Computer Assisted Centrifugal Elutriation. II

Multiparametric statistical analysis

Sloot, P.M.A.; van der Donk, E.H.M.; Figdor, C.G.

DOI
10.1016/0169-2607(88)90101-0

Publication date
1988

Document Version
Final published version

Published in
Computer Methods and Programs in Biomedicine

Computer Methods and Programs in Biomedicine 27 (1988) 37–46
Elsevier

CPB 00915

Section I. Methodology

Computer-assisted centrifugal elutriation. II. Multiparametric statistical analysis

Peter M.A. Sloot, Emile H.M. Van der Donk and Carl G. Figdor

Division of Biophysics, Netherlands Cancer Institute, Amsterdam, the Netherlands

A combination of *non-interactive* statistical methods is discussed to analyze multiparametric light-scatter data obtained by means of computer-assisted centrifugal elutriation.

Statistics; Centrifugal elutriation; Blood cells

1. Introduction

Computer-assisted centrifugal elutriation (CACE) is a new technique to monitor the separation of large numbers of human peripheral blood cells. It facilitates tuning of the separation process by means of on-line information on the number and type of cells that are elutriated. In addition it allows the detailed study of light-scatter phenomena of well-defined (sub)populations of cells in flow. In a previous report, we described the development of both the optical system and the stand-alone computer constituting the CACE equipment [1]. During the centrifugal elutriation process, three 6-bit parameters (forward-scatter (FS), side-scatter (SS) and back-scatter (BS)), of each sampled cell, are detected and accumulatively stored into a local memory (512 kByte). The content of this memory is continuously displayed for on-line interpretation and can be dumped, by means of a local network, to a host computer * for off-line analysis. In this paper we describe the special-purpose off-line software developed to analyze and interpret, by *non-interactive* methods, the large amount of information present in the data obtained from elutriation.

First, we introduce a modified 'linear separation method' to estimate the initial parameters of each population, with no essential limitations on the number of distributions constituting the histograms. Subsequently, a 2-parameter expectation-maximization (EM) algorithm is applied to optimize the estimation of the statistical parameters describing the various (sub)populations. Preliminary results obtained with the complete CACE system, including the off-line software, indicate that CACE is well suited to monitor and optimize the centrifugal elutriation process. In addition, the off-line software allows rapid and reliable differentiation of the cells in the eluted fractions which closely resembles (time-consuming) histological differentiation after May Grünwald Giemsa staining.

2. Background

In this section arguments are provided for the specific methods applied to analyze multidimen-

Correspondence: P.M.A. Sloot, Division of Biophysics, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands

* LoVME 68010 (Microproject BV, the Netherlands) running the UNIX operating system.

sional data obtained from a CACE experiment. Both nonparametric or parametric analysis may be used to unravel these multidimensional histograms.

Nonparametric analysis can be applied when no presumption is allowed on the statistics of the distributions [2,3]. For example, *interactive* procedures, which determine the clusters constituting the histogram and which estimate the number of cells within a selected area, are commonly used [4–6]. Several nonparametrical methods such as (simulated) randomization tests have been applied in the literature to study the contrasted groups within a multidimensional histogram. These methods allow characterization of statistically different subpopulations [3,7,8]. Furthermore, various statistical tests have been proposed to analyze different sets of histograms [2,3,9–12]. However, a major disadvantage related to nonparametrical analysis, in our application, is that it cannot detect hidden distributions. As a consequence, the influence of small morphological changes on the light-scatter characteristics of the cell populations, cannot be studied in detail. Moreover, detailed investigation on the influence of the elutriation parameters (fluid flow and density, rotor speed, temperature and rotor chamber design) on the complete data-set, requires an even more accurate characterization of the subpopulations.

A second approach to unravel the CACE data is to apply parametric analyses [13,14]. Here it is assumed that the density function, which describes the biological spread within each individual cell population, can be represented by a Gaussian distribution [15,16], and that small deviations from Gaussian properties of the intensity functions are due to alignment difficulties of the cells in flow [17], or errors in the beam-shaping optics [18,19]. Since the construction of the CACE equipment facilitates compensation for possible variations in the hydro-focused sample flow, and a uniform laser-field is guaranteed by cylindrical lenses [1], no significant deviations from normal-distributed light-scatter intensities are expected. Besides, the measurement of light-scatter signals instead of fluorescence signals (commonly used in flow cytometry (FCM)) and linear amplification in contrast to logarithmic amplification, will contribute to the (symmetric-) Gaussian profile of the distributions. From these considerations, parametric analysis of the intrinsic multivariate normal distributions is justified.

