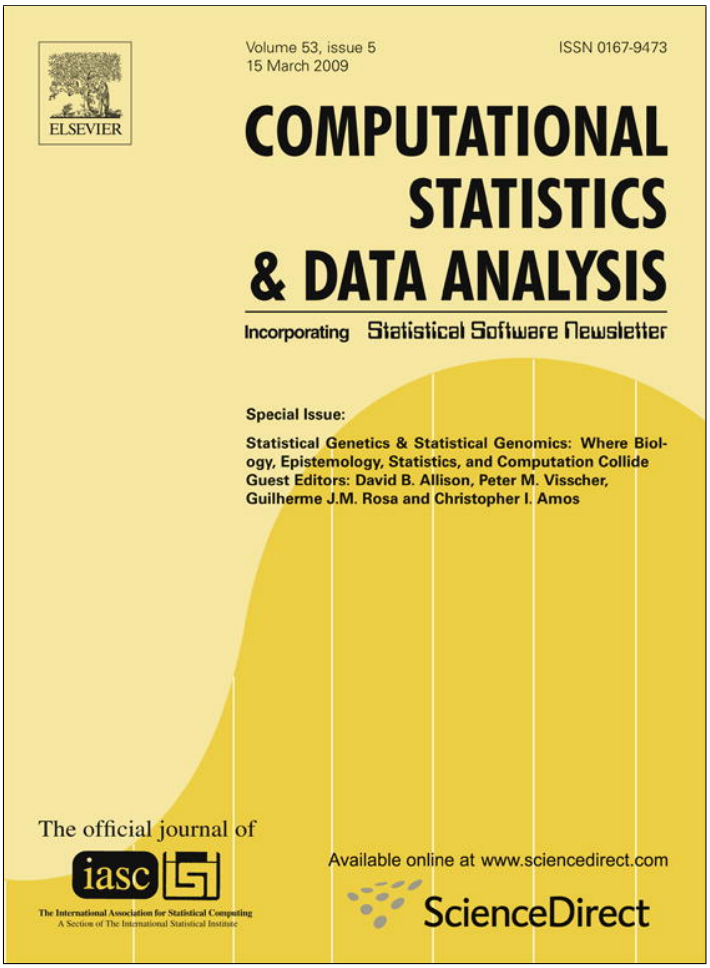
Bingham, Ella; Mannila, Heikki

Complexity control in a mixture model by the Hardy-Weinberg equilibrium

Published in:
Computational Statistics and Data Analysis

Published: 01/01/2009

Please cite the original version:
Bingham, E., & Mannila, H. (2009). Complexity control in a mixture model by the Hardy-Weinberg equilibrium. *Computational Statistics and Data Analysis*, *53*(5), 1711-1719.



Computational Statistics and Data Analysis 53 (2009) 1711–1719

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Complexity control in a mixture model by the Hardy–Weinberg equilibrium

Ella Bingham [a,*], Heikki Mannila [a,b]

[a] Helsinki Institute for Information Technology, University of Helsinki, P.O. Box 68, FIN-00014 Helsinki, Finland
[b] Helsinki Institute for Information Technology, Helsinki University of Technology, P.O. Box 5400, FIN-02015 HUT, Finland

ARTICLE INFO

Article history:
Available online 22 July 2008

ABSTRACT

A method of complexity control in multinomial mixture modeling of multiple-marker genotype data, imposing the Hardy–Weinberg equilibrium (HWE) between the genotype values, is studied. This is a very natural restriction, and known to hold at population level under modest assumptions. The hypothesis under study is that imposing this restriction will prevent overfitting and lead to a better model. This is shown to indeed be case. Experimental results on chromosomes 1 and 17 of the HapMap data demonstrate that the restricted model generalizes better to unseen data, and also finds clusters that correspond better to the ethnic groups of the HapMap, when compared with a model without the HWE restriction.

1. Introduction

In this paper we discuss the problem of clustering genotype data that consists of multiple markers. We adopt the statistical model-based approach, assuming that the data are generated by a multinomial mixture model, and further assuming that the parameters of such a model can be identified by the maximization of the data likelihood.

