doi: 10.1046/j.1529-8817.2004.00109.x

Inferring the Most Likely Geographical Origin of mtDNA Sequence Profiles

T. Egeland[1,2], H.M. Bøvelstad[3], G.O. Storvik[3] and A. Salas[4]

[1]*Biostatistics, Rikshospitalet, 0027 Oslo, Norway. thore.egeland@basalmed.uio.no*
[2]*Section of Medical Statistics, University of Oslo*
[3]*Department of Mathematics, University of Oslo*
[4]*Unidad de Genética Forense, Instituto de Medicina Legal, Universidad de Santiago de Compostela, Galicia, Spain*

Summary

In a number of practical cases it is important to determine the likely geographical origin of an individual or a biological sample. A dead body, old bones or a sample of semen may be available. Information on where the sample might come from can assist investigation or research. The first part of this paper is independent of specific data structure. We formulate the problem as a classification problem. Bayes' theorem allows different sources of information or data to be reconciled conveniently. The main part of the paper involves high dimensional data for which simple, standard methods are not likely to work properly. Mitochondrial DNA (mtDNA) data is a typical example of such data. We propose a procedure involving essentially two steps. First, principal component analysis is used to reduce the dimension of the data. Next, quadratic discriminant analysis performs the actual classification. A cross validation procedure is implemented to select the optimal number of principal components. The importance of using separate data sets for model fitting and testing is emphasized. This method distinguishes well between individuals with a self reported European (Icelandic or German) origin and SE Africans. In this case the error rate is 2.0%.

Introduction

Several authors have discussed variations of the problem addressed in the title of this paper. To indicate some related topics we quote a few publications, including some titles: 'Assessing ethnicity from human mitochondrial . . . ' (Connor & Stoneking, 1994), 'Probable Race of a Stain Donor' (Brenner, 1997), 'Inference of population structure using multilocus genotype data' (Pritchard *et al.* 2000), 'Inferring ethnic origin by means of an STR profile' (Lowe *et al.* 2001), 'An annotated mtDNA database', (Röhl *et al.* 2001), 'Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms' (Romualdi *et al.* 2002), 'Informativeness of genetic markers for inference of ancestry' (Rosenberg *et al.* 2003). There are also computer programs available, for instance Röhl *et al.* (2001) describe MTRADIUS while Pritchard *et al.* (2000) explain STRUCTURE.

Of the mentioned papers, Connor & Stoneking (1994) and Röhl *et al.* (2001) are the only ones to use data similar to ours. As far as methods go, the approach we propose combines tools from multivariate statistics in a way that differs from those in the above-mentioned publications.

We assume that distinct population groups are defined. For each of these groups a database is available. A new sample of unknown origin is presented and we are asked to assign the sample to one of the defined groups. More generally, a list of probabilities indicating how likely the different groups are could be requested and delivered. The classification should be based on all available data. This includes a mtDNA profile and perhaps also information on the population structure of the area where the sample was found. As opposed to Pritchard *et al.* (2001), we do not aim to cluster populations or estimate how many different groups there are.

© University College London 2004

The indicated papers use various genetic data (STR, ALU repetitions, mtDNA, Y-chromosomes etc.). Different data may serve different purposes. If it is desired to classify to a few, large continental regions, slowly varying DNA, i.e., DNA which very rarely mutates, may be appropriate. If classification is within a continent, or perhaps even within a smaller region, DNA with finer resolution is needed. Generally, various sorts of data may be combined using Bayes' theorem, as explained later. Our approach is to construct methods appropriate for high dimensional data. We try to be no more restrictive than necessary as far as data assumptions are concerned; the approach is not restricted to mtDNA data. Essentially we propose a two-step approach. The first reduces the dimension of the data and transforms the data. The second step involves standard discriminant analysis; the transformation of the first part makes the assumptions required for step 2 reasonable. Throughout, we emphasize the importance of using separate data for model fitting and testing. Error rates along with associated estimates of uncertainty are presented, based on cross-validation procedures explained in more detail in the appendix. Some conclusions, alternative methods, and comments are provided in the last section.

Material and Methods

The data template is presented first and followed by a review of the optimal classification rule. Then we briefly explain Principal Component Analysis (PCA), Quadratic Discriminant Analysis (QDA) and Cross Validation (CV). The methods are implemented in S-PLUS 6.1 for Windows. In particular, the library MASS is used (Venables & Ripley, 1997). Some additional details of the statistical methods and their implementation are supplied in Appendix I.

The Data Template

There are n individuals, each with data from k sites. The data is organized as a matrix. Each line corresponds to an individual. The first column indicates the class the person belongs to. An entry of the following column is 0 if this value coincides with the rCRS (revised Cambridge Reference Sequence; Andrews *et al.* 1999) of this site and 1 otherwise. The remaining columns are defined similarly, i.e., there is one column for each different site in the database. An example of a small data set is provided in the appendix.