Parametric analysis of one-dimensional (FCM) histograms, including iterative procedures has been reported in the literature [12,15,20,21]. In the sequel, we extend the parametric analysis in one-dimension to parametric analysis of multivariate normal distributions by application of the expectation-maximization (EM) algorithm [22–26]. Estimation of the initial parameters and of the number of distributions constituting the multivariate mixture, and approximation of the relevant area to which a subpopulation is confined, greatly determines the reliability of the iteratively calculated parameters. Furthermore, the time required for an iterative procedure to converge to an optimal estimation of the parameters is extremely dependent on the quality of the initial guesses. Therefore, we developed a method, derived from the 'fixed increment rule for linear separation' [27,28], which facilitates rapid *non-interactive* numerical calculation of the number of subpopulations and of the initial estimates of the multivariate parameters.

3. Theory and computational methods

In a CACE experiment, three parameters (FS, SS and BS) of a detected cell are stored [1]. This 4-dimensional (3-parameter) density problem can be described by a mixture of multivariate normal distributions $P_i(\bar{r} \mid \phi)$, as was argued in the previous section:

$$P(\bar{r} \mid \Phi) := \sum_{i=0}^{m-1} \alpha_i \mathrm{p}_i(\bar{r} \mid \phi_i); \quad \bar{r} = (x,\ y,\ z) \in \mathbb{R}^3 \tag{1}$$

Where $0 \leq x,\ y,\ z < 64$ represents the FS, SS and BS channels respectively. The number and fraction of the component populations is represented by m and α. The multivariate distribution is

parameterized by:

$$\Phi := (\alpha_0, \alpha_1, \ldots \alpha_{m-1}; \quad \phi_0, \phi_1, \ldots \phi_{m-1}) \tag{2}$$

where

$$\phi_i := \left(\bar{\mu}_i, \bar{\bar{\Sigma}}_i\right) \tag{3}$$

The bar denotes a vector in $\mathbb{R}^3$ and the double bar denotes a matrix in $\mathbb{R}^3$.

Hence, ϕ_i completely describes the statistical parameters of the subpopulations which must be calculated.

$\bar{\mu}_i$ is the vector of the mean and $\bar{\bar{\Sigma}}_i$ is the (co)variance matrix of a single distribution i, defined by:

$$\bar{\bar{\Sigma}}_i := \begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix} \tag{4}$$

σ_{pq} is the variance ($p = q$) or the covariance ($p \neq q$) term of the distribution.

3.1. Estimation of the initial parameters

Since it is convenient in a CACE experiment to study two-parameter 'scatter-plots' [1], a linear dimension reduction is applied to the raw 3-parameter data:

$$\mathrm{p}_{x,y}(x, y) := \sum_{z=0}^{63} p(\bar{r})$$

$P_{x,y}(x, y)$ is defined by:

$$p_{x,y}(x, y) := \alpha \frac{1}{2\pi |\bar{\bar{\Sigma}}|^{1/2}} \exp\left(-1/2\left(\bar{\mu}\bar{\bar{\Sigma}}^{-1}\bar{\mu}^T\right)\right) \tag{5}$$

where

$$\bar{\mu} := \frac{\sum_{\bar{r}} \bar{r} D(\bar{r})}{\sum_{\bar{r}} D(\bar{r})}$$

and

$$\sigma_{xy} := \frac{\sum_{\bar{r}} (x - \mu_x)(y - \mu_y)^T D(\bar{r})}{\sum_{\bar{r}} D(\bar{r})} \tag{6}$$

$D(\bar{r})$ represents the number of cells in channel $\bar{r}(\bar{r} := (x, y))$ and $|\bar{\bar{\Sigma}}_i|$ is the determinant of the positive definite symmetric 2×2 matrix containing the (co)variance terms.

It is assumed that the intrinsic Gaussian properties of the distributions are conserved. Other (nonlinear) dimension-reduction algorithms may be considered, e.g. quenching the off-diagonal elements of Eq. 4. However, this type of dimension reduction is not applied, since the *interactive* procedures involved are time consuming and lack clarity when mixed distributions are concerned [21,29].

Next, the clusters confining a single population are estimated. Interpretation of the data, however, is hampered by the presence of noise superimposed on the light-scatter signal. Both instrumental and stochastic noise may contribute to the histograms. Instrumental noise is present as a consequence of (known) unavoidable nonsystematic instrumentational errors [1,18,30], whereas stochastic noise may arise from a statistically insufficient number of cells [31]. In a previous paper we described a special-purpose low-pass digital filter to reduce this predominantly high-frequency noise from the data [32]. After application of this filter procedure, the following clustering algorithm is applied for each $P_{p,q}(\bar{r} \mid \phi)$ ($p, q \epsilon \{x, y, z\}$, $p \neq q$):

First, a two-dimensional scan calculates the zero's of the first-order derivatives of the density distribution, with respect to the two principal axes. This set of zero's is modified according to the following criteria:

No local distribution is expected when:

(i) The corresponding density is less than a user-defined percentage of the highest density encountered.

(ii) The Euclidean distance to the surrounding zero's is less than a user-defined resolution value (this results in merging of the distributions).

(iii) The width of the distribution, defined by the distance to the nearest minimum, is less than a preset value.