We study the effect of requiring that the Hardy–Weinberg equilibrium (HWE) (Hardy, 1908; Weinberg, 1908) holds for the genotype values. Our hypothesis is that imposing the HWE restriction acts as a means to complexity control, and helps to reduce overfitting. Often in studies on genotype data, the number of observations (patients) is quite small compared with the number of variables (markers) in the data, making model estimation prone to overfitting.

Genotype data of a diploid organism can be presented as unordered pairs of the maternal and paternal haplotypes: $\{A, A\}$, $\{a, a\}$ and $\{A, a\}$. The data are thus categorical and we do not assume any ordering between the values. Multinomial mixtures is a well-known technique aimed at modeling such categorical data, similarly to Bernoulli mixtures for 0/1 data (Blekas and Likas, 2004; Redner and Walker, 1984; Meilă and Heckerman, 2001; Song et al., 2007; Rufo et al., 2007; Patist, 2006).

Hardy–Weinberg equilibrium (HWE) is one of the key concepts in genetics. Let us denote the frequency of allele A by α; then the frequency of allele a is $1-\alpha$. The Hardy–Weinberg equilibrium says that the frequency of $\{A, A\}$ is α^2, the frequency of $\{a, a\}$ is $(1-\alpha)^2$ and the frequency of $\{A, a\}$ is $2\alpha(1-\alpha)$. Given a marker with any frequency distribution of the genotypes $\{A, A\}$, $\{A, a\}$, and $\{a, a\}$ in a large population, the simple assumptions of random mating and no selective effects lead to HWE in one generation. The genotype frequencies will remain unchanged over successive generations.

In mixture modeling of genotype data of multiple markers, one can either enforce the HWE for the model parameters or disregard it. Here we study the effect of this choice from the point of view of identifying populations. Using HapMap data, http://www.hapmap.org, we show that enforcing the HWE leads to a multinomial mixture model that finds the ethnic groups in the data more easily than a multinomial mixture model that disregards HWE. In addition to identifying populations, we will also show that the HWE-enforced model, hereafter HWEmultinomial, fits better to unseen data. The

* Corresponding author. Tel.: +358 9 191 51377; fax: +358 9 191 51120.
E-mail addresses: ella@iki.fi (E. Bingham), heikki.mannila@tkk.fi (H. Mannila).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.07.023

original multinomial model, with no HWE imposed, has two unknown parameters for each marker: the probabilities of two genotypes (the probability of the third genotype value is not a free parameter, as we know that the probabilities must sum to 1). In contrast, the HWEmultinomial model has only one free parameter per marker. As a result, the HWEmultinomial has less freedom in fitting to observed data, and does not overfit as easily to the possibly noisy details of the data. The HWE restriction controls the flexibility of the HWEmultinomial model, leading into better generalization capability and less overfitting.

Linkage disequilibrium, dependence between nearby markers, is another biological constraint in genotype data. We concentrate our study on the effect of HWE, independently of linkage disequilibrium. To accomplish this, we select markers randomly from a chromosome; as a result the markers are on average quite far from each other, and the effect of linkage disquilibrium is small.

Multinomial models with HWE have been presented before in a single-attribute case, as opposed to our multivariate approach. Glickman and Kao (2005) study the role of the Apo-E gene in the onset of cardiovascular diseases. They report that imposing the HWE restriction results in poorer fit, which is not surprising as the model has fewer free parameters. HWE of HapMap data is studied in Wigginton et al. (2005) who give exact statistical tests, Barrett et al. (2005) who present an analysis software, McCarroll et al. (2006) who discover common deletion polymorphisms, Fung et al. (2006) who study the genotyping of Parkinson disease and Weinberg and Morris (2003) who comment on testing for HWE and possible reasons for HWE violations. Monitoring deviations from HWE can be used as tools for quality control (Hosking et al., 2004), and for identifying interesting genomic locations; see Kocsis et al. (2004) for a study on the use and underuse of HWE.