We will assume that the data have been obtained by random sampling of individuals from each population group, or something reasonably close to that. This may be a problem for mtDNA data if, say, data originates from paternity cases including a mother and her child. In this case, several copies of haplotypes may be included and it may or may not be possible to preprocess the database to make it useful and representative. A great number of errors or inaccuracies in mtDNA databases have been reported (c.f. Röhl *et al.* 2001), indicating that care is needed to establish a good database. Bandelt *et al*. (2002) present methods to detect suspicious data sets.

The Optimal Solution

The problem at hand can be described as a classification problem with the geographical groups corresponding to the classes. The solution may rely on many, diverse, sources of information and these need to be reconciled. Let $g = 1, \ldots, G$ denote the geographical groups and P(g) the prior probability that the donator belongs to group g. If D_1 denotes data and $P(D_1|g)$ the likelihood then the posterior probability $P(g|D_1)$ is related to the prior probability by means of Bayes' theorem

$$P(g\,|\,D_1) = \frac{P(D_1|g)\,P(g)}{\sum_{g'} P(D_1|g')\,P(g')}. \qquad (1)$$

For the special case where nothing is known of the geographic origin prior to the DNA sample being obtained, specifying $P(g)$ to be equal for all groups may be reasonable. In this case the above expression simplifies to

$$P(g\,|\,D_1) = \frac{P(D_1|g)}{\sum_{g'} P(D_1|g')}. \qquad (2)$$

The optimal rule (with respect to minimization of the classification error rate) is to classify to the group with largest posterior probability. We will refer to this as the Bayes rule. With $P(g)$ equal for all groups this corresponds to classifying to the group maximizing the likelihood.

Equation (1) may be used repeatedly if there are various sources of information. If a new piece of data, D_2,

 © University College London 2004

is available, the probabilities are updated as

$$P(g\,|\,D_1,\,D_2) = \frac{P(D_2|\,D_1, g)\,P(g\,|\,D_1)}{\sum_{g'} P(D_2|\,D_1, g')\,P(g'|\,D_1)} = \frac{P(D_2|g)\,P(g\,|\,D_1)}{\sum_{g'} P(D_2|g')\,P(g'|\,D_1)} \quad (3)$$

where the last equality holds if D_1 and D_2 are independent given g.

Specification of $P(D_1|g)$

The main difficulty in using the Bayes rule is the specification of the likelihood $P(D_1|g)$. This includes both specification of the form *and* estimation of the parameters involved. Assuming now $D_1 = x_i$, we need to consider the multivariate distribution for a specific individual i from class g,

$$P(X_i = x_i|\,\text{class} = g) = P(X_{i1} = x_{i1}, \ldots, X_{ik} = x_{ik}|\text{class} = g).$$

The problem is that data from different sites are dependent in a complex way and so there is no simple model. With a limited amount of data to estimate the unknown parameters, simplified models should be preferred.

Standard models from probability, such as Markov chains, do not describe the structure adequately. Considering locus i for an individual from class g, the likelihood is $p_{gi}^{\,x_{gi}}(1 - p_{gi})^{1-x_{gi}}$ where p_{gi} is the probability that site i deviates from the rCRS for a sample from group g.

Unrealistically assuming independence between sites, these likelihoods may be multiplied. Taking logarithms, we arrive at

$$l(x_i|g) = \log\left(\prod_{i=1}^{k} P(X_i = x_i|\text{class} = g)\right) = \sum_{i=1}^{k} (x_{gi}\log(p_{gi}) + (1 - x_{gi})\log(1 - p_{gi})) \quad (4)$$

Then we can apply Bayes rule and classify to the group maximizing the likelihood.

In the case of only two groups, (4) can be seen as a special case of logistic regression. The use of logistic regression in the two-class situation has been demonstrated by Connor & Stoneking (1994).

PCA-QDA Approach

This method relies on two steps. The first, PCA, reduces the dimension of the data set. The second, QDA, performs the classification. In the likelihood formulation this corresponds to defining D_1 to be a linear transformation of x_i to a lower-dimensional space and assuming the transformed variable follows a multivariate normal distribution. Because some data in this case is thrown away, it will not be an optimal Bayes rule, but since estimation of parameters will be easier in the reduced space, good performance can still be achieved.

We first motivate and briefly describe the present PCA implementation. The need for some sort of dimension reduction arises when the number of variables, sites in this case, is large compared to the number of individuals. In our case, it would be possible to increase the number of individuals in our main example beyond our number (2017). However, the number of sites will still be relatively large. Moreover, problems of collinearity will prevail since different groups of sites may contain essentially the same information.