These criteria are justified since a cellular subpopulation has a certain homogeneous biological spread [30,32] and since the resolution which can be obtained from optical-scatter methods is limited [18,33,34]. Next, a table is produced that contains the overlap vectors. An overlap vector is defined by the percentage of overlap of the distribution under consideration with a surrounding local maximum (i.e. the relative height of the minimum between the distributions), and the relative orientation of this maximum. The overlap table is used to calculate the span and the 'purest' part of each distribution. Finally, the initial parameters of the populations are calculated in accordance with Eq. 6, and extrapolated to describe the complete distribution, by using symmetry conditions for the bivariate Gaussian density function (Eq. 5). The advantage of this local integration technique is clear when the estimated parameters are used to calculate the bivariate fit of each detected distribution, since limitation of the integration area substantially reduces the computational requirements. To reduce possible errors in this initial parameter estimation, and to detect hidden distributions, the procedure is repeated after subtraction of the fitted distributions from the complete data-set. It was concluded, from preliminary experiments, that the data of a CACE experiment never contained more than 20 distributions. If more distributions are detected, the sensitivity of the clustering algorithm is automatically reduced by means of a more restricted definition of a (sub)population. In addition, the cut-off frequency of the low-pass digital filter is diminished. Application of this clustering algorithm results in a set of Φ (Eq. 2) that parametrize the three mixed bivariate normal density distributions (FS vs. SS, FS vs. BS and SS vs. BS). Optimization of the initial parameters is established by means of a two-dimensional iteration procedure, as discussed below.

3.2. The EM algorithm for bivariate data-sets

With the development of computing facilities, interest has grown in maximum-likelihood (ML) techniques for estimating the parameters that determine a mixture density [22,26]. ML procedures applied to univariate histograms were recently discussed in literature [20,25,35]. ML methods guarantee, in contrast to least-squares minimum-distance methods, statistical consistency and efficiency. Furthermore, invariance with respect to one-to-one transformation of the parameters is in general not fulfilled for least-squares techniques.

In this section, we applied a special iterative procedure, the expectation-maximization (EM) algorithm, to determine numerically the ML estimates of multivariate mixture densities. The global convergence of the procedure is discussed, and an empirically derived function is described which is helpful to predict the number of iterations (i.e. computation time) to obtain a required accuracy.

From Eq. 1 we obtain the log-likelihood function [22,23,36]

$$\ln L(\Phi) = \sum_{\bar{r}} D(\bar{r}) \ln \sum_{i=0}^{m-1} \alpha_i P_i(\bar{r} \mid \phi_i) \tag{7}$$

where the bivariate normal density $P_i(r \mid \phi_i)$ is defined by Eq. 5.

Next, by setting the partial derivatives of the log-likelihood function to zero, the values of the parameters which maximize this function are calculated:

$$\bar{\nabla}_{(\alpha_i, \bar{\mu}_i, \bar{\bar{\Sigma}}_i)} L(\Phi) := 0$$

Solving this equation for α_i, $\bar{\mu}_i$ and $\bar{\bar{\Sigma}}$, results in the following iteration scheme:

$$\alpha_i^{(r+1)} = \alpha_i^{(r)} \frac{\sum_{\bar{r}} \frac{D(\bar{r})}{f(r)} P_i^{(r)}(\bar{r} \mid \phi_i)}{\sum_{\bar{r}} \frac{D(\bar{r})}{f(\bar{r})} P_0^{(r)}(\bar{r} \mid \phi_i)} \tag{8a}$$

$$\bar{\mu}_i^{(r+1)} = \frac{\sum_{\bar{r}} \frac{D(\bar{r})}{f(\bar{r})} P_i^{(r)}(\bar{r} \mid \phi_i) \bar{r}}{\sum_{\bar{r}} \frac{D(\bar{r})}{f(\bar{r})} P_i^{(r)}(\bar{r} \mid \phi_i)} \tag{8b}$$

$$\overset{=(r+1)}{\sum_i} = \left[\sum_{\bar{r}} \frac{D(\bar{r})}{f(\bar{r})} P_i^{(r)}(\bar{r}\,|\,\phi_i)\left(\bar{r} - \bar{\mu}_i^{(r+1)}\right) \times \left(\bar{r} - \bar{\mu}_i^{(r+1)}\right)^T\right] \times \left[\sum_{\bar{r}} \frac{D(\bar{r})}{f(\bar{r})} P_i^{(r)}(\bar{r}\,|\,\phi_i)\right]^{-1} \tag{8c}$$

where:

$$f(\bar{r}) = \sum_{i=0}^{m-1} \alpha_i P_i(\bar{r}\,|\,\phi_i)$$

and

$$\alpha_0 := 1 - \sum_{i=1}^{m-1} \alpha_i \tag{8d}$$

$(r+1)$ is a label that indicates the next iteration step.