Methods for identifying population structure in multilocus genotype data have been presented by several authors. Rannala and Mountain (1997) and Cornuet et al. (1999) discuss assigning individuals of unknown origin into potential (known) source populations. Pritchard et al. (2000) and further Falush et al. (2003, 2007) give a Bayesian formulation for finding the populations, and assigning individuals to them. The model is implemented as the *structure* program that uses MCMC simulations to estimate the model parameters. Their model accounts for the presence of Hardy–Weinberg disequilibrium by grouping individuals into populations within which the HWE more or less holds. We will discuss the *structure* program in more detail in the experimental section, in which we show comparisons between it and the method presented in this paper. Somewhat similar models are presented by Dawson and Belkhir (2001) and by Anderson and Thompson (2002). Corander et al. (2003) give a Bayesian method for estimating hidden population substructure that uses geographical sampling information and again assumes HWE and linkage equilibrium within populations. Excoffier et al. (2005) use approximate Bayesian computation to estimate a model that is capable of explicitly handling mutations; their model was defined previously by Bertorelle and Excoffier (1998) and (in a maximum likelihood formulation) by Wang (2003). Wu et al. (2006) give a maximum likelihood approach, partially based on the method presented by Tang et al. (2005), and an efficient implementation.

This paper is organized as follows. In Section 2 we present multinomial models, both the original and the HWE-restricted version. In Section 3 we show experimental results on both clustering and generalization, using the two multinomial models and the *structure* program. Section 4 concludes the paper with a brief discussion.

2. Models

We assume multivariate genotype data, and denote by $t = 1, \ldots, T$ the variables (attributes, markers, columns) and by $n = 1, \ldots, N$ the observations (individuals, rows) in the data. For brevity, we will use the symbols 1, 2 and 0 to denote the genotype values $\{A, A\}$, $\{a, a\}$ and $\{A, a\}$, respectively. Note that these are just symbols and we do not assume any ordering between the values 1, 2 and 0. Thus $x_{tn} \in \{0, 1, 2\}$ is the value of the t-th marker of the n-th observation. Each marker is considered separately of others, and hence observing, say, value 1 at one marker gives us no information of value 1 at another marker. (The markers will be chosen far away from each other so that linkage disequilibrium can be neglected.) Finally, we denote by $k = 1, \ldots, K$ the mixture components that are assumed to have generated the data. In practice, the mixture components k will correspond to different populations or groups to which the individuals belong.

Let

$$p_{tk} = \text{prob}(\text{variable } t = 1 | \text{component } k) = p(x_{tn} = 1 | k) \tag{1}$$

for any n. Similarly,

$$r_{tk} = \text{prob}(\text{variable } t = 2 | \text{component } k) = p(x_{tn} = 2 | k) \tag{2}$$

and

$$1 - p_{tk} - r_{tk} = \text{prob}(\text{variable } t = 0 | \text{component } k) = p(x_{tn} = 0 | k). \tag{3}$$

Also let π_k be the prior probability of mixture component k, and $\sum_k \pi_k = 1$. The log likelihood of the multinomial mixture model is then

$$\mathcal{L} = \sum_n \log \sum_k \pi_k \prod_t p_{tk}^{I(x_{tn}=1)} r_{tk}^{I(x_{tn}=2)} (1 - p_{tk} - r_{tk})^{I(x_{tn}=0)} - \beta \left(\sum_k \pi_k - 1 \right) \tag{4}$$

where $I(x_{tn} = \ell)$ is an indicator function, and β is a Lagrange multiplier, using which we ensure that the prior probabilities sum to 1 (it is easily seen that $\beta = N$).

In this report we restrict the multinomial mixture model by additionally requiring that the HWE holds. Using the notation in Section 1, this translates to $p_{tk} = \alpha^2$, $r_{tk} = (1-\alpha)^2$ and $1 - p_{tk} - r_{tk} = 2\alpha(1-\alpha)$. These equations can be combined into

$$r_{tk} = (1 - \sqrt{p_{tk}})^2 \tag{5}$$

which is the restriction we want to apply to our parameters. The log likelihood (4) then becomes

$$\mathcal{L} = \sum_n \log \sum_k \pi_k \prod_t p_{tk}^{I(x_{tn}=1)} (1 - \sqrt{p_{tk}})^{2I(x_{tn}=2)} (2\sqrt{p_{tk}} - 2p_{tk})^{I(x_{tn}=0)} - \beta \left(\sum_k \pi_k - 1 \right). \tag{6}$$

