random variables Z^k_j are elliptically symmetric the resulting classifiers are extensions of the classifiers IEC_{φ} based on the assumptions of independent objects and elliptically-contoured feature vectors introduced in Ritter et al. [1995]. We call the new classifiers $IECO_{\varphi\,\psi}$, the letter O standing for *outlier* and the subscripts φ and ψ indicating types of radial functions of the regular and outlier densities, respectively.

The most useful radial functions are the functions $\varphi_{normal}(r) = \beta_{nor}\, e^{-r^2/2}$, $\varphi_{exponential}(r) = \beta_{exp}\, e^{-\sqrt{d+1}\,r}$, and $\varphi_{Pearson_{VII}}(r) = \beta_{Pear}\,(1+r^2/\eta)^{-\lambda/2}$; they belong to the normal distribution, the spherically–symmetric distribution with exponential tail, and distibutions with asymptotic Pareto tails $\beta\, r^{-\lambda}$ like the Pearson-VII distribution, respectively. In the latter distibutions the exponent $\lambda > d+2$ is a free parameter and η depends on it. For a more detailed description of radial functions we refer the interested reader to Ritter et al. [1995], Subsection 3.2.

If, after estimation of the class $\sigma(i)$ of object i, we wish to estimate its category a possible estimator is $argmax_k p_{\sigma(i),k} f_{Z^k_{\sigma(i)}}(x_i)$. The estimator of class *and* category obtained in this way is, however, not the Bayesian estimator.

4 Application to automatic chromosome classification

We finally apply the foregoing material to automatic classification of human chromosomes in the presence of outliers where each chromosome is represented by d numerical features.

4.1 Possible numerical aberrations

A normal human cell contains 46 chromosomes, 44 of which consist of 22 matching pairs of autosomal chromosomes 1..22 1..22 and the sex chromosomes XY and XX in male and female cells, respectively. However, some cells represent abnormal constellations and sometimes there are artifacts of preparation and culture. These aberrations usually cause a cell to contain fewer or additional chromosomes.

A biologically pathological constellation with one *missing* chromosome is *Turner's syndrome* 1..22X1..22. Other cases of missing chromosomes are usually artifacts; here the missing chromosome may be of any class. *Extra* chromosomes are contained in cells with autosomal trisomies or pathological constellations of sex chromosomes. The main trisomies are Down's syndrome (triple 21), Edward's syndrome (triple 18), Pätau's syndrome (triple 13), and the trisomies triple 14, triple 15, triple 16; they occur in both males and females. The pathological constellations of sex chromosomes are XXX, XYX, and XYY. Other syndromes, which are usually not viable up to the 12th week, are not considered here.

From this discussion it is clear that we need three classifiers: One for cells with 46 chromosomes and two classifiers for cells with one missing and one extra chromosome, respectively. We now turn to the design of these classifiers.

4.2 Cells containing 46 chromosomes

A cell containing 46 chromosomes will be assumed to be a normal male or female cell. In the female case it consists of 23 matching pairs of homologous chromosomes, the 23rd pair consisting of two X-chromosomes. In a male cell one X-chromosome is replaced by a Y-chromosome. Thus there are 24 classes, namely 1 to 22 and X,Y which will be identified with classes 23 and 24, respectively. For the sake of applying the estimator of Subsection 3.4, where only one object per class is allowed, we introduce the 46 (virtual) classes 1..46, classes j and $j+23$ being identical for $j \in 1..23$ in the female case and for $j \in 1..22$ in the male case. Therefore, $m = n = 46$ here. We assume that chromosomes of each class may be of two qualities: regular (REG) or outlier (OUT). Each chromosome is represented by d real-valued features, i.e., $S = \mathbf{R}^d$ and ρ is d-dimensional Lebesgue measure λ^d.

