Classification Of Disorders Of Anemia On The Basis Of Mixture Model Parameters

Technical Report No. 01–56,
Information and Computer Science Department,
University of California, Irvine

C. E. McLaren[1], I. V. Cadez[2], P. Smyth[2] and G. J. McLachlan[3]
[1]Division of Epidemiology, Department of Medicine,
University of California, Irvine, CA 92697, U.S.A.
[2] Department of Information and Computer Science,
University of California, Irvine, CA 92697, U.S.A.
[3] Department of Mathematics,
The University of Queensland,
Brisbane, Australia
cmclaren@uci.edu, icadez@ics.uci.edu, smyth@ics.uci.edu, gjm@maths.uq.edu.au

November 21, 2001

Abstract

Over one billion people in the world are anemic and at risk for major liabilities. Previous models proposed to differentiate between disorders of anemia on the basis of red blood cell measurements have been limited by the need to use printed output from automated blood sample analyses. We developed electronic methods to capture multivariate red cell data measured by flow cytometric blood cell counting instruments and devised a general bilevel framework for classification that includes (1) fitting mixture densities to the multivariate grouped and truncated distribution for each individual, and (2) discrimination between patient subgroups on the basis of distribution parameter estimates. Data were collected from 90 healthy individuals and 146 patients with anemia. For each subject the joint distribution of red blood cell volume and hemoglobin concentration was modeled as a mixture of two multivariate lognormal distributions. The Expectation-Maximization algorithm was used to fit parameters to grouped and truncated data. Classification into controls and two patient subgroups by fitting normal density models, with leave-one-out cross validation, achieved 98% and 100% correct classification for controls and patients, respectively.

1 Introduction

According to population estimates, over one billion people in the world have anemia, defined as a reduction in the circulating red cell mass that may diminish the oxygen-carrying capacity of the blood. Iron deficiency anemia, attributed to an imbalance between dietary iron supply and physiological requirements for growth and reproduction, is the most common nutritional anemia (DeMaeyer and Adiels-Tegman, 1985.) Major liabilities including mental and motor developmental defects in infants (Oski, 1993) and weakness, weight loss, and impaired work performance in adults (Basta et *al.*, 1979; Edgerton et *al.*, 1979). Other nutritional anemias include vitamin B_{12} deficiency and folate deficiency. The anemia of chronic disorders found in infectious diseases such as tuberculosis, typhoid, and smallpox and in noninfectious disorders including rheumatoid arthritis, Hodgkin disease, metastatic carcinoma, is usually moderate and rarely symptomatic, while thalassemia, a form of severe anemia caused by mutations (or deletions) in or around the globin chain DNA and accompanied by a disturbance of hemoglobin synthesis, may lead to organ damage and premature death. Chronic alcohol ingestion is often associated by anemia as a result of poor nutrition, gastrointestinal bleeding, or the toxic efffect of alcolohol on the production of erythrocytes. Alcoholics may also develop coincident iron deficiency and folate deficiency (Williams, et *al.*, 1990).

Hematologic examination of the blood is a routine procedure to determine the presence of anemia in patients with major illnesses. Typical morphologic classification of anemias is made on the basis of the mean cell volume (Williams, et *al.*, 1990). For example, iron deficiency anemia, α- and β-thallasemias, and in some cases, anemia of chronic disease, may be characterized by microcytosis in which the mean cell volume is below the normal range, while vitamin B_{12} deficiency, folate deficiency, and alcoholic liver disease are often accompanied by macrocytosis with a mean cell volume above the normal range. However, this classification does not take into account the variations in hemoglobin concentration that are also important in the clinical diagnosis of anemia.

Flow cytometric blood cell counting instruments make measurements on *each* red cell using a laser light scattering system. This technology provides the red cell volume distribution, hemoglobin concentration distribution, and the joint red cell volume and hemoglobin concentration distribution. Previous investigators have demonstrated that new flow cytometric technology accurately measures both volume and hemoglobin concentrations over a wide range of mean cell volume (30 to 120 fl) and mean cell hemoglobin concentration (27 to 45 g/dL) values. Mohandas and colleagues (1986) documented that volume and hemoglobin concentration distributions can vary independently of each other in pathologic red cell samples.

Measurement of volume and hemoglobin concentration in individual resealed erythrocytes revealed populations of microcytic and macrocytic cells. Additionally subpopulations of cells were found with either decreased hemoglobin concentration (hypochromic cells) or increased hemoglobin concentration (hyperchromic cells) (Green et *al.*, 1987). While illustrating the potential of the new information inherent in laser light scattering technology, previous studies have been limited by the necessity to use the printed statistical and graphical output provided with blood sample analyses. No generally accepted techniques are currently available for the analysis and interpretation of the underlying joint distributions of cell volume and hemoglobin concentration. Since different causes of anemia may result in characteristic alterations in these distributions, we hypothesized that classification based on modeling of the multivariate distribution of red cell volume and hemoglobin concentration would be useful for diagnostic evaluation of anemia. Our study is the first to model and classify these multivariate distributions.

To provide standardized methodological uniformity for cell size studies, the International Commission for Standardization in Hematology (ICSH) first outlined general principles for analyzing cell size curves (ICHS, 1982). Methods were then developed for fitting a single reference lognormal distribution and assessing its goodness of fit. (ICSH, 1990). Analysis of red blood cell volume data demonstrated the reproducibility of the ICSH reference method and the ability to detect sequential changes in distributions (McLaren, et *al.*, 1993). Methods have been developed for detection of two-component mixtures of lognormal distributions and utilized to characterize and quantify subpopulations of red blood cells in developing iron deficiency anemia and subsequent treatment for the disease (McLaren, 1996; McLaren et *al.*, 2000). While these methods have been applied to analysis of univariate red blood cell volume distributions, no suitable statistical methods are currently available for analysis of multivariate distributions arising from multiple measurements made on a single blood cell, such as the volume and hemoglobin concentration of a red blood cell.