We will use an EM algorithm (Dempster et al., 1977) for estimating the parameters of the model. The update equation for the parameter π_k is similar to the update rule in the original multinomial mixture model (Redner and Walker, 1984; Blekas and Likas, 2004; Patist, 2006):

$$\pi_k = \frac{1}{N} \sum_n s_{kn} \tag{7}$$

where s_{kn} is the posterior probability of component k having created observation n. The update rule of s_{kn} in turn is different from its update rule in the original multinomial mixture model; this time it is computed as

$$s_{kn} = \frac{\pi_k \prod_t p_{tk}^{I(x_{tn}=1)} (1 - \sqrt{p_{tk}})^{2I(x_{tn}=2)} (2\sqrt{p_{tk}} - 2p_{tk})^{I(x_{tn}=0)}}{\sum_k \pi_k \prod_t p_{tk}^{I(x_{tn}=1)} (1 - \sqrt{p_{tk}})^{2I(x_{tn}=2)} (2\sqrt{p_{tk}} - 2p_{tk})^{I(x_{tn}=0)}}. \tag{8}$$

The update equation for the parameter p_{tk} is in turn

$$\sqrt{p_{tk}} = \frac{\sum_n s_{kn} (I(x_{tn} = 1) + \frac{1}{2} I(x_{tn} = 0))}{\sum_n s_{kn}}. \tag{9}$$

For comparison, the update equation in the original multinomial mixture model is

$$p_{tk} = \frac{\sum_n s_{kn} I(x_{tn} = 1)}{\sum_n s_{kn}}. \tag{10}$$

Also for curiosity let us see how the update equation for r_{tk} would look like, derived from (5):

$$\sqrt{r_{tk}} = \frac{\sum_n s_{kn} (I(x_{tn} = 2) + \frac{1}{2} I(x_{tn} = 0))}{\sum_n s_{kn}}. \tag{11}$$

An EM algorithm for genotype data, taking into account the Hardy–Weinberg equilibrium, is now given by (7)–(9).

Our hypothesis is that incorporating the HWE restriction is a convenient and biologically well motivated way to reduce overfitting. The experimental section gives supporting evidence for this.

3. Results

We show clustering and generalization results on both the original multinomial model and the HWE-restricted multinomial model. The experiments are conducted on Matlab, in which the two multinomial models are implemented. In addition, we will show clustering results on the *structure* program (Pritchard et al., 2000), using the implementation given at http://pritch.bsd.uchicago.edu. Let us start by describing the data sets used in the experiments.

3.1. Basic properties of the data

In the HapMap project (The International HapMap Consortium, 2003), http://www.hapmap.org, the complete genotype information of 270 persons was identified and delivered in the public domain. The 270 persons consist of 90 European, 90 African and 90 Asian individuals. More specifically, the "European" samples are Utah residents with ancestry from northern and western Europe; the "African" samples are from the Yoruba people of Ibadan, Nigeria; and the "Asian" samples consist of 45 Japanese individuals from Tokyo and 45 Han Chinese individuals from Beijing. We used the phase II data, release 21.

We took chromosome 1 and screened the markers such that all markers having missing values in any of the individuals were removed. We then selected $T = 20$, 50, 150 or 500 markers randomly along the chromosome. We repeated this 200 times, ending up with 200 data sets of random markers at each T. Similar samples were drawn from chromosome 17.