We illustrate the range of the MAP-estimator 3.4(10),(11) by handling, besides classes, also outliers and sex. Classes 1..45 are divided into two categories, namely the qualities REG and OUT, class 46 possesses the four possible categories of quality and sex f or m,

$$C_j = \begin{cases} \{REG, OUT\}, & 1 \le j \le 45, \\ \{REG, OUT\} \times \{f, m\}, & j = 46. \end{cases}$$

In Ritter et al. [1995], we have shown that the *elliptically-symmetric* family, cf. Fang et al. [1990], is a successful type of class–conditional distributions for chromosome classification. Thus, we

assume

$$(12)\quad f_{Z_j^k} = \begin{cases} f_j^{(q)}, & 1 \le j \le 23, \quad k = q, \\ f_{j-23}^{(q)}, & 24 \le j \le 45, \quad k = q, \\ f_{23}^{(q)}, & j = 46, \quad k = (q,f), \\ f_{24}^{(q)}, & j = 46, \quad k = (q,m), \end{cases}$$

with

$$(13)\quad f_j^{(q)}(z) = \frac{1}{2} ln\, det\, V_j^{(q)} - ln\, \varphi_j^{(q)}((z - e_j^{(q)})^T (V_j^{(q)})^{-1/2}(z - e_j^{(q)})),$$

$1 \le j \le 24$, $q \in \{REG, OUT\}$, $z \in \mathbf{R}^d$. The parameters $e_j^{(q)} \in \mathbf{R}^d$, $V_j^{(q)} \in GL(d)$, and $\varphi_j^{(q)} : \mathbf{R}_+ \to \mathbf{R}_+$ are the expectations, the variances, and the radial functions of the biological classes j of quality q, respectively.

We also need the prior probabilities $p_{j,REG}$ and $p_{j,OUT} = 1 - p_{j,REG}$ of a chromosome of class $j \in 1..46$ to be regular or an outlier and the probabilities $p_{46,(q,s)} = p_{46,q} p_s$, $q \in \{REG, OUT\}$, $s \in \{f, m\}$, cf. Subsection 3.2. No α_j's for missing classes are needed here.

4.3 Cells containing 45 chromosomes

We assume that a cell containing 45 chromosomes is a normal female or male cell with one chromosome of any class missing. Therefore, the parameters of Subsection 4.2 apply here, too, in particular, there are the probabilities $p_{j,q}$, $1 \le j \le 46$ and $p_{46,(q,s)} = p_{46,q} p_s, q \in \{REG, OUT\}, s \in \{f, m\}$. All cases of missing classes could be covered by assuming that one of the classes 1..23 is missing. However, this approach would not permit to introduce the correct prior probability of Turner's syndrome. We, therefore, open class 46, too, as a possibly missing class arranging matters so that the constellation 1..22 1..22Y is reflected by the missing class 23 and Turner's syndrome by missing class 46.

Therefore, the difference between the present case and that described in Subsection 4.2 is $m = 45$ and the appearance of parameters α_j, $j \in 1..23 \cup \{46\}$, to be chosen in such a way that the random mapping Φ_{n-1} appearing in 3.3(6) satisfies $P[j \notin Im\, \Phi_{n-1}] = \alpha_j$. This means that α_j is the probability of absence of class j if $j \in 1..22$, the overall probability of constellation 1..22 1..22Y if $j = 23$, and the overall probability of constellation 1..22 X 1..22 (Turner's syndrome or artifact) if $j = 46$. Denoting by p_{23} the probability of the missing X-chromosome in a *male* cell, we have $\alpha_{23} = p_{23} p_m$ and denoting by p_T the probability of occurrence of the constellation 1..22 X 1..22 in a *female* cell, we also have $\alpha_{46} = p_T p_f$.

Leaving aside qualities in the following discussion there are three assignments competing for Turner's syndrome, namely

$$\begin{array}{llllll} i & \to & 23, & 46 & \to & (46, f), \\ i & \to & 23, & 46 & \to & (46, m), \text{ and} \\ 46 & \to & 23, & i & \to & (46, f). \end{array}$$

The weights contributed by chromosome i and the dummy chromosome 46 are

$$\begin{array}{l} -(ln\, f_{23}(z_i) + ln\, \alpha_{46} + ln\, p_{46,f}), \\ -(ln\, f_{23}(z_i) + ln\, \alpha_{46} + ln\, p_{46,m}), \text{ and} \\ -(ln\, f_{23}(z_i) + ln\, \alpha_{23} + ln\, p_{46,f}), \end{array}$$

respectively. Since $\alpha_{46} > \alpha_{23}$ and $p_{46,m} > p_{46,f}$, both inequalities for biological reasons, the second one will always be chosen.