We now describe a general framework that includes the following: (1) development of techniques to model multivariate mixtures of distributions from grouped and truncated data, (2) description of the bivariate distribution of red cell volume and hemoglobin concentration in patients with anemia and controls, and (3) classification of patient subgroups on the basis of distribution parameter estimates. Analysis of data from 90 healthy individuals and 146 patients with documented disorders of anemia showed that mixture modeling on parameter estimates with leave-one-out cross validation, achieved 98% and 100% correct classification for controls and patients, respectively. We conclude that these methods provide a means for automated screening for disorders of anemia and monitoring the response to therapy.

2 Methods

2.1 Patients and Reference Group

This study was performed at the Western Infirmary, Glasgow, Scotland after Institutional Review Board approval was obtained. We collected blood samples from a reference group of healthy individuals and patients with documented disorders of anemia. Diagnoses and body iron status were confirmed by examination of blood films, iron studies, and red cell indices. Reference ranges for red cell indices were as follows: hemoglobin (HGB) 13.5-17.5 g/dL (males), 12-16 g/dL (females); mean cell hemoglobin concentration (MCHC) 33.4-35.3 g/dL; and mean cell volume (MCV) 80-100 fl. We analyzed data from 90 healthy individuals and 146 patients. Patients were divided into two subgroups, those with microcytosis including iron deficiency anemia (n=82), thalassemia (n=8) and anemia of chronic disease (n=16), and those with macrocytosis including vitamin either B_{12}/folate deficiency (n=12), and alcoholic liver disease (n=28). For blood cell analysis, we used a flow cytometric blood cell counting instrument Technicon H*1 (Bayer Diagnostics, Tarrytown, New York, USA), with a laser light scattering system that estimates both the volume and hemoglobin concentration of each red cell and provides the red cell volume distribution, hemoglobin concentration distribution, and joint red cell volume and hemoglobin concentration distributions. For this system a photoelectric cell detects light which is refracted, diffracted, or scattered by cells passing through a small illuminated area in the optical system. The detector generates electrical pulses of magnitude proportional to the size of the particle. Individual red blood cell volume and hemoglobin concentration are determined by simultaneous measurement of light scatter at two different angles. This methodology permits identification of red blood cell populations with different hemoglobin concentrations (Ross and Bentley, 1986; Bollinger et *al.*, 1987). Complete blood count (CBC) values and frequency counts representing the number of cells within two-dimensional regions of red cell volume and hemoglobin concentration were captured electronically and stored for further analysis. All analyses described in this paper were performed on the raw data before instrument processing. For each sample, measured in duplicate, the data obtained consisted of a cytogram, i.e. bivariate histogram, with a range of 0 to 200 fl for cell volume and 0 to 50 g/dl for hemoglobin concentration.

2.2 Mixture Modeling

We developed techniques to model the joint distribution of red cell volume and hemoglobin concentration as a mixture of two multivariate lognormal distributions. A mathematical model of red cell production predicted a lognormal form for the distribution of red blood cell volume (McLaren *al.*, 1986) and red cell hemoglobin concentration distributions exhibit a right-skew (Noe and Bell, 1985). Finite mixture models have been fit to univariate grouped and truncated data by maximum likelihood via the Expectation-Maximization (EM) algorithm (McLachlan and Jones,1988; McLaren et *al.*, 1991). McLachlan and Krishnan (1997) review the principles and methodology of the EM algorithm and discuss aspects of its implementation in many contexts. A comprehensive account of the theory and applications of modeling via finite mixture distributions is provided by McLachlan and Peel (2000). For our studies the Expectation-Maximization (EM) algorithm was extended to evaluate multidimensional integrals over two-dimensional regions and numerical integration techniques were employed to improve computational efficiency [See (Cadez et *al.*, 2002)].

For analysis of data from healthy individuals and patients, we developed a new bilevel modeling technique. We first identified mixtures of two subpopulations of cells within a single blood sample by fitting a two-component lognormal mixture model to each individual distribution. The parameter estimates from each fitted distribution were recorded. These included the mixing weight for the larger proportion, volume and hemoglobin concentration means and variances for each component, and the estimated correlation between volume and hemoglobin concentration for each component. Second, for discrimination between patient subgroups, a supervised classifier for the parameter sets of all subjects from the same disease subgroup (control, microcytic anemia, macrocytic anemia) was used to fit a normal density model to each group.

2.3 Classification

Classification was performed using Bayes rule for the posterior class distribution:

$$p(c_i|x) = \frac{p(x|c_i)p(c_i)}{p(x)} \tag{1}$$

where $p(x|c_i)$ is a probability density function under the i-th single normal subpopulation, $p(c_i)$ is the prior for the disease, estimated by the ratio of patients in the i-th group to the total number of patients, and $p(x)$ is a normalizing constant. Both $p(x|c_i)$ and $p(c_i)$ are estimated from the data as described in Section 2.4. To estimate the overall accuracy of discrimination of patient subgroups, we used leave-one-out cross validation. For this type of cross validation we removed a data point from the data set and trained the three single normal density models. The excluded data point was then evaluated using Bayes rule and assigned to a class. This process was repeated for each of the data points in the data set. The overall percent of correctly classified distributions was calculated.

2.4 Bilevel Model

The bilevel model used in this paper consists of two "levels." The lower (individual) level model consists of a two-component lognormal mixture fitted to each of the individual cytograms. Maximum likelihood estimates of the lognormal mixture parameters are obtained for each individual cytogram using the EM procedure as outlined below. Variability among parameters of different individuals is then modeled at the higher (group) level of the hierarchy by a multivariate normal density function for each of the groups.