Note that these estimators differ fundamentally from the one-parameter analogue recently reported in literature [20,25]:

$$\sigma^{(r+1)} = \frac{\sum_x \frac{D(x)}{f(x)} P_i^{(r)}(\bar{r}\,|\,\phi_i)\left(x - \mu^{(r)}\right)^2}{\sum_x \frac{D(x)}{f(x)} P_i^{(r)}(\bar{r}\,|\,\phi_i)}$$

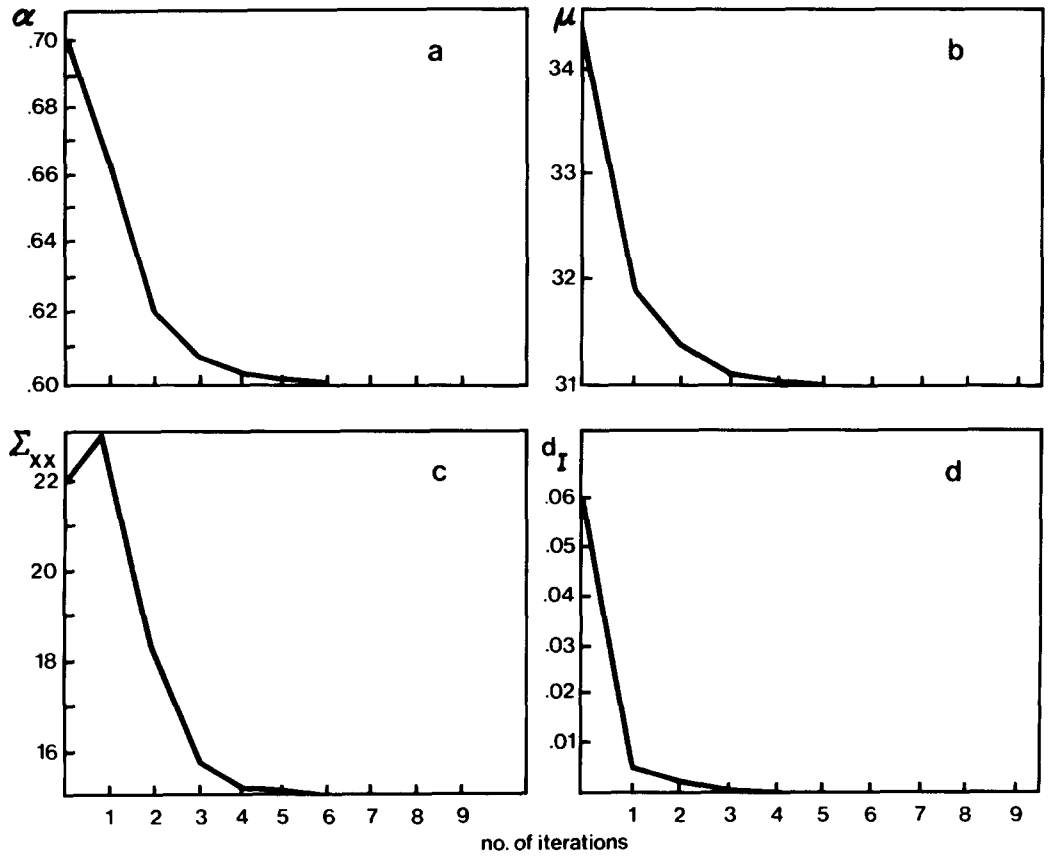

Fig. 1. Various distribution parameters versus number of iterations. Initial distance between the two mixed distributions: $d_B = 2.0$. d_I indicates the relative Euclidean distance between two successively iterated distributions.

Where a one-parameter modification of our notation is used. Here, $\sigma_i^{(r+1)}$ results from calculation of $\mu_x^{(r)}$, whereas in our application $\sigma_i^{(r+1)}$ results from $\mu_x^{(r+1)}$. It is obvious that insertion of a previously calculated $\mu_x^{(r+1)}$ into the calculation of $\sigma_i^{(r+1)}$ results in a faster iteration process; therefore, Eq. 8c is applied. Deviation in the calculation of Eq. 8a from one-parameter analogues reported in the literature [22,23,25] is due to the definition of the normalization equation of $\{\alpha\}$ in Eq. 8d.

In the remaining part of this section, the characteristics of the iteration procedure defined by Eqs. 8a–d are studied. We define a measure for Euclidean distance between two populations $P_1(\bar{r}\,|\,\phi_1)$ and $P_2(\bar{r}\,|\,\phi_2)$ by means of the so-called Bhattacharyya distance (dB) [37]:

$$d_B := -\ln\int_{-\infty}^{+\infty}\left(P_1(\bar{r}\,|\,\phi_1)P_2(\bar{r}\,|\,\phi_2)\right)^{1/2}\,\mathrm{d}\bar{r} \qquad (9)$$