Principal component analysis in this setting requires sufficient variability in the data related to differences between groups to be useful. Linear transformations of the data with maximal variability should therefore be useful quantities for discriminating. The first principal component is defined as the linear combination of the original variables $x_1, \ldots, x_k$ that accounts for the maximal variance of the x-variables among all such combinations. The second principal component is defined similarly and is required to be uncorrelated with the first. Remaining principal components are defined along the same lines. See Appendix I for more details and implementation. The PCA step converts discrete data into continuous data. From the above definition, it is reasonable that the principal components might be close to normally distributed. The central limit theorem does not apply directly to show this since the weights in the linear combinations depend on data. Still, the weights should be reasonably stable, justifying normal distributions. Figure 1 shows that the first two principal components distinguish reasonably well between SE Africa, and Germany and Iceland.

© University College London 2004

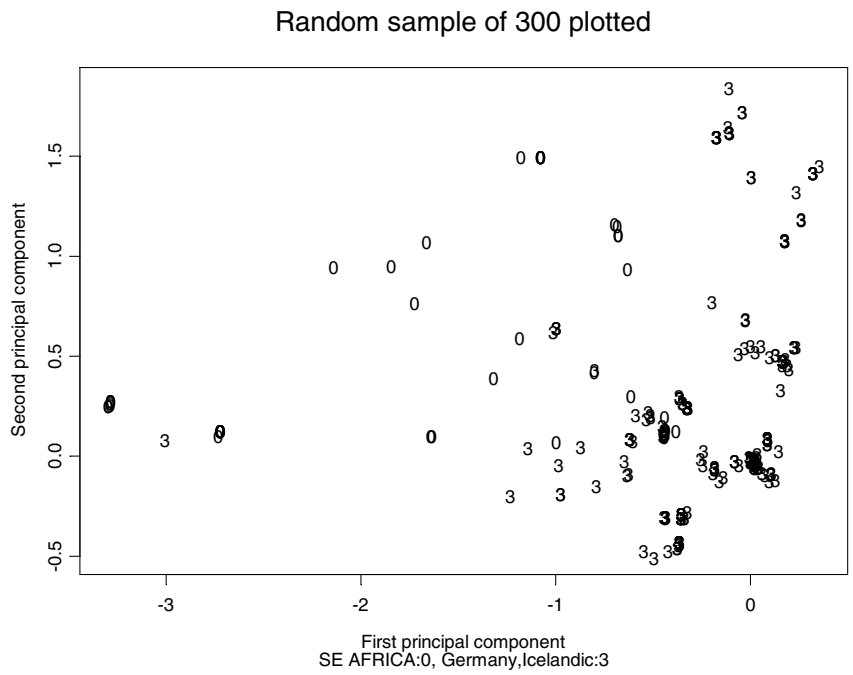

Figure 1 The plot shows that the first two principal components distinguish SE African samples reasonably well from the others.

Once the principal components have been determined, we use discriminant analysis to perform the classification. Let μ_g and Σ_g denote the mean and covariance of class g in the space spanned by the chosen principal components. For linear discriminant analysis, the covariance matrices are supposed to be identical across groups, whereas this is not the case for quadratic discriminant analysis. An observation y in the PCA-space is classified by QDA to the class g minimizing the Mahalanobis distance

$$(y - \mu_g)^T \Sigma_g^{-1} (y - \mu_g) \tag{5}$$

where ^{T} denotes transposed. This approach corresponds to the Bayes rule (based on y), provided the data from the groups are normal with the parameters μ_g and Σ_g.

Cross Validation Procedure to Determine the Number of Principal Components

A separate problem is to decide on the number of principal components to retain. A cross-validation scheme for this purpose is explained next; Breiman & Spector (1992) explore similar approaches in more general frameworks.

1. First a part of the data is selected randomly and in equal proportions from each group. This part of the data is set aside for test purposes and called test data whereas the remaining part is the training data. We have pragmatically chosen 90% of the data for training.
2. The training data is divided into T parts indexed by $t = 1, \ldots, T$. In our implementation, $T = 10$.
3. Let max.pri denote the maximum number of principal components, max.pri must be smaller than the number of sites.
4. For $i = 1, \ldots,$ max.pri {
   For $t = 1, \ldots, T$ {
   Make a QDA classification rule based on training data apart from part t and based on the first i principal components.
   Perform classification on part t and save the error rate as rate[t]
   }
   Let err[i] = average of rate[1], . . . ,rate[T].
   }
5. Choose the number of principal components, n.pri, to minimize err[i].