2.4.1 Modeling at the Individual Level: EM for Fitting Mixtures to Cytograms

The binned, and in some cases, truncated nature of the cytogram data for each individual requires that the standard EM estimation framework for finite mixtures be somewhat modified. The theory for fitting finite mixture models to such data in the univariate case was developed in full by McLachlan and Jones (1988). Here we present a brief summary of the underlying ideas. The model can be written as:

$$f(x;\Phi) = \sum_{i=1}^{g} \pi_i f_i(x;\theta), \tag{2.1}$$

where the π_i's are weights for the individual components, the f_i's are the component density functions of the mixture model parametrized by θ, and Φ is the set of all mixture model parameters, $\Phi = \{\pi, \theta\}$. The overall sample space $\mathcal{H}$ is divided into v disjoint subspaces $\mathcal{H}_j$, (bins) of which only the counts on the first r bins are observed, while the counts on the last $v - r$ bins are missing. The (observed) likelihood associated with this model (up to irrelevant constant terms) is given by Jones and McLachlan (1990).

$$\ln L = \sum_{j=1}^{r} n_j \ln P_j - n \ln P, \tag{2.2}$$

where n_j is the count in bin j, n is the total observed count $n = \sum_{j=1}^{r} n_j$, and the Ps represent integrals of the probability density function (PDF) over bins:

$$P_j \equiv P_j(\Phi) = \int_{\mathcal{H}_j} f(x;\Phi)dx, \tag{2.3}$$

$$P \equiv P(\Phi) = \int_{\mathcal{H}} f(x;\Phi)dx = \sum_{j=1}^{r} P_j. \tag{2.4}$$

The form of the likelihood function above corresponds to a multinomial distributional assumption on bin occupancy.

In Cadez et *al.* (1999, 2002), we provide a detailed description of how the EM algorithm can be implemented efficiently in the multidimensional case for binned and truncated data and we also demonstrate the application of the method to mixture modeling of cytograms. Numerical efficiency is achieved by leveraging a variety of computational short-cuts at various stages of the EM algorithm. For example, for any fixed sample size, a multivariate histogram will be much sparser than any marginal univariate counterpart, in terms of counts per bin (i.e., marginals). This sparseness can in turn be taken advantage of for the purposes of efficient numerical integration.

2.4.2 Group Level Modeling in Parameter Space

The output of the EM mixture modeling is a set of 11 parameters for each individual that describes a two-component mixture distribution for the red blood cells of that individual. The parameters of this mixture model consist of a mixing weight α, a two-dimensional mean vector μ, and three covariance parameters for each of the two components. For each individual, the two mean vectors μ_1 and μ_2 represent the mean cell volume and mean cell hemoglobin concentration of each of the mixture components. The two covariance matrices Σ_1 and Σ_2 describe the variability in parameter space of the cell volume and the cell hemoglobin concentration for each of the two mixture components.

For the purposes of the higher-level model we use only the weight parameters and the mean parameters in the model, i.e., α, μ_1 and μ_2, for a total of five dimensions. The covariance parameters Σ_1 and Σ_2 did not appear to contain much discriminatory power between the groups and, thus, were omitted from the modeling at this level. Defining $\alpha_{\max} = \max\{\alpha, 1-\alpha\}$ as the larger of the mixing weights, we transform $\alpha_{\max}$ to a log-odds scale, i.e., $\log \frac{\alpha_{\max}}{1-\alpha_{\max}}$. We then model the class-conditional distribution of the 5-dimensional parameter vector (consisting of the log-odds plus two sets of 2-dimensional means) for each group as being a multivariate Normal density, resulting in a 3-component or 6-component Normal mixture depending on the number of groups defined. Since the group labels are known a priori, maximum likelihood parameter estimation for each group is performed in closed form.

The mixing proportions $p(c_i)$, are also estimated by maximum likelihood and represent the prior probabilities of belonging to each group. The maximum likelihood estimates are the proportion of patients in the i-th group. The estimated 5-dimensional mean for a particular group represents the typical log-odds and mean parameters that will be fit to cytograms from that group. The 5×5 estimated covariance matrix for each group represents the variability in the mixing weight and mean vector parameters estimated from the cytogram, for that group. Since a full covariance matrix for five dimensions requires the estimation of 15 parameters per group, we assumed a diagonal covariance model for groups that had less than 50 individuals and used a full covariance matrix for groups with 50 or more individuals. Posterior probabilities of group membership can then be calculated via Bayes rule using the estimated models for each group and the group with the maximum probability is then chosen.

2.4.3 Experimental Methodology

In the experiments in this paper, leave-one-out cross-validation was used to generate the classification results, where each of the 236 individuals were removed from the data set one at a time, the classification models were estimated as described above using the other 235 individuals, and a prediction was made on the individual not used in the estimation process. The results reported in this paper were obtained using MATLAB 6.0 Release 12.1 (The MathWorks, Inc., Natick, MA, USA), and a set of custom Matlab scripts we developed specifically for this task.

3 Results

3.1 Distribution Modeling and Classification

Tables 1 and 2 give descriptive statistics (mean and standard deviation) for complete blood count values measured in the 163 females (Table 1) and 73 males (Table 2) in the reference group.

Table 1: Complete Blood Count Values in 58 Healthy Females and 105 Female Patients with Anemia

CBC Value	**Controls (n=58)**	**IDA (n=68)**	**ACD (n=11)**	**THT (n=3)**	**ALD (n=15)**	**BFD (n=8)**
HGB g/dL	13.2 (0.90)	9.0 (2.28)	10.7(1.06)	11.5(0.76)	12.0 (1.67)	9.0 (2.8)
HCT %	0.40 (0.09)	0.30 (0.06)	0.34 (0.04)	0.38 (0.06)	0.4 (0.05)	0.3 (0.09)
MCV fl	92.8 (3.70)	75.5 (6.59)	77.9 (5.92)	68.4 (3.07)	117.9 (8.90)	122.9 (10.5)
MCHC g/dL	33.4 (1.01)	29.4 (2.39)	31.2 (1.56)	20.9 (2.43)	32.6 (1.59)	33.1(2.40)
RDW %	12.8 (0.43)	16.4 (1.73)	16.0 (1.77)	15.8 (0.12)	15.9 (2.00)	18.9 (3.20)
HDW g/dL	2.3 (0.20)	3.0 (0.42)	2.7 (0.31)	3.1 (0.11)	2.2 (0.33)	2.8 (0.75)

Table 2: Complete Blood Count Values in 32 Healthy Males (Controls) and 41 Male Patients with Anemia