Where again our notation is used. Calculation of dB from Eqs. 5 and 9 results in:

$$d_B = \frac{1}{8}\left[(\bar{\mu}_2-\bar{\mu}_1)\left(\frac{\bar{\bar{\Sigma}}_1+\bar{\bar{\Sigma}}_2}{2}\right)^{-1}(\bar{\mu}_2-\bar{\mu}_1)^T + \tfrac{1}{2}\ln\left(\frac{\left|\frac{\bar{\bar{\Sigma}}_1+\bar{\bar{\Sigma}}_2}{2}\right|}{\left|\bar{\bar{\Sigma}}_1\right|^{1/2}\left|\bar{\bar{\Sigma}}_2\right|^{1/2}}\right)\right]$$

Fig. 1 shows the relative changes in the estimators determined from 9 successive iteration steps for a particular separation of two simulated multivariate distributions with $d_B = 2.0$. The Bhattacharyya distance between the simulated data and the initial fit is denoted by d_F ($= 0.06$), whereas the Bhattacharyya distance between consecutively iterated distributions is denoted by d_I. It is observed that the relative fraction α, the vector of means ($\bar{\mu}$) and the elements of the (co)variance matrices ($\bar{\bar{\Sigma}}$) converge after approximately 6 iterations. Fig. 1d, in addition, shows self-stabilization of d_I. Since, d_I contains the weighted (statistical) parameters of the successive generated distributions, it can be applied as an unambiguous criterion for the termination of the iteration process. Calculation for other values of d_B, d_F and d_I gave comparable results (data not shown).

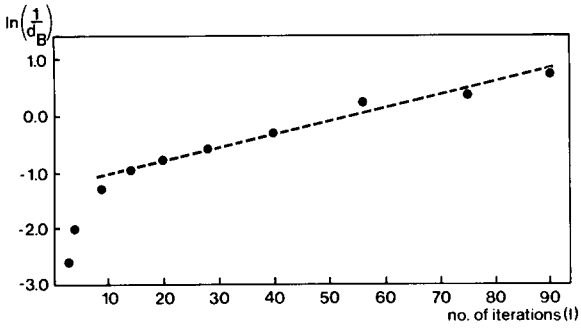

Fig. 2. Logarithm of the inverse distance d_B versus the number of iterations between two mixed distributions: ● = measured values; – – – = linear regression fit for $I > 5$.

The global convergence with respect to the number of iterations was calculated for different values of d_B with a fixed d_F (0.18) and a predefined stop-criterion $d_I = 2 \times 10^{-5}$. Calculation of the coefficient of correlation (ρ), for $\ln(1/d_B)$ versus the number of iterations (I), resulted in $\rho = 0.96$ (Fig. 2) (for $I > 5$, 10 observations). Therefore, the number of iterations necessary to obtain self-stabilization can be predicted, if the distance d_B is known. From pilot experiments, it became apparent that a typical minimum Bhattacharyya distance, of approximately $d_B = 2.0$ can be expected between two adjacent distributions. Consequently, the corresponding number of iterations is approximately 20 (Fig. 2). Finally, the influence of the initial estimates on the iteration process has been studied by calculating the number of iterations for a specific d_B at various values of d_F. No significant changes in the number of iterations for a relevant range of d_F could be detected.

4. Results

The methodology described in the preceding sections is illustrated by the performance of a typical