 © University College London 2004

Table 1 A listing of the most discriminating sites

site	SE Africa	GeIc	diff
223	303	129	0.91
278	156	43	0.48
189	182	252	0.45
187	111	6	0.36
230	100	4	0.32
311	145	289	0.30
320	93	29	0.29
148	88	8	0.28
309	85	2	0.28
188	88	30	0.27
172	92	80	0.25
294	109	223	0.22
126	4	372	−0.20
129	69	146	0.14
290	43	7	0.14
286	34	5	0.11
304	1	180	−0.10

The leftmost columns identify the site by its number (minus 16000). Next follows the number of differences compared to rCRS for the SE African sample (307 all together) and the combined Germanand Icelandic sample, abbreviated as GeIc (1710 samples in total). The next column is 0.91 for site 16223 and is calculated by forming the difference between 303/307 = 0.987 and 129/1710 = 0.075.

6. Make a QDA classification rule based on n.pri principal components and the entire training data.
7. Evaluate the classification rule on the separate test data.

Steps 1 to 7 can be repeated to estimate uncertainty in the estimate for the number of principal components and error rate. The main idea of this procedure is to consider the estimation of the number of principal components as part of constructing a classification rule. By using *CV inside the training set* a reliable estimate of the error rate can be obtained. Sometimes, we call the procedure a *double CV*.

Results

Example. Simulated Data

The first example is constructed in order to evaluate how well the PCA-QDA method performs in a situation where the full multivariate distribution of x_i is known and defined by Equation (4). We have simulated data of a size and structure similar to the data of our main example, the main difference being that sites are assumed independent for the simulations. Table 2 shows the results and the legend provides more details. The main conclusion is that the suggested PCA-QDA approach is close to the optimal in this case.

In addition, the error rate is seen to increase from 0 to around 20% as the populations become closer.

Example. Real Data

Our mtDNA consists of 1314 Germans (Pfeiffer *et al.* 2001), 396 individuals of Icelandic origin (Helgason *et al.* 2000) and 307 of SE African background (Salas *et al.* 2002). Only the 151 sites of the HVS-I region displaying variability are used. A more complete discussion of the data centering on Table 1 follows in the Discussion.

Table 3 indicates that we can distinguish reasonably well between Germany and Iceland on one side and SE Africa on the other. This is as anticipated. Considering the raw data in greater detail verifies that we should not expect to be able to separate German and Icelandic samples accurately. For instance, 602 of the 1314 German samples are also observed in the Icelandic data set. Recall that not all samples have different sequences; in fact there are only 634 unique sequences in the complete data set of 2017 samples.

Next we reduce the problem to distinguish SE African samples from the others. Figure 2 is based on the cross validation procedure and shows the optimal choice to be 48 principal components. The error rate increases after the minima reflecting overfitting; too many parameters are fitted. Only 4 out of 202 observations are misclassified in the subsequent run with this choice (Table 4). To investigate further the uncertainty of *the method*, the number of principal components was fixed to 48 and the cross validation procedure was repeated 100 times with test sets chosen randomly each time. Figure 3 shows that error rates range from 1 to 5%. Figure 4 displays the amount of variance explained as a function of the number of principal components.

As mentioned previously, logistic regression may be used when there are only two groups (Connor & Stoneking, 1994). We have constructed a procedure based on logistic regression which resembles the PCA-QDA approach. First, 90% of the data was randomly chosen. A model was fitted based on this data by

© University College London 2004

Table 2 Test results for simulated data

	Error rate (%)			
Method	(0.05,0.20,0.35)	(0.05,0.15,0.30)	(0.05,0.10,0.15)	(0.01,0.05,0.09)
PDA-QDA	0 (2)	1.4 (2)	21.9 (4)	17.1 (2)
likelihood	0	1.4	20.0	17.1

Four data sets of size and structure similar to the example data have been simulated. There are 150 independent sites. The test data consists of 700 samples from each of the three populations. The heading (0.05,0.10,0.15) indicates that the probability for a 1 is 0.05,0.10 and 0.15 respectively for the three populations and similarly for the other columns. The error rate for the PDA-QDA is 21.9% for the mentioned case. The PDA-QDA classifications are performed with the optimal number of principal components and these are given in parentheses. In this case the likelihood approach is optimal and the PDA-QDA approach is very close.

Table 3 PDA-QDA approach for three groups

	Sample from		
Assigned to	SE Africa	Germany	Iceland
SE Africa	31	11	6
Germany	0	94	24
Icelandic	0	26	10

The result of the PDA-QDA approach based on the optimal choice, 11 principal components, is shown. The test data consists of the 10% of the data (202 samples) not used for model fitting. For instance all SE Africans are correctly classified whereas the results for Germany and Iceland are much worse. The overall error rate is 33.2%.

running stepwise regression and an error rate was estimated using the test data. This procedure led to twice the error rates of Table 4.