CBC Value	**Controls (n=32)**	**IDA (n=14)**	**ACD (n=5)**	**THT (n=5)**	**ALD (n=13)**	**BFD (n=4)**
HGB g/dL	15.0 (0.89)	9.1 (2.23)	11.9 (0.58)	11.6 (0.97)	12.0 (3.19)	10.0 (1.96)
HCT %	0.5 (0.03)	0.3 (0.06)	0.4 (0.02)	0.4 (0.02)	0.4 (0.09)	0.3 (0.07)
MCV fl	91.5 (3.53)	74.6 (6.83)	79.9 (2.12)	69.3 (4.31)	117.0 (7.51)	122.7 (19.8)
MCHC g/dL	33.7 (1.13)	29.1 (2.10)	31.2 (0.88)	21.7 (1.11)	32.9 (2.02)	32.2 (1.17)
RDW %	12.8 (0.35)	16.7 (2.55)	15.3 (2.59)	15.9 (0.84)	16.2 (1.80)	17.1 (2.96)
HDW g/dL	2.4 (0.16)	3.2 (0.35)	2.9 (0.34)	3.0 (0.32)	2.2 (0.27)	2.7 (0.44)

Table 3: Classification of Healthy Individuals (Controls) and Three Patient Subgroups

		Number of Cases Classified into Group			
Subgroup	**Percent Correct**	Control	Microcytosis	Macrocytosis	Total
Control	97.8%	88	2	0	90
Microcytosis	100.0%	0	106	0	106
Macrocytosis	100.0%	0	0	40	40

Figures 1 and 2 show histograms from two representative subjects. Each distribution represents about 40 000 red blood cells measured on a single blood sample. Figure 1 shows the distribution from a healthy male with estimated geometric mean cell volume and hemoglobin concentration of 89 fl and 34.4 g/dL respectively. For comparison, the distribution shown in Figure 2 is from a female with developing iron deficiency anemia. The bivariate distribution contained a hypochromic, microcytic subpopulation of 58% of cells with an estimated geometric mean red cell volume of 73.9 fl and geometric mean hemoglobin concentration of 28.4 g/dL, both below normal. A hypochromic, normocytic subpopulation (with an estimated 42% of the total cells) had geometric mean red cell volume of 80.9 fl, within the normal range, and geometric mean hemoglobin concentration of 30.2 g/dL.

The distribution and contour plot from a female with alcoholic liver disease with hypochromic, macrocytic anemia are shown in Figures 3 and 4. The larger subpopulation contains 95% of the cells with an estimated geometric mean cell volume of 140 fl and moderately reduced geometric

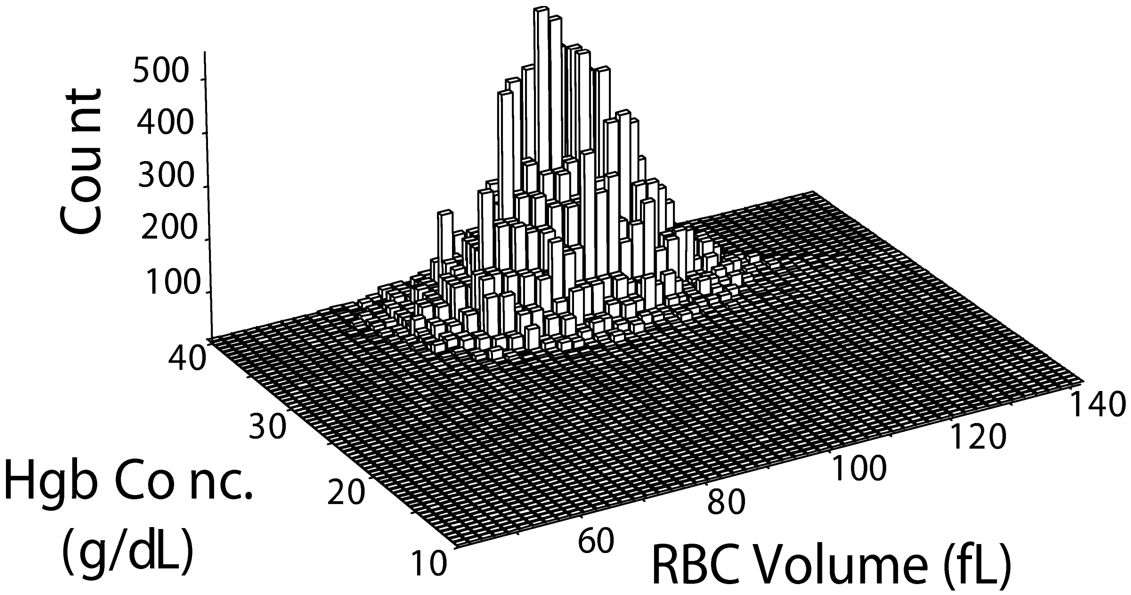

Figure 1: Red blood cell volume and hemoglobin concentration distribution in a healthy male. Parameter estimates: mixing proportion = 1.0, geometric mean cell volume = 89 fl, geometric mean cell hemoglobin concentration = 34.4 g/dL.