4.4 Cells containing 47 chromosomes

We assume that a cell containing 47 chromosomes belongs to either one of the known trisomies (Down, Edwards, Pätau, 14, 15, 16) or one of the constellations XXX, XYX, XYY (cf. Subsection 4.1). We handle these situations by putting $m = n = 47$. Classes 1..46 and categories $C_j, j \in 1..46$, are as before and class 47 consists of the categories $C_{47} = \{REG, OUT\} \times \{13, 14, 15, 16, 18, 21, X, Y\}$. Likelihood functions are, mutatis mutandis, as described in (12). There are no classes missing and hence no α_j's. The prior probabilities $p_{47,(q,t)}$, $(q,t) \in C_{47}$ can be represented as products $p_{47,q}q_t$, where q_t is the probability of occurrence of a trisomy t relative to all trisomies, if $t \in \{13, 14, 15, 16, 18, 21\}$, q_X is the relative probability of occurrence of the XXX constellation, and q_Y is the relative probability of occurrence of the XYY constellation. Instead of relative probabilities one can take overall probabilities (i.e., ith respect to all cells).

Similarly as in the case of a missing chromosome there is ambiguity concerning the constellation XYX. Again leaving aside qualities, it can be represented by two assignments

$$i_1 \to 23, \quad i_2 \to (46, m), \quad i_3 \to \{47, X\} \text{ and}$$
$$i_1 \to 23, \quad i_3 \to (46, f), \quad i_2 \to \{47, Y\}.$$

The weights contributed by i_1, i_2, i_3 are

$$-(ln\, f_{23}(z_{i_1}) + ln\, f_{24}(z_{i_2}) + ln\, p_{46,m} + ln\, f_{23}(z_{i_3}) + ln\, q_X)$$

in the first case and

$$-(ln\, f_{23}(z_{i_1}) + ln\, f_{23}(z_{i_3}) + ln\, p_{46,f} + ln\, f_{24}(z_{i_2}) + ln\, q_Y)$$

in the second case. If $p_{46,m}q_X > p_{46,f}q_Y$, which we assume, then the second assignment causes the larger total weight and does not appear as a minimal solution and the first leads to the correct weight of this constellation.

5 Experimental results

We have implemented the algorithms described in Sections 4 and 3 in the Programming Language C on a SUN workstation SPARC 10. IEC works at a speed of about 4 cells/sec. By (10), the IECO-models require both likelihoods $f_j^{(REG)}(x_i)$ and $f_j^{(OUT)}(x_i)$, cf. (13), for computing one entry $c_{i,j}$, $i \leq m$, of the table of weights; hence, the new method takes almost twice as long.

5.1 Data set

For our experiments we used the Edinburgh feature data of the large Copenhagen data set Cpr; cf. Piper [1992]. It consists to date of 2,804 karyotyped metaphase amnion cells 1,344 of which are female and 1,460 are male, i.e., contain a Y-chromosome. These 2,804 cells consist of 2,740 cells with 46 chromosomes, there is one chromosome missing in 26 cells, 37 cells possess one extra chromosome, and one cell contains 48 chromosomes. Besides these normal cells and 'normal aberrations' there are also strange constellations. There are, e.g., cells containing 46 chromosomes but only one chromosome of some autosomal class and a trisomy. (Since these do not contribute substantially to the overall classification error we did not consider them as possible cases in our classifiers although this could be done.) Among the cells with a missing chromosome

there are 8 Turner syndromes, the set of cells with 1 extra chromosome is composed of 15 Down, 7 Edwards, and two Pätau syndromes, one trisomy 15, one trisomy 16, two Klinefelter syndromes, four XYY-constellations, and five cells containing unclassified chromosomes.