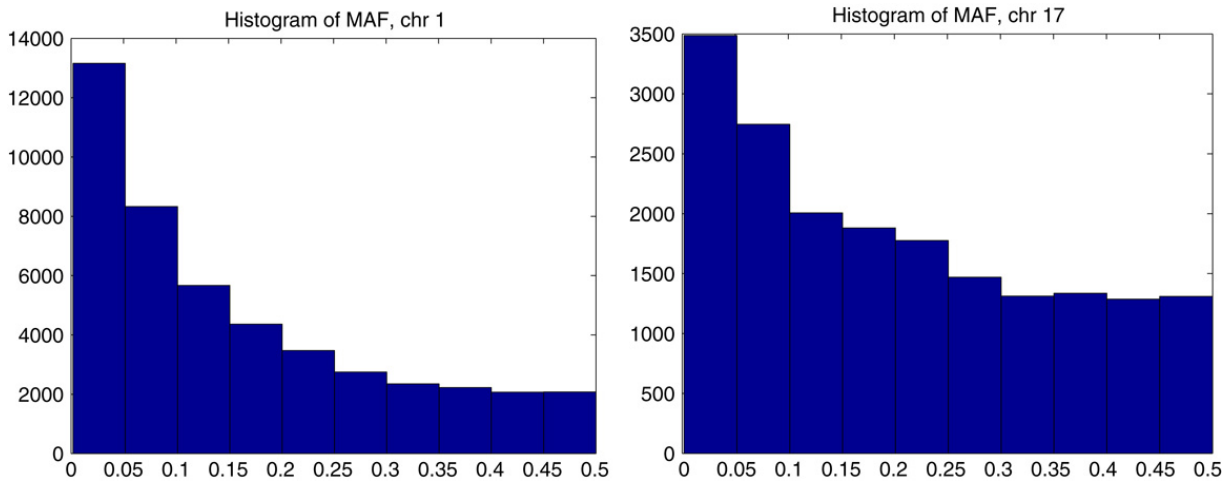


Fig. 1. Histogram of the minimum allele frequency in the markers of chromosome 1 (left) and chromosome 17 (right). In total, there are 46 468 markers in chromosome 1 and 18 611 markers in chromosome 17 in our data.

There are some issues of data quality that one should keep in mind. The MAF, minimum allele frequency, is quite small at some markers, meaning that the value of the marker is nearly constant. This kind of a marker does not help differentiating between the rows (persons) of data. However, we do not remove such markers from the data. Fig. 1 (left panel) shows the histogram of the MAF values in chromosome 1. We see that out of a total of 46 468 markers, about 13 000 or 28% have MAF ≤ 0.05. In chromosome 17 the MAF is not as small on the average, but still out of 18 611 markers, about 19% have MAF ≤ 0.05 (Fig. 1, right panel).

Moreover, the Hardy–Weinberg equilibrium does not always hold in the data. We test the HWE at each marker by the chi squared goodness-of-fit test, and find that the P value of the test is often small: in chromosome 1, 18% of markers have $P < 0.05$, and still 9% of markers have $P < 0.001$. In chromosome 17, the violations are even more frequent: 25% of markers have $P < 0.05$ and 13% of markers have $P < 0.001$. However, the issue of multiple testing has to be taken into account: when performing tens of thousands of tests, it may well be that some of them produce a very small P value just by chance. After Bonferroni correction, there are still 3%–5% of markers having $P < 0.05$ and 2%–3% of markers having $P < 0.001$ in chromosomes 1 and 17. So the data do not always obey the Hardy–Weinberg equilibrium. This is perhaps due to finite population size and nonrandom mating, or to genotyping errors. Indeed, many other methods for finding the population structure (Pritchard et al., 2000; Dawson and Belkhir, 2001; Anderson and Thompson, 2002; Corander et al., 2003) use the HWE violations. For a discussion on the detection of genotyping errors by HWE violations see Hosking et al. (2004). Still, in our experiments shown in the sequel it is seen that the HWE is a useful constraint. In particular, we do not remove the markers that are in Hardy–Weinberg disequilibrium, but keep them in the data.

3.2. Choosing the number of components

The model order, that is, the number of multinomial components must be chosen by the user. Popular ways to do this are the Bayesian Information Criterion (BIC) (Schwarz, 1978), Akaike Information Criterion (AIC) (Akaike, 1973) and the peak of the out-of-sample likelihood curve. BIC and AIC consider the ability of the model to fit to the training data; this is well motivated when the aim is to study the properties of the data at hand and obtain a parsimonious data explanatory model, instead of predicting the behavior of unseen data. Ripley (1996) motivates the use of AIC in models estimated by likelihood maximization. The BIC and AIC optimal numbers of multinomial components in our data are typically 3 or 4. This is not surprising, as the HapMap data are known to contain observations of 4 ethnic groups, two of which might be similar to each other (the Japanese in Tokyo and Han Chinese in Beijing).

On the other hand, the prediction or out of sample performance is often considered important. In this case the model order is selected based on the maximum of the cross validated out of sample likelihood. Smyth (2000) gives a nice analytic motivation for this. Again, not surprisingly, the optimal number of multinomial components in our data sets falls to 3 or a few more.

In the experiments we show results for $K = 3, \ldots, 8$ components.