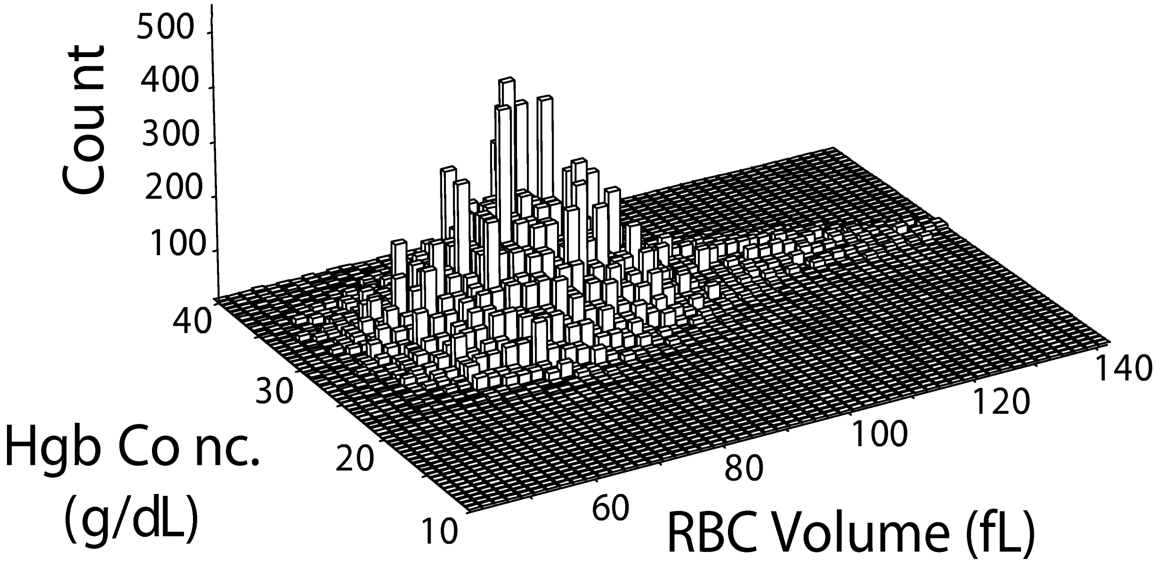

Figure 2: Red blood cell volume and hemoglobin concentration distribution in developing iron deficiency anemia. Parameter estimates: mixing proportion = .58, geometric mean red cell volume = 73.9 fl, geometric mean cell hemoglobin concentration = 28.4 g/dL; mixing proportion = .42, geometric mean red cell volume = 80.9 fl, geometric mean cell hemoglobin concentration = 30.2 g/dL

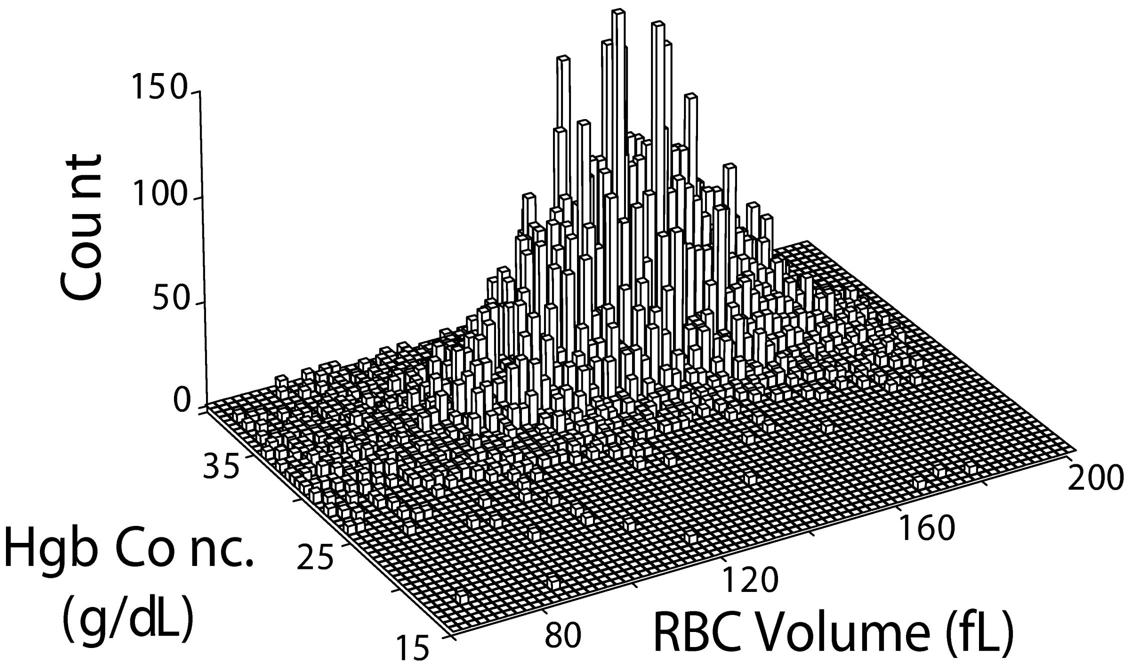

Figure 3: Bivariate distribution for red blood cell volume and hemoglobin concentration in alcoholic liver disease (ALD). Parameter estimates: mixing proportion = .95, geometric mean red cell volume = 140 fl, geometric mean cell hemoglobin concentration = 33.0 g/dL; mixing proportion = .05, geometric mean red cell volume = 95.9 fl, geometric mean cell hemoglobin concentration = 32.0 g/dL

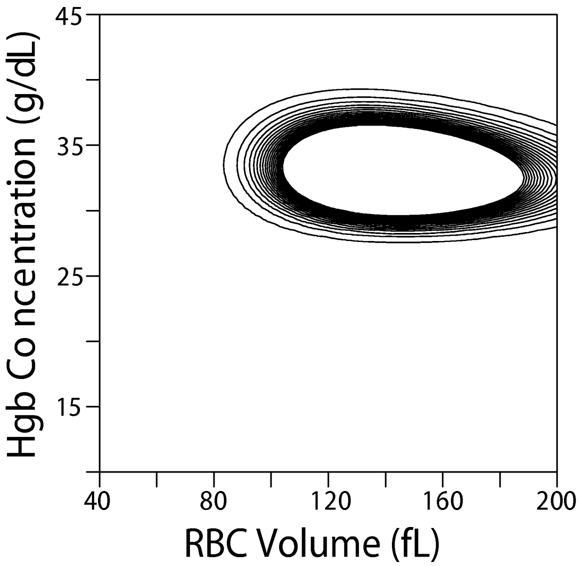

Figure 4: Contour plot for cell volume and hemoglobin concentration in ALD.