Example. Including Prior Information

So far all classification has been carried out using a flat prior. In other words, *a priori* different classes are assumed to be equally likely. Likelihood based classification ignoring prior information would lead to the same results. Prior information is taken care of by Bayes' theorem as explained previously. The practical implementation is shown in Appendix I.

Discussion

Table 1 summarizes some characteristics of the data. Briefly, within the worldwide mtDNA phylogeny, haplogroups L0, L1, and L2 constitute macrohaplogroup L (the most ancient groups) and encompass the majority of the sub-Saharan mtDNAs (c.f. Chen *et al*. 2000; Watson *et al*. 1997; Salas *et al*. 2002). From the African L3 paragroup radiates the Eurasian macrohaplogroups M and N. From N derive the European haplogroups H, I, J, N1b, T, U, V, W, and X (more than 95% of the European lineages; Richards *et al*. 2000). All these derived N lineages lack the transition w.r.t. the rCRS at position 16223. This site tops the list of the sites making the most striking difference in the statistical analysis (Table 1). But many other sites, with special relevance for the PCA, are also important diagnostic positions for haplogroups characterizing the populations used as examples in the present work. For instance: a) 16278 separates L2'3 from L3* (note that we are only considering HVS-I information), b) non - L2'3 haplotypes bear 16187, 16189 and 16311; c) 16230 characterizes L0, while 16148 and 16320 lead to the rather frequent southeastern L0-African Bantu derived lineages L0a, d) 16309 characterized L2a1, also an important lineage well represented in SE African populations, e) site 16126 characterizes the sub-Saharan African haplogroup L1b, but it is also an important keyposition for the European lineages T (base motif: 16126 16294) and J (16069 16126), etc. In general, there is a good agreement between the statistical approach developed here and the mtDNA phylogeny. In particular, the four misclassified samples (data not shown) of Table 3 are of ambiguous phylogenetical origin.

There are various other reasonable approaches to the classification problem we have discussed. PLS (Partial Least Squares) is a method similar to PCA, reported to work particularly well when there is high correlation among the variables. Other suggestions for discrimination include tree-based methods like CART (Venables & Ripley, 1997) and neural nets (Romualdi *et al*. 2002).

 © University College London 2004

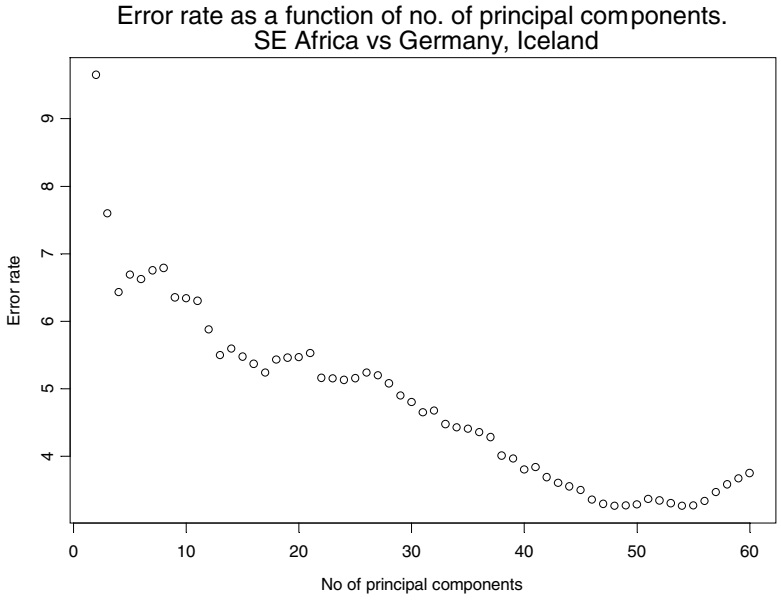

Figure 2 The error rate based on the training data is shown as a function of the number of principal components. The figure is based on a double cross-validation; the described cross validation procedure is repeated 100 times and the figure shows the average. A minimum error rate is obtained for 48 principal components.

Table 4 The quality of the classification demonstrated using the test data

	Sample from	
Assigned to	SE Africa	Germany/Iceland
SE Africa	31	4
Germany/Iceland	0	167

There are 4 misclassifications in the separate test set based on the PDA-QDA approach with 48 principal components (the optimal choice).

We found our approach to outperform logistic regression. One principal difference in philosophy between the two approaches is that while logistic regression only uses a subset of sites or variables, the PCA-QDA approach employs all sites with different weights. In cases where there is information in all sites, it seems intuitively reasonable to use all the data.

Throughout, we consider 0-1 data. There could be more information in the data than we utilize. There could also be insertions and deletions, as well as information on the type of substitution. This type of data can be handled, for instance by introducing *dummy* variables.

Forensic data may provide mother-father-child triplets. In this case the mother and her child will have the same data, disregarding mutations. For this sampling it may be wise to remove duplicate data within each group. However, adjusting for non-random sampling is never straightforward.