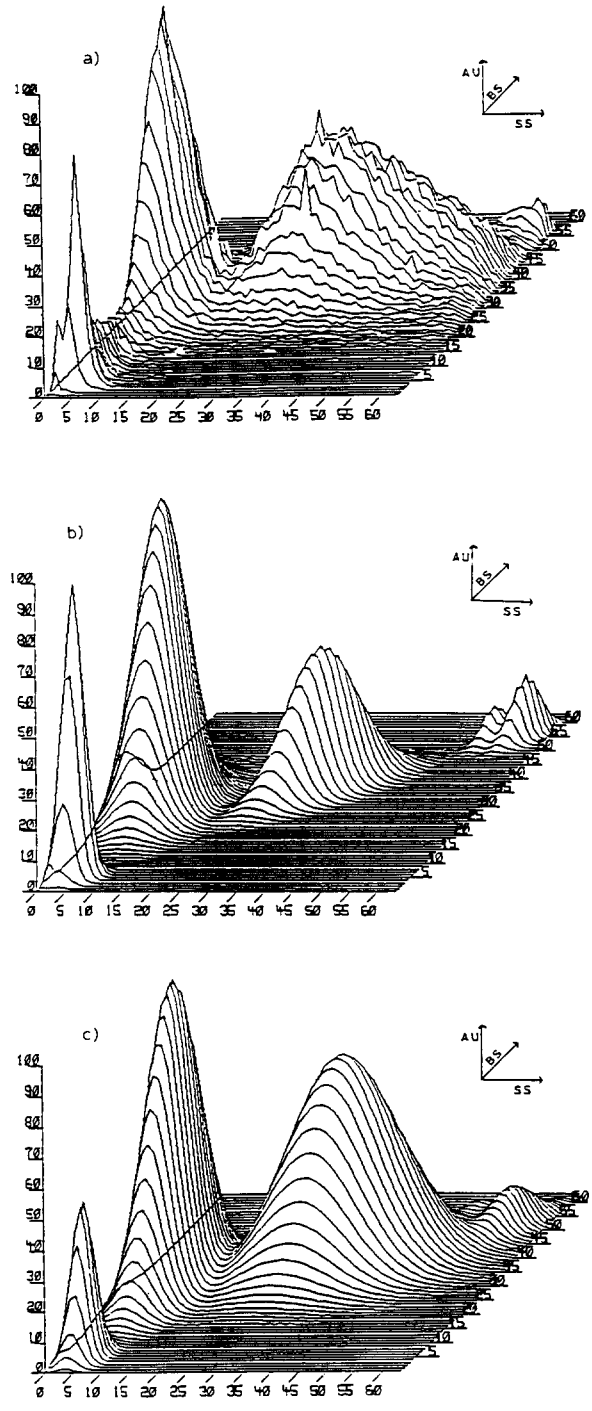

Fig. 3. Bivariate histogram of a mixed monocyte/neutrophilic fraction. SS, side-scatter; BS, back-scatter. Number of cells in arbitrary units (AU): (a) initial histogram; (b) fitted distribution after application of the clustering procedure; (c) fitted distributions after 16 iterations.

CACE experiment [1]. A mononuclear cell suspension was prepared from 500 ml of human peripheral blood by means of a blood-component separator, as described previously [38]. Mononuclear cells were suspended in phosphate-buffered saline supplemented with 0.15% bovine serum albumin (fraction V; Sigma, St. Louis, MO) penicillin (100 IU/ml) and streptomycin (50 μg/ml), and introduced into the elutriator rotor. CE was performed by means of two JE-6 elutriation rotors, equipped with standard separation chambers, in series. Approximately 800×10^6 leukocytes and 1200×10^6 erythrocytes were injected into a cooled sample introduction unit and loaded into the first rotor at a flow rate of 12 ml/min and at a rotor speed of 3800 rpm. After introduction of the cell sample the flow rate was set at 18 ml/min. Fractionation was obtained by stepwise decreasing the rotor speed, whereas the speed of a second rotor was maintained at 4500 rpm to collect and concentrate the cells elutriated from the first rotor. The fractionation resulted in a subsequent elution of erythrocytes, lymphocytes, monocytes and neutrophilic granulocytes. Details of the (CA)CE equipment, the cell separation procedure and the data aquisition, are described elsewhere [1,39].

To facilitate calculations, linear dimensions reduction is applied to the raw data, the result of which, for BS vs. SS, is shown in Fig. 3a. Two major populations and at least two minor clusters can be identified. The initial fit, after application of the clustering algorithm, is shown in Fig. 3b. The algorithm discriminates between 5 different distributions D[0]..D[4], where D[1] corresponds to the monocyte population and D[2] corresponds to the neutrophilic granulocyte population (Table 1). After the clustering procedure, the ML–EM iterative algorithm is applied. Self-stabilization occurred after 16 iterations, as is shown in Fig. 3c. The data clearly indicate the extreme improvement of the iterated distributions in comparison with the initial fit (χ^2 changes from 23342 to 5771 (Table 1)). Distribution D[0] contains mainly erythrocytes, whereas the biological characterization of the small distributions D[3] and D[4] is still under investigation. The abundance of the major populations corresponds qualitatively to data obtained from differentiation according to microscopical methods.

TABLE 1

Statistics of the distributions detected in a monocyte/granulocyte fraction

Initial parameters: Chi-square = 23 342; number of iterations = 0					
Distribution number	D[0]	D[1]	D[2]	D[3]	D[4]
Vector of mean channel number (SS, BS)	(4,5)	(7,30)	(32,36)	(54,54)	(59,55)
Covariance matrix					
(Σ_{ss} Σ_{sb})	3 1	14 0	23 15	2 2	3 3
(Σ_{bs} Σ_{bb})	1 2	0 29	15 19	2 4	3 5
Relative proportion (α)	0.064	0.683	0.222	0.009	0.021
After iteration termination: Chi-square = 5 771; number of iterations = 16					
Distribution number	D[0]	D[1]	D[2]	D[3]	D[4]
Vector of mean channel number (SS, BS)	(4,6)	(8,29)	(35,37)	(51,52)	(57,51)
Covariance matrix					
(Σ_{ss} Σ_{sb})	3 1	10 5	76 18	8 1	5 5
(Σ_{bs} Σ_{bb})	1 4	5 33	18 34	1 9	5 17
Relative proportion (α)	0.032	0.323	0.614	0.01	0.020

TABLE 2

Triangular part of symmetric Bhattacharyya-distance matrices

Initial values					After 16 iterations			
	D[0]	D[1]	D[2]	D[3]	D[0]	D[1]	D[2]	D[3]
D[1]	5.2				3.6			
D[2]	11.8	4.3			7.3	2.2		
D[3]	146.5	37.1	5.1		78.0	26.2	1.6	
D[4]	136.3	41.7	7.1	2.3	88.2	40.7	1.9	0.9