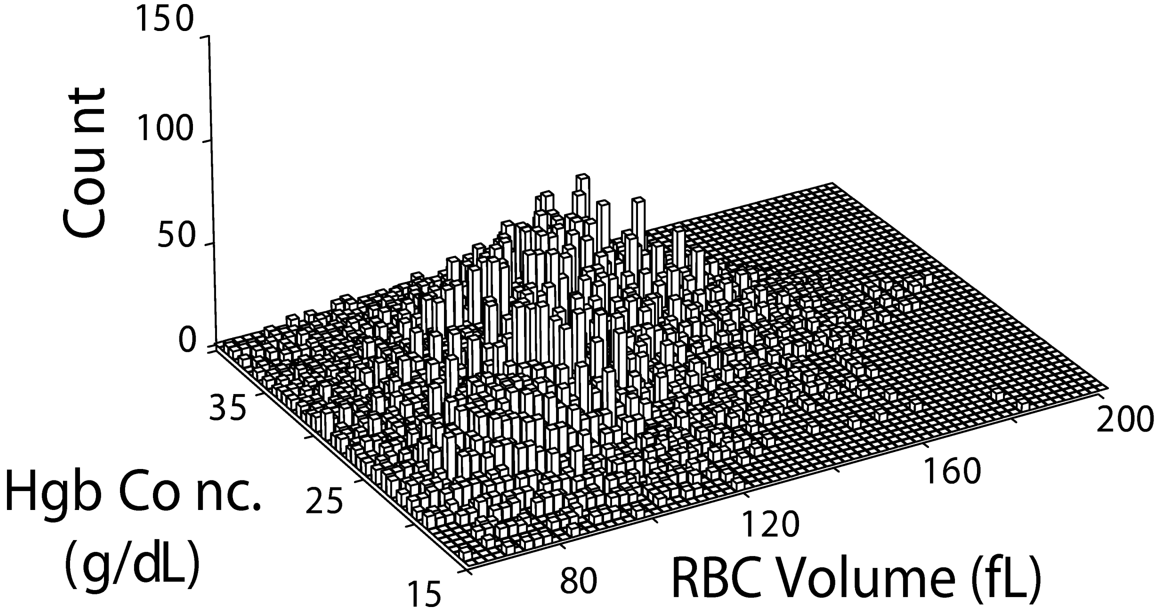

Figure 5: Bivariate distribution for red blood cell volume and hemoglobin concentration in B_{12} deficiency and folate deficiency (BFD). Parameter estimates: mixing proportion = .48, geometric mean red cell volume = 120.7 fl, geometric mean cell hemoglobin concentration = 32.4 g/dL; mixing proportion = .52, geometric mean red cell volume = 94.0 fl, geometric mean cell hemoglobin concentration = 26.0 g/dL

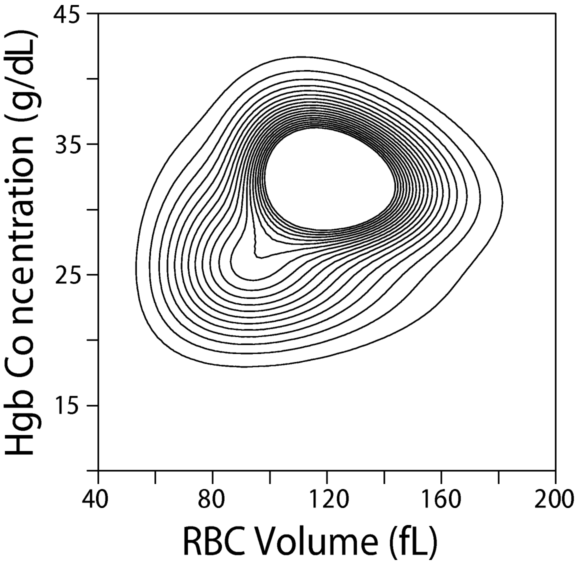

Figure 6: Contour plot for cell volume and hemoglobin concentration in BFD.

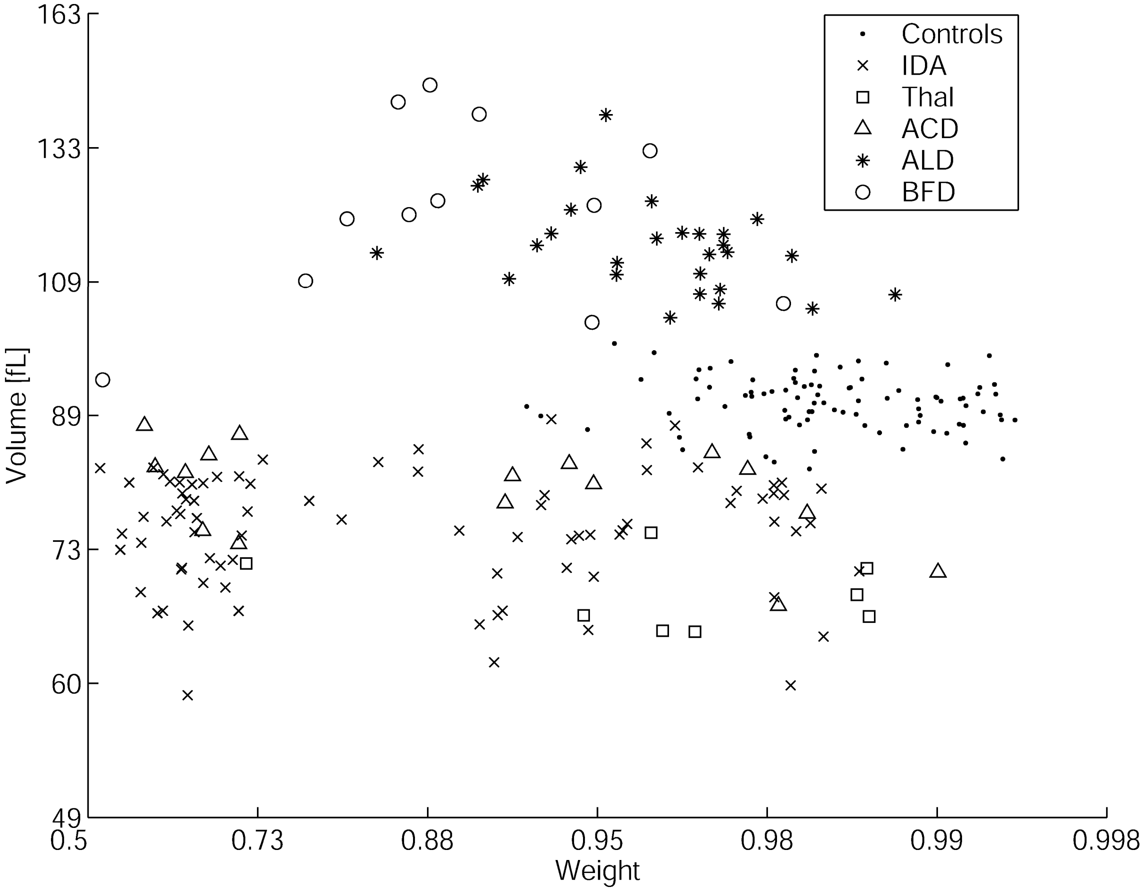

Figure 7: Scatter plot of mean volume and log-odds of the mixture weight α for the largest mixture component.