Each chromosome is described by 30 normalized features $0, \ldots, 29$ extracted by the Edinburgh MRC chromosome analysis system described in Piper and Granum [1989]. These features contain information about size, density, convex hull perimeter, centromeric index, shape, and band pattern; cf. Granum [1982] and Piper et al. [1980] 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 so that $d = 24$.

5.2 Estimation of model parameters

The elliptically-symmetric family, cf. (13), was chosen for $\hat{\mu}_{BAS}$ and $\hat{\mu}_{OUT}$. A natural choice for the indicator of outliers χ, cf. Subsection 2.2, is the Mahalanobis distance. The usual sample expectations and sample covariances were used as parameters. For *classification* (but not for parameter estimation), the variances of both the basic and the outlier populations were multiplied outside the diagonal by the Stein factors shown in Tables 5.1 and 5.2. Best results are achieved if $\hat{\mu}_{BAS}$ is assumed to be normal. Classification is insensitive to the radial function of $\hat{\mu}_{OUT}$ at low values of r, values smaller than 5, say. It is, however, important to match the empirical radial function at higher values. Since the graph of the negative logarithm of the empirical radial function of the outlier population is sigmoidal, radial functions with Pareto tails for $\hat{\mu}_{OUT}$ yield the best classification results; cf. the end of Subsection 3.4.

Properly speaking, the free exponent λ_{OUT} of the Pareto tail depends on the outlier population and hence on the cutoff cut_{OUT} chosen. It, however, turns out that it is only very weakly sensitive to the cutoff. The exponent can be estimated by plotting the negative logarithm of the empirical radial function. This, in turn, can be estimated with the aid of the empirical radial density, cf. Ritter et al. [1995], Subsection 4.2. The exponent $\lambda_{OUT} = 30.5$ matches the tail of the outliers best. The empirical radial function, which can be estimated by histogram-and-smoothing techniques, did not yield better results. The optimal cutoffs are $cut_{BAS} = 8.5$ and $cut_{OUT} = 7.0$; cf. Subsection 2.2. A study of the sensitivity of classification results to these parameters is shown in Table 5.3.

For the sake of comparison we present in Tables 5.1 and 5.2 also the results for normal and exponentially–tailed $\hat{\mu}_{OUT}$ and also the results for IEC_{φ}. The exponent $\lambda = 28$ was chosen in the case IEC_{Par}. This completes the description of the estimates of class–conditional distributions.

As prior probabilities (cf. Subsections 4.2–4.4) we chose $p_{j,REG} = 0.95$ for all $j \in 1..46$ and, in accordance with standard cytogenetical tables:

- $p_m = 0.55$, $p_f = 0.45$;
- $p_j = 1/4,000$, $1 \leq j \leq 22$, $p_{23} = 1/8,000$, $p_T = 1/2,000$;
- $q_{13} = 1/5,000$, $q_{14} = q_{15} = q_{16} = 1/10,000$, $q_{18} = 1/3,000$, $q_{21} = 1/700$, $q_X = q_Y = 1/1,000$.

5.3 Estimated probabilities of misclassification

The (optimistic) error rate obtained by using the training set also as test set (in the present case all cells) is called the *resubstitution* error rate. For assessing *holdout* (cross–validation) error

rates we used the following jackknifing method: We divided the 2,803 cells with at most 47 chromosomes at random into 7 disjoint test sets of 400 or 401 cells each. For each of these test sets the parameters of the classifiers were estimated by means of the remaining cells. The holdout error rate is the arithmetic mean of the errors obtained from the 7 test sets. The *Bayes error* of the methods can be expected to lie between the holdout and the resubstitution error rates; cf. Fukunaga [1990]. The overall error rates are displayed in Table 5.1.