There are some pragmatic choices in the cross validation algorithm. One could imagine formal, simulation based methods to address these issues. Our method may underestimate the required number of principal components since the estimation is performed on a reduced data set. On the other hand, it may be slightly surprising that such a large number of principal components is required. This may reflect subtle dependencies between sites of the data. Using 48 principal components without validation on external data may be highly dubious and once more we emphasize the importance of separating training and testing. CV is not an essential part of the method, it could be replaced by other ways of estimating the number of principal componts and by alternative test procedures.

There is an important distinction between our application and most other cases where PCA analysis is used: the uncertainty in the data can be ignored. For instance, if the length of a baby is measured several times within a short period, different values will

© University College London 2004

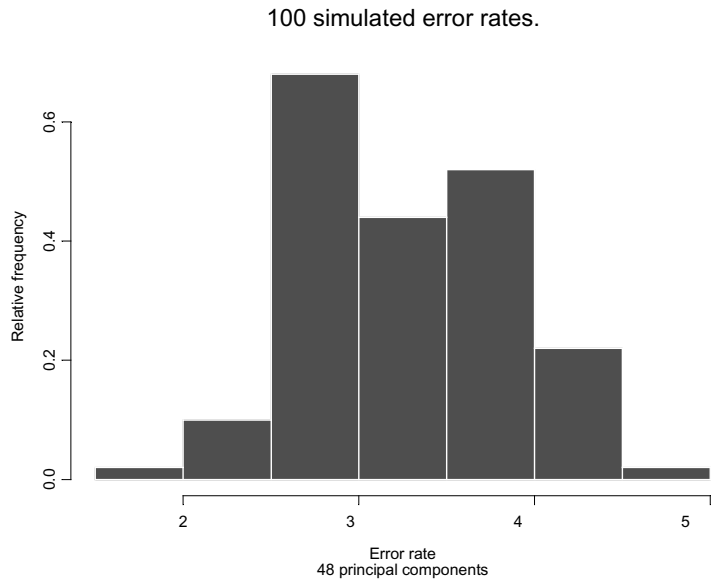

Figure 3 The histogram indicates the uncertainty in the estimate of the error rate. The numbers of principal components have been fixed to the optimal 48, and 100 double cross validations have been performed.

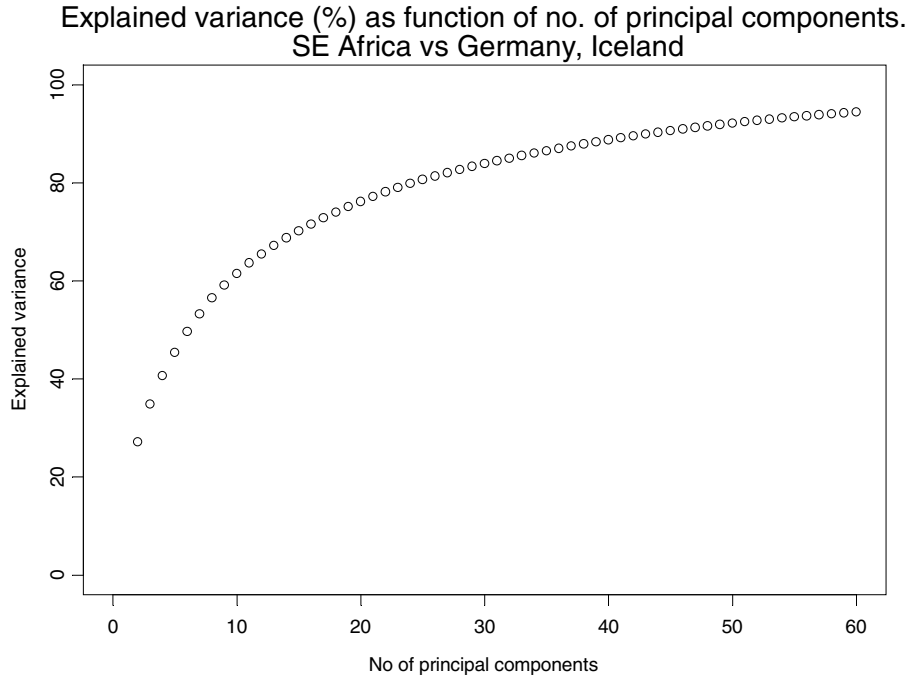

Figure 4 The amount of variance explained as a function of the number of principal components. The chosen number, 48, is seen to explain slightly less than 90% of the variance.

be obtained. We assume that the mtDNA sequence profile will be the same if repeated analyses are performed on different samples. In many PCA applications, scaling is a problem. For our examples, this is not an issue.