The influence of the iteration procedure on the mutual Bhattacharyya distances is shown in Table 2. The mutual distance between all estimated distributions is reduced by the iteration procedure. This is in line with observations of other experiments (data not shown). Increasing the integration area of each distribution during the clustering procedure slightly reduces this phenomenon. The computational time required to obtain the same accuracy, however, becomes unacceptably large. It can be derived from Table 2 that the computational time is mainly determined by the overlap of distribution D[4] with its nearest neighbour distributions ($d_B = 2$ implies approximately 20 iterations; see Fig. 2). To obtain an accuracy defined by the iteration criterion $d_1 = 2 \times 10^{-5}$, a mean computational time of approximately 15 seconds per iteration is required. Since the mean number of iterations is approximately 10, most mixed bivariate distributions can be calculated within 1.5–3.0 minutes. The programs were written in the language C on different computers running the UNIX operating system. The sources and a detailed outline of the algorithms are available upon request from the authors.

5. Conclusions

In this paper we discussed multiparametric statistical analysis of intrinsic Gaussian distributions by means of *non-interactive* methods.

It is observed that application of a special-purpose clustering algorithm resulted in a reliable first-order estimation of the initial parameters of the bivariate normal distributions. A modified maximum-likelihood–expectation-maximization algorithm (ML–EM) has been developed from statistical considerations. The convergence of this iterative procedure was studied numerically by means of simulated bivariate data. We introduce a new differential distance measure, that includes all significant statistical parameters of two successive distributions (d_I), to define a criterion for self-stabilization of the iteration process. Furthermore, a linear relationship was detected for the number of iterations versus $\ln(1/d_B)$, where d_B indicates the Euclidean distance between two adjacent distributions ($d_B > 3.0$). Numerical calculations showed that global convergence occurred irrespective of the initial misfit (d_F) introduced by the clustering algorithm. Consequently, the main purpose of the clustering procedure is to estimate the exact number of initial distributions. The programs were modified in accordance with this observation (small integration areas and reliable detection of hidden distributions). The applicability of the off-line software described here was tested in a number of CACE experiments. A typical example is discussed in the preceding sections. It was demonstrated that the combination of rapid clustering followed by ML–EM results in a powerful method to discriminate statistically between the various (sub)populations detected in a CACE experiment, with no limit to the number of distributions contained in a multivariate histogram. Moreover, application of the algorithms to other computational fields where multiparametric normal distributions are studied may be considered.

In the near future, a number of items will be studied:

(1) The assumption that the bivariate projections may be regarded as normal distributions is not proved extensively.

(2) The information obtained after iteration can be applied to determine the eigen-vectors of the (co)variance matrices. Subsequently, non-linear dimension reduction by diagonalization may be considered.

(3) Detailed comparison between the biological characterization of the (sub)populations must be established.

(4) Improvement of the computational speed may be accomplished by means of an extremely small integration area for each detected distribution. In addition, time-consuming modules will be converted to (optimized) assembly code.

(5) The extension of the noninteractive procedures described in the preceding sections to three-parameter analogues will be considered.

Acknowledgements

The authors thank Ir. A.A.M. Hart and Dr. R.W. De Boer for critical reading of the manuscript. The research was financially supported by STW grant No. LGN 260353.

References

[1] P.M.A. Sloot, M.J. Carels, P. Tensen and C.G. Figdor, Computer-assisted centrifugal elutriation. I. Detection system and data acquisition equipment, Comput. Methods Programs Biomed. 24 (1987) 179–188.

[2] M. Hollander and D.A. Wolfe, Nonparametrical Statistical Methods (John Wiley and Sons, New York, 1970).

[3] W.J. Conover, Practical Nonparametric Statistics. (John Wiley and Sons, New York, 1971).

[4] R.F. Murphy, Automated identification of subpopulations in FCM list mode data cluster analysis, Cytometry 6 (1985) 302–308.

[5] V. Kachel and H. Schneider, Flow cytometry and other applications, Cytometry 7 (1986) 25–40.

[6] V. Kachel, Interactive multi-window integration of two-parameter flow cytometric data fields, Cytometry 7 (1986) 89–92 (Technical Note).

[7] M. Recchia and M. Rocchetti, The simulated randomization test, Comput. Programs Biomed. 15 (1982) 111–116.

[8] R.C. Mann and R.E. Hand Jr., The randomization test applied to flow cytometric histograms, Comput. Programs Biomed. 17 (1983) 95–100.

[9] C.B. Bagwell, J.L. Hudson and G.L. Irvin III, Nonparametric flow cytometry analysis, J. Histochem. Cytochem. 27 (1979) 293–296.

[10] I.T. Young, Proof without prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources, J. Histochem. Cytochem. 25 (1977) 935–941.

[11] G. Clayton, On the random-pairs method of density estimation, Biometrics 40, (1984) 199–202.

[12] J.V. Beck and K.J. Arnold, Parametric Estimation in Engineering and Science (John Wiley and Sons, New York, 1970).