Table 4: Classification of Controls and 6 Disease Categories. Variability among paramaters of different individuals was modeled by a multivariate density function.

Subgroup	**Percent** Correct	Number of Cases Classified into Group Control	IDA	THAL	ACD	ALD	BFD	Total
Control	97.8%	88	2	0	0	0	0	90
Iron Deficiency Anemia	93.9%	0	77	1	3	0	1	82
Thalassemia	62.5%	0	3	5	0	0	0	8
Anemia of Chronic Disease	0.0%	0	15	1	0	0	0	16
Alcoholic Liver Disease	82.1%	0	0	0	0	23	5	28
B12 Folate Deficiency	58.3%	0	0	0	1	7	4	12

mean hemoglobin concentration of 33.0 g/dL. A remaining 5% of the cells had geometric mean volume (95.9 fl) and hemoglobin concentration (32.0 g/dL) consistent with that of controls. The bivariate distribution (Figure 5) and contour plot (Figure 6) from a female patient with B_{12}/folate deficiency represent a hypochromic, macrocytic, subpopulation of 48% of cells with an estimated geometric mean red cell volume of 120.7 fl and geometric mean hemoglobin concentration of 32.4 g/dL and a hypochromic, normocytic, subpopulation containing 52% of cells with geometric means for red cell volume and hemoglobin concentration of 94.0 fl and 26.0 g/dL respectively.

Figure 7 shows a scatter plot of the mean volume and the log-odds of the mixture weight α for the largest mixture component resulting from the mixture distribution analysis applied to distributions from each of the 236 individuals. Separation of the control, microcytic, and macrocytic populations is reflected in terms of separation along the volume axis. The control population lies roughly between 80 and 100 fl, the reference range for mean red cell volume. The microcytic disorders (IDA, Thal, and ACD) are below 80 fl in general, and the macrocytic disorders (ALD and BFD) are generally above 100 fl. Thus, the control, microcytic, and macrocytic groups appear to be relatively well-separated and accurate classification of these three groups should be possible. However, within the microcytic and macrocytic groups there is considerable overlap of the different sub-groups (between IDA, Thal, and ACD; and between ALD and BFD) in the two-dimensional volume/log-odds space. Similar overlap is also seen to exist in other two-dimensional scatter plots (not shown). Thus, the existence of the ability to perform accurate within-group classification is not apparent from two-dimensional scatter plots of the data.

The control group is further distinguished by having a mixing weight α that is typically closer to 1 (high log-odds, the horizontal axis in Figure 1) than for the other groups. This lends support to the hypothesis that the mixture model is fitting one component to the "normal" population of cells and a second (typically smaller) component to the abnormal population of cells, for each individual. For the control group the larger component tends to be given a much higher weight (more normal cells) than for many of the individuals in the other groups.

Results of classification into controls and two patients subgroups are shown in Table 3, including the number and percent of all patients correctly classified using leave-one-out cross validation.

Distributions from two healthy individuals were misclassified as having anemia. These results are not unexpected, because production of red blood cells is a dynamic process making it difficult to classify distributions falling at the upper or lower limits of a particular subgroup.

For comparison, we also attempted classification of distributions into those of controls and six patient subgroups (Table 4). Among the distributions from patients with microcytic disorders, correct classifications were achieved for 77 of 82 (93.9%) patients with iron deficiency anemia and 5 of 8 (62.5%) patients with thalassemia. None of the 16 distributions from patients with the anemia of chronic disease were correctly classified, but were recognized as similar to those of patients with thalassemia or the anemia of chronic disease. Among the distributions from patients with macrocytic disorders, correct classifications were achieved for 23 of 28 (82.1%) patients with alcoholic liver disease and 7 of 12 (58.3%) patients with B_{12}/folate deficiency. The small sample sizes for some patient subgroups affected the disease classification rates. Including sex in the model might be desirable from the standpoint of clinical interpretation, however we did not do this because of the limited amount of data per class. Our results in general indicate that classification into controls and 2 patient subgroups, for example, can be carried out quite reliably without taking sex of the individual into account.

4 Discussion

We have successfully developed techniques to model and classify multivariate distributions from grouped and truncated red blood cell data. Our study is unique in using mixture models in a bilevel fashion, first for description on an individual subject level and then for classification on a group level. On the individual subject level, the range of distributions in healthy individuals with normocytic, normochromic cells is defined, while in patients with anemia, distributions containing subpopulations of cells with microcytic, normocytic, or macrocytic red blood cell volume and hypochromic or normochromic hemoglobin concentration are described. On the group level, parameter estimates for individual mixture models are then used to distinguish between healthy individuals and patients with anemia.

Discrimination between patient subgroups on the basis of the distribution parameters for hemoglobin concentration and red blood cell showed that controls are well separated from other patients with disorders of anemia (Table 3: 98% overall correct classification). In previous studies, classification of patients with thalassemia trait and iron deficiency anemia (Jimenez et *al.*,1995) or vitamin B12/folate deficiency, alcohol excess/liver disease and reticulocytosis (Harkins et *al.*, 1994) utilized printed statistical and graphical output from flow cytometric blood cell counting instruments and additional clinical information. Our study is the first to model the joint distribution of red cell volume and hemoglobin concentration reflecting measurements of individual blood cells.

Our results demonstrate that bivariate volume and hemoglobin concentration measurements can be used to discriminate control, macrocytic, and microcytic groups to high accuracy. More detailed discrimination at the level of specific anemias appears to be less accurate based on results obtained in this particular study. Whether this is primarily due to inherent ambiguity of the class origins of cytograms in volume and hemoglobin concentration space, or due to inaccuracies in the classification model due to the small sample sizes available at the specific anemia level in this study, remains an open question.