In practical cases there may be missing data; information on exactly the same sites as we have used may not be available. A practical solution to this problem is to develop the procedure based on the sites available in the sample to be classified. If this is not

 © University College London 2004

possible, there are statistical approaches to this missing value problem. We omit details here, but note that stronger distributional assumptions may be required in this case. It may be of interest to extend the procedure to include a class 'doubt'. Then, it is conceivable that we may conclude that a new sample cannot reasonably be allocated to any of the known populations.

Finally, it must be kept in mind that the most likely geographical origin of an mtDNA profile is not equivalent to the most likely geographical origin of an individual. Neither should we expect to find a complete correlation between the profile and a specific (African, western European, etc) phenotype. For instance, the mtDNA haplogroup composition of the self-defined 'white' Brazilian population in Alves–Silva *et al.* (2000) is: 33% Native American/Asian, 28% African, and 39% European. However, some correlation exists, and it is this correlation that can make this method useful for police investigation in criminal casework and also other problems of interest.

Acknowledgements

We would like to thank for helpful comments from two anonymous reviewers. The Leverhulme Trust supported the first author's contribution to this paper. Financial support was provided by the Ministerio de Sanidad y Consumo (Fondo de Investigación Sanitaria; Instituto de Salud Carlos III, PI030893; SCO/3425/2002). AS is supported by the Isidro Parga Pondal program (Xunta de Galicia).

References

Alves-Silva, J., Silva Santos, M., Guimaraes, P.E., Ferreira, A.C., Bandelt H.-J., Pena, S.D. & Prado, V.F. (2000) The ancestry of Brazilian mtDNA lineages. *Am J Hum Genet* **Aug 67**, 444–61. *Erratum in Am J Hum Genet* (2000) **67**, 775.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M. & Howell, N. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**, 147.

Bandelt, H.-J., Quintana-Murci, L., Salas, A. & Macaulay, V. (2002) The fingerprint of phantom mutations in mitochondrial DNA data. *Am J Hum Genet* **71**, 1150–60.

Breiman, L. & Spector, P. (1992) Submodel selection and evaluation in regression. The x-random case. *International Statistical Review* **60**, 291–319.

Brenner, C. (1997) Probable Race of a Stain Donor. Proceedings from the Seventh International Symposium on Human Identification 1996, Promega Corp, 48–52.

Chen, Y.S., Olckers, A., Schurr, T.G., Kogelnik, A.M., Huoponen, K. & Wallace, D.C. (2000) mtDNA variation in the South African Kung and Khwe and their genetic relationships to other African populations. *Am J Hum Genet* **66**, 1362–1383.

Connor, A. & Stoneking, M. (1994) Assessing ethnicity from human mitochondrial DNA types determined by hybridization with sequence-specific oligonucleotides. *J Forensic Sci S* **39**, 1360–1371.

Helgason, A., Sigurðardóttir, S., Gulcher, J.R., Ward, R. & Stefansson, K. (2000) mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am J Hum Genet* **66**, 999–1016.

Lowe, A.L., Urquhart, A., Foreman, L.A. & Evett, I.W. (2001) Inferring ethnic origin by means of an STR profile. *Forensic Sci Int* **119**, 17–22.

Pfeiffer, H., Forster, P., Ortmann, C. & Brinkmann, B. (2001) The results of an mtDNA study of 1,200 inhabitants of a German village in comparison to other Caucasian databases and its relevance for forensic casework. *Int J Legal Med* **114**, 169–72.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–59.

Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellitto, D., Cruciani, F., Kivisild, T., Villems, R., Thomas, M., Rychkov, S., Rychkov, O., Rychkov, Y., Golge, M., Dimitrov, D., Hill, E., Bradley, D., Romano, V., Cali, F., Vona, G., Demaine, A., Papiha, S., Triantaphyllidis, C., Stefanescu, G., Hatina, J., Belledi, M., Di Rienzo, A., Novelletto, A., Oppenheim, A., Nørby, S., Al–Zaheri, N., Santachiara–Benerecetti, S., Scozzari, R., Torroni, A. & Bandelt, H.-J. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* **67**, 1251–1276.

Röhl, A., Brinkmann, B., Forster, L. & Forster, P. (2001) An annotated mtDNA database. *Int J Legal Med* **115**, 29–39.

Romualdi, C., Balding, D., Nasidze, I.S., Risch, G., Robichaux, M., Sherry, S.T., Stoneking, M., Batzer, M.A. & Barbujani, G. (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res* **12**, 602–12.

Rosenberg, N.A., Li, L.M., Ward, R. & Pritchard, J.K. (2003) Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* **73**, 1402–22.

Salas, A., Richards, M., De la Fé, T., Lareu, M.V., Sobrino. B., Sánchez-Diz, P., Macaulay, V. & Carracedo, A. (2002) The making of the African mtDNA landscape. *Am J Hum Genet* **5**, 1082–111.