[13] M.A. Van Dilla, Ph.N. Dean, O.D. Laerum and M.R. Melamed (eds), Flow Cytometry: Instrumentation and Data Analysis (Academic Press, New York, 1985).

[14] K. Fukunaga, Introduction to Statistical Pattern Recognition (Academic Press, New York, 1972).

[15] O.D. Laerum, T. Lindmo and E. Thorud (eds), Chapter 5: Mathematical analysis of DNA histograms, in: Flow Cytometry Vol. IV, (Universitetsforlaget, Bergen-Oslo-Tromso, 1980).

[16] G. Valet, H. Hofmann and G. Ruhenstroth-Bauer, The computer analysis of volume distribution curves: demonstration of two erythrocyte populations of different size in the young guinea pig and analysis of the mechanism of immune lysis of cells by antibody and complement, J. Histochem. Cytochem. 24 (1976) 231–246.

[17] T.M. Jovin, S.J. Morris, G. Striker, H.A. Schultens, M. Digweed and D.J. Arndt-Jovin, Automatic sizing and separation of particles by ratios of light scattering intensities, J. Histochem. Cytochem. 24 (1) (1976) 269–283.

[18] J.A. Steinkamp, M.J. Fulwyler, J.R. Coulter, R.D. Hiebert, J.L. Horney and P.F. Mullaney, A new multiparametric separator for microscopic particles and biological cells, Rev. Sci. Instrum. 44 (1973) 1301.

[19] P.N. Wild and J. Swithenbank, Beamstop and vignetting effects in particle size measurements by laser diffraction, Appl. Optics 25 (1986) 3520–3527.

[20] R.C. Mann, R.E. Hand Jr. and G.R. Braslawsky, Parametric analysis of histograms measured in flow cytometry, Cytometry 4 (1983) 75–82.

[21] R.C. Mann, D.M. Popp and R.E. Hand Jr., The use of projections for dimensionality reduction of flow cytometric data, Cytometry 5 (1984) 304–307.

[22] R.A. Redner and H.F. Walker, Mixture densities, maximum likelihood and the EM algorithm, SIAM Rev. 26 (1984) 195–239.

[23] V. Hasselblad, Estimation of parameters for a mixture of normal distributions, Technometrics 8 (1966) 431–444.

[24] K. Meyer, Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices, Biometrics 41 (1985) 153–165.

[25] M. Agha and M.T. Ibrahim, Maximum likelihood estimation of mixtures of distributions, Appl. Stat. 203 (1984) 227–332.

[26] B.C. Peters Jr. and H.F. Walker, An iterative procedure for obtaining maximum likelihood estimates for a mixture of a normal distributions, SIAM J. Appl. Math. B 35 (1978) 362–378.

[27] G.C. Salzman, J.M. Crowell, C.A. Goad, K.M. Hansen, R.D. Hiebert, P.M. LaBauve, J.C. Martin, M.L. Ingram and P.F. Mullaney, A flow-system multiangle light-scattering instrument for cell characterization, Clin. Chem. 21 (1975) 1297–1304.

[28] M. Minsky and S. Papert, Perceptrons, An Introduction to Computation Geometry (MIT Press, Cambridge MA, 1969).

[29] P. Jolicoeur, Principal components, factor analysis, and multivariate allometry: a small-sample direction test, Biometrics 40 (1984) 658–690.

[30] P. Ubezio and A. Andreoni, Linearity and noise source in flow cytometry, Cytometry 6 (1985) 109–115.

[31] W.H. Schuette, S.E. Shackney, M.A. MacCollum and C.A. Smith, A count-dependent filter for smoothing flow cytometric histograms, Cytometry 5 (1984) 487–493.

[32] P.M.A. Sloot, P. Tensen and C.G. Figdor, Spectral analysis of flow cytometric data: design of a special-purpose low-pass digital filter, Cytometry 8 (1987) 545–551.

[33] P.M.A. Sloot and C.G. Figdor, Elastic light scattering from nucleated blood cells: rapid numerical analysis, Appl. Optics 25 (1986) 3559–3565.

[34] J.A. Steinkamp, Flow cytometry, Rev. Sci. Instrum. 55 (1984) 1357.

[35] F.G. Boese, Comments on 'Parametric analysis of histograms measured in flow cytometry' by R.C. Mann et al., Cytometry 7 (1986) 224–226.

[36] J.R. Green and D. Margerison, Statistical Treatment of Experimental Data (Elsevier, Amsterdam, 1978).

[37] H.C. Andrews, Introduction to Mathematical Techniques in Pattern Recognition (Wiley-Interscience, New York, 1972).

[38] C.G. Figdor et al., Rapid isolation of mononuclear cells from buffy coats prepared by a new blood cell separator, J. Immunol. Methods 55 (1982) 221–229.

[39] C.G. Figdor et al., A centrifugal elutriation system of separating small numbers of cells, J. Immunol. Methods 68 (1984) 73–87.