Table 5.1. Comparison of overall holdout and resubstitution error rates for the classifiers IEC and IECO with the Stein factors 0.9 and 1.0; cf. Subsection 2.3. All classifiers are constrained on the correct number of chromosomes in each class; cf. Sections 3 and 4. The classifiers IEC are based on the elliptically–contoured distributional models of feature vectors, their types 'normal', 'exponential', and 'Pareto–tailed' being indicated as subscripts, cf. the end of Subsection 3.4. The new classifiers IECO are based on mixtures of normal and elliptically–contoured distributions, instead; cf. Subsection 3.4. The expression p/q means that p% of chromosomes and q% of cells were misclassified. Standard deviation with respect to q is roughly 1. In the mean, there are between three and four chromosomes misclassified in a misclassified cell. It follows that a standard deviation with respect to p is roughly 0.1. The table shows that the classifiers $IECO$ are superior to the classifiers IEC. The best classifier reported here is $IECO_{norPar}$ with a Stein factor of 0.9.

(a) IEC

	IEC_{nor}	IEC_{exp}	IEC_{Par}
holdout (%) Stein factor 0.9	2.68/34.5	2.13/28.8	1.84/25.2
holdout (%) Stein factor 1.0	2.97/37.3	2.31/30.7	1.94/26.3
resubstitution (%) Stein factor 0.9	2.55/33.3	1.99/27.3	1.72/24.0
resubstitution (%) Stein factor 1.0	2.75/35.6	2.12/28.8	1.83/25.1

(b) IECO

	$IECO_{nor\,nor}$	$IECO_{nor\,exp}$	$IECO_{nor\,Par}$
holdout (%) Stein factor 0.9	1.50/21.4	1.39/20.4	1.32/19.3
holdout (%) Stein factor 1.0	1.62/22.8	1.50/21.6	1.45/20.9
resubstitution (%) Stein factor 0.9	1.19/17.9	1.13/17.1	1.09/16.5
resubstitution (%) Stein factor 1.0	1.21/18.1	1.14/17.3	1.09/16.6

Error rates in the cells containing 45 or 47 chromosomes are about twice as large as the overall error rates shown in Table 5.1. The numbers of correctly recognized constellations among the 32 pathological constellations with one extra chromosome are shown in Table 5.2. Table 5.3 shows a study of the *sensitivity* of classification results to the parameters cut_{BAS}, cut_{OUT}, and λ_{OUT} in

the case of the classifier $IECO_{norPar}$. It shows that classification results are almost insensitive to alterations of cut_{BAS} and λ_{OUT} by ± 1; an alteration of cut_{OUT} by ± 1 increases the error rate by only 0.1.

Table 5.2. Number of correctly recognized pathological constellations with 47 chromosomes. The data set Cpr contains 32 pathological constellations with one extra chromosome. By 'correctly recognized' we mean that the three chromosomes determining the particular constellation have been correctly classified; other chromosomes in this cell may be misclassified. Standard deviation in Table 5.2 is 2. The table shows that the classifiers described in Subsection 4.4 recognize most of these constellations correctly.

(a) IEC

	IEC_{nor}	IEC_{exp}	IEC_{Par}
holdout (%) Stein factor 0.9	27	28	28
resubstitution (%) Stein factor 0.9	27	28	28

(b) IECO

	$IECO_{nor\,nor}$	$IECO_{nor\,exp}$	$IECO_{nor\,Par}$
holdout (%) Stein factor 0.9	29	29	29
resubstitution (%) Stein factor 0.9	30	30	29

Table 5.3. Sensitivity of classification results to the parameters cut_{BAS}, cut_{OUT}, and λ_{OUT}, cf. Subsections 2.2 and 3.4, in the case of the estimator $IECO_{norPar}$. The first line represents the optimal parameters; in the other lines, one parameter differs from the optimal one by at most ± 1. The table shows that classification results are almost insensitive to small deviations of cut_{BAS} and λ_{OUT} from their optimal values and only weakly sensitive to the parameter cut_{OUT}.

cut_{BAS}	cut_{OUT}	λ_{OUT}	$holdout(\%)$	$resubstitution(\%)$
8.5	7.0	30.5	1.32/19.3	1.09/16.5
8.5	7.0	29.5	1.33/19.6	1.07/16.3
8.5	7.0	31.5	1.33/19.4	1.08/16.3
8.5	6.0	30.5	1.40/20.4	1.19/17.6
8.5	6.5	30.5	1.36/19.6	1.10/16.6
8.5	7.5	30.5	1.37/19.9	1.07/16.3
8.5	8.0	30.5	1.43/20.5	1.07/16.3
7.5	7.0	30.5	1.33/19.2	1.10/16.3
9.5	7.0	30.5	1.33/19.6	1.08/16.4