As described in the Sections 2.4, the bivariate mixture models for each individual can be estimated in a straightforward and computationally efficient manner, using the Expectation-Maximization algorithm. The method can be readily implemented in software on a standard PC workstation, or could equally well be embedded within a flow cytometric blood cell counting instrument. We conclude that for individual subjects, these methods may provide a means for monitoring the response to therapy or for automated screening for disorders of anemia.

Acknowledgements

This work has been supported in part by research grants from the National Institutes of Health (R15-HL48349, R43-HL46037) and a Wellcome Research Travel Grant awarded by the Burroughs Wellcome Fund (CEM). Additional support was provided by a National Science Foundation CAREER award, IRI-9703120 (PS, IVC) and a grant from the Australian Research Council, A10027060 (GJM). We are grateful to Brian Ortner and Dr. Albert Greenbaum for technical assistance. We thank Thomas H. Cavanagh for providing laboratory facilities.

References

Basta, S. S., Soekirman, M. S., Karyada, D., Scrimshaw, N.S. (1979) Iron deficiency anemia and the productivity of adult males in Indonesia. *American Journal of Clinical Nutrition,* **32**, 916–925.

Bollinger P.B., Drewinko, B., Brailas, C.D., Smeeton N.A., Trujillo, J. M. (1987) The Technicon H*1–an automated hematology analyzer for today and tomorrow. Complete blood count parameters. *American Journal of Clinical Pathology,* **87**, 71-8.

Cadez, I. V., McLaren, C. E., Smyth, P. and McLachlan G. J. (1999) Hierarchical Models for Screening of Iron Deficiency Anemia. In Bratko, I., Dzeroski, S. eds. *Proceedings of the 1999 International Conference on Machine Learning.* Los Gatos, CA: Morgan Kaufmann, 77–86.

Cadez, I. V., Smyth, P., McLachlan, G. J. and McLaren, C. (2002) Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. *Machine Learning*, in press.

DeMaeyer E. and Adiels-Tegman M. (1985) The prevalence of anemia in the world. *World Health Statistical Quarterly,* **38**, 302–316.

Edgerton V. R., Gardner G. W., Ohira Y., Gunawardena K. A., Senewiratne B. (1979) Iron-deficiency anaemia and its effect on worker productivity and activity patterns. *British Medical Journal,* **2**, 1546–1549.

Green, R., King, R. R., Jacobsen, D. W. and Luce, K. (1987) Direct measurement of volume and hemoglobin concentration of individual resealed erythrocytes sing laser light scattering. In *Advances in the Biosciences*, **67**, 233–241.

Harkins, L. S., Sirel, J. M., McKay, P. J. and Wylie, R. C., Titterington D. M. and Rowan R, M. (1994) Discriminant analysis of macrocytic red cells. *Clinincal and Laboratory Haematolology,* **16**, 225–234.

International Committee for Standardization in Hematology (1982) ICSH recommendations for the analysis of red cell, white cell, and platelet size distribution curves: I. General principles. *Journal of Clinical Pathology,* **35**, 1320–1322.

International Committee for Standardization in Hematology (1990) ICSH recommendations for the analysis of red cell, white cell, and platelet size distribution curves. Methods for fitting a reference distribution and assessing goodness-of-fit. *Clinical and Laboratory Haematology,* **12**, 417–431.

Jimenez, C. V., Minchinela, J. and Ros, J. (1995) New indices from the H*2 analyser improve differentiation between heterozygous b or db thalassaemia and iron-deficency anaemia. *Clinical and Laboratory Haematology,* **17**, 151–155.

Jones, P. N. and McLachlan, G. J. (1990) Maximum Likelihood Estimation from Grouped and Truncated Data with Finite Normal Mixture Models. *Journal of the Royal Statistical Society, Applied Statistics (Series C)*, **39, 273–282.**

McLachlan, G.J. and Jones P.N. (1988) Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics,* **44**, 571–578.

McLachlan, G.J. and Krishnan T. (1997) *The EM Algorithm and Extensions.* New York: John Wiley.

McLachlan, G.J. and Peel, D. (2000) *Finite Mixture Models.* New York: John Wiley.

McLaren, C.E. (1996) Mixture models in hematology. *Statistical Methods in Medical Research,* **5**, 129–153.

McLaren, C.E., Brittenham, G.M. and Hasselblad, V. (1986). Analysis of the volume of red blood cells: application of the expectation- maximization algorithm to grouped data from the doubly-truncaated lognormal distribution. *Biometrics,* **42**, 143–158.

McLaren, C.E., Houwen, B., Koepke, J., Rowan, R.M., Ortner, B.A. and Bishop, M.L. (1993) Analysis of red blood cell volume distributions using the ICSH reference method: detection of sequential changes in distributions determined by hydrocynamic focusing. *Clinical and Laboratory Haematology,* **15**, 173-184.

McLaren, C.E., Wagstaff, M., Brittenham, G.M. and Jacobs, A. (1991). Detection of two component mixtures of lognormal distributions in grouped doubly-truncated data: analysis of red blood cell volume distributions. *Biometrics,* **47**, 607–622.

McLaren, C.E., Kambour, E., McLachlan, G.J., Lukaski, H.C., X. Li, Brittenham, G.M., and

McLaren, G.D. (2000) Patient-specific analysis of sequential hematological data by multiple linear regression and mixture distribution modeling. *Statistics in Medicine,* **19**, 83–98.

Mohandas, N., Kim, Y. R., Tycko, D. H., Orlik, J., Wyatt, J. and Groner, W. (1986) Accurate and independent measurement of volume and hemoglobin concentration of individual red cells by laser light scattering. *Blood,* **68**, 506–13.

Noe, D.A. and Bell, W.R. (1985) A Model of the heterogeneity of red cell hemoglobin concentration. *Computers and Biomedical Research,* **18**, 544-552.

Oski, F. A. (1993) Iron deficiency in infancy and childhood. *New England Journal of Medicine,* **329**, 190–193.

Ross, D.W. and Bentley, S.A. (1986) Evaluation of an automated hematology system (Technicon H*1). *Archives of Pathohology and Laboratory Medicine,* **110**, 803–8.

Williams, W. J., Beutler, E., Erslev, A. J., and Lichtman, M. A. (1988) *Hematology*, fourth edition. New York: McGraw-Hill.