© University College London 2004

Venables, W.N. & Ripley, B.D. (1997) Modern applied statistics with S-PLUS. Springer, NY, 2. Ed.

Watson, E., Forster, P., Richards, M. & Bandelt, H.-J. (1997) Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet* **61**, 691–704.

Appendix I: PCA-QDA Approach

This section explains the combination of Principal Component Analysis (PCA) and Quadratic Discriminant Analysis (QDA) used in the paper. S-PLUS commands are also included, but other programs like the shareware program R can be used. Consider the following superficially small data set:

pop	x_1	x_2	x_3
1	0	0	0
1	0	1	0
1	1	0	0
2	1	1	1
2	1	0	1
2	0	1	1

There are six samples, each from three sites. The three first samples are known to be from population 1, the others from population 2. PCA is a way of transforming x_1, x_2, and x_3 to new variables y_1, y_2, and y_3 with the following properties:

1. Each y is a linear combination of x_1, x_2, and x_3, i.e.,

   $$y_1 = a_1 * x_1 + a_2 * x_2 + a_3 * x_3$$

   and similarly for y_2 and y_3.
2. The sum of the squares of the a-coefficients is unity.
3. y_1 is the linear combination of x_1, x_2 and x_3 with greatest variance. Of all possible linear combinations uncorrelated with y_1, y_2 has the greatest variance. Similarly, y_3 has the largest variance of all linear combinations uncorrelated with y_1 and y_2.

The S-PLUS command performing PCA is prcomp(data). Most textbooks on multivariate statistics will explain the algebra underlying the implementation and these details are omitted here. The rotation matrix, rotation, is part of the output:

0.8165	0.0000	−0.5774
−0.4082	0.7071	−0.5774
0.4082	0.7071	0.5774

In the above notation $a_1 = 0.8165$, $a_2 = -0.4082$ and $a_3 = 0.4082$.

Observe that requirement 2 above holds since

$$0.8165^2 + (-0.4082)^2 + (0.4082)^2 = 1.000$$

and similarly for the two other columns. Requirement 3 corresponds to demanding the columns of the above matrix to be orthogonal. This is seen to be true for the two first principal components since

$$0.8165 * 0 + (-0.4082) * 0.7071 + 0.4082 * 0.7071 = 0$$

and similarly for the other pairs of columns.

The main idea of PCA is that most of the information is retained using some of the components. In the paper, we outline a cross validation procedure to decide on how many components to use. Assume that we decide on two in this artificial example. Then the data is transformed to the space spanned by the two first principal components. This is done by matrix multiplication in S-PLUS producing the transformed data

0.0000	0.0000
−0.4082	0.7071
0.8165	0.0000
0.8165	1.4142
1.2247	0.7071
0.0000	1.4142

For instance, item 2,1 of this matrix is

$$0 * 0.8165 + 1 * (-0.4082) + 0 * 0.4082 = -0.4082.$$

Next, we turn to the QDA section. The paper describes this briefly. Again, textbooks provide more details. The implementation in S-PLUS 6.1 is

```
z1 <- qda(transformed.data,c(1, 1, 1, 2, 2, 2),
prior = c(0.5,0.5))
```

The instruction *z*1<-implies that the results are stored to *z*1. There are three parameters, the transformed.data, a vector (1, 1, 1, 2, 2, 2) indicating the known populations these six samples come from, and finally the prior

 © University College London 2004

probability that a sample is from population 1 or 2, 0.5 for each in this case.

So far, the PCA part and the QDA have been explained. Consider a test sample, say the observation (1, 0, 0). We would like to calculate the *posterior* probability that this sample is from population 1 (for now we ignore that we expect it to be from 1). First we need to transform the observation to the two dimensional PCA space by matrix multiplication resulting in obs = (0.816, 0). (Technically, this is done by multiplying (1, 0, 0) with the matrix consisting of the first two columns of the rotation matrix.) The output from the call

```
predict(z1,obs, prior =c(0.5,0.5))
```

contains the posterior probability that this is an observation from population 1, and is 0.9997. If, for some reason, it is *a priori* unlikely to be a population 1 observation, say there is a prior of 0.01, then the calls become

```
z1< − qda(transformed.data,c(1, 1, 1, 2, 2, 2),
  prior=c(0.01,0.99))
predict(z1, obs, prior =c(0.01,0.99))
```

and the posterior probability is slightly reduced, 0.9679.

This concludes the small tutorial. Obviously, the reader should consult other sources for a more complete story. We emphasize that the approach we have outlined is intended for cases where there may be many populations and a large number of correlated sites.

Received: 10 December 2003
Accepted: 8 March 2004

© University College London 2004