Acknowledgements. We thank Dr Jim Piper for kindly supplying the Edinburgh feature data set and for a number of conversations. We are indebted to Herr Harth of Zytogenetisches Labor, Passau, for his assistance in medical and biological questions. We finally thank the referees whose suggestions improved the paper.

References

Anscombe, F.J. (1960). Rejection of outliers. Technometrics 2, 123-147.

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

Barnett, V., T. Lewis (1994). Outliers in Statistical Data, 3rd Edition. Wiley, Chichester, New York, Brisbane, Toronto, Singapore.

Bock, H. H. (1996). Probabilistic models in cluster analysis. Computational Statistics and Data Analysis, 23, 5–28.

Cheeseman, P., J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman (1988). AutoClass: A Bayesian Classification System. Proceedings 5th Int. Conf. Machine Learning, The University of Michigan, Ann Arbor, June 12–14, 1988. 54–64.

Davies, L., U. Gather (1993). The Identification of Multiple outliers. J. Amer. Stat. Association 88(423), 782–792.

Dempster, A.P., N.M. Laird, and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1–38.

Dixon, W. J. (1953). Processing Data for Outliers. Biometrics, 9, 74–89.

Fang, K.-T., S. Kotz and K.W. Ng (1990). Symmetric Multivariate and Related Distributions. Chapman and Hall, London.

Friedman, J.H. (1989). Regularized discriminant analysis. J. Amer. Statistical Ass. 84, 165–175.

K. Fukunaga (1990). Introduction to Statistical Pattern Recognition. Academic Press, Boston, San Diego, New York, London, Sydney, Tokyo, Toronto.

Gallegos, M.T., G. Ritter (1996). Outlier Treatment: A New Statistical Method for Automatic Chromosome Classification. COMPSTAT 96, Conference Abstract.

Granum, E. (1982). 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 (Proc. NATO ASI Series), Reidel, Dordrecht, 373–398.

Hawkins, D.M. (1980). Identification of Outliers. Chapman and Hall, London, New York.

James, W., and C. Stein (1961). Estimation with quadratic loss. Proc. Fourth Berkeley Symp. Math. Statist. and Prob., 1, 361–379.

Kleinschmidt, P., I. Mitterreiter, and J. Piper (1994). Improved chromosome classification using monotonic functions of Mahalanobis distance and the transportation method. ZOR 40, 305–323.

Ledley, R.S. and F.H. Ruddle (1966). Chromosome analysis by computer. Scientific American, 214 (4), 40–46.

Mardia, K.V., J.T. Kent, and J.M. Bibby (1979). Multivariate Analysis. Academic Press, London, New York, Toronto, Sydney, San Francisco.

Papadimitriou, C.H., and K. Steiglitz (1982). Combinatorial Optimization. Prentice–Hall, Englewood Cliffs, New Jersey.

Piper, J. (1987). The effect of zero feature correlation assumption on maximum likelihood based classification of chromosomes. Signal Processing 12, 49–57.

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

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

Piper, J. , E. Granum, D. Rutovitz, and H. Ruttledge (1980). Automation of chromosome analysis. Signal Processing 2, 203–221.

Piper, J., I. Poole, and A. Carothers (1994). Stein's Paradox and Improved Quadrat-ic 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) 529–532.

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

Rousseeuw, P.J. and A.M. Leroy (1987). Robust Regression and Outlier Detection. Wiley, New York, Chichester, Brisbane, Toronto, Singapore.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Neumann J. (ed) Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, December 1954. University of California Press, Berkeley, CA.

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

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

Tukey, J. W. (1960). A Survey of Sampling from Contaminated Distributions. In Olkin, I. (ed.) Contributions to Probability and Statistics. University Press, Stanford, California.