3.3. Clustering accuracy with respect to the HapMap ethnic groups

3.3.1. HWE-restricted multinomial versus ordinary multinomial

We cluster the individuals into K groups using the multinomial models. More specifically, we select for each individual n the component (that is, cluster) k for which the posterior probability s_{kn} is the largest over $k = 1, \ldots, K$. For each cluster, we check which of the 3 ethnic groups of the HapMap project it best represents. (For simplicity of the presentation of results, we will merge the Japanese and Chinese groups together, to get 3 groups of equal size. However, the multinomial models

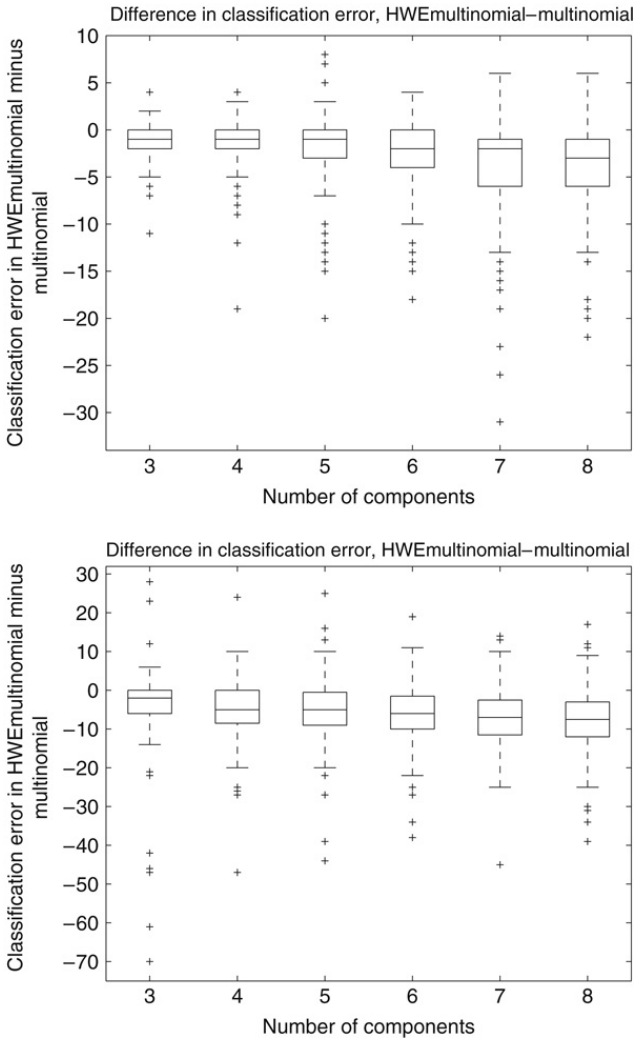


Fig. 2. Classification error in HapMap ethnic groups. Number of classification errors in the HWEmultinomial model minus number of classification errors in the multinomial model. The difference is statistically significant. The box plot shows the distribution of the difference: the boxes have lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the boxes out to the most extreme data value within 1.5 times the interquartile range of the sample. 200 data sets of HapMap chr 1 (top) and chr 17 (bottom), each containing 50 (top) or 20 (bottom) markers at random locations. Horizontal axis: number of components in the multinomial models.

have no intrinsic preferences towards equal group sizes.) Then we count the number of individuals that do not belong to the chosen group, and interpret them as erroneously classified. Let us emphasize that the clustering is completely unsupervised, and the information on the ethnic groups is only used to assess the results.

The estimation of the model parameters by the EM algorithm is prone to local maxima of the likelihood, and depends on the initialization of the model parameters. A way to overcome this was suggested by Blekas et al. (2003) who propose an incremental learning scheme. However, as the estimation of a single model is computationally quite simple, we rely on repeated runs of the EM algorithm using random initializations. We then simply choose the set of parameters which yields the highest in-sample log likelihood. In the experiments presented in this subsection, we initialized the EM algorithm 10 times and chose the best parameters among those runs.

Both multinomial models are reasonably good in identifying the populations. One can conclude that the data lend themselves easily to clustering, and the peculiarities of the data discussed in Section 3.1 do not pose difficulties. Markers having a very small minimum allele frequency, and thus having almost a constant value over the individuals, are not problematic, and neither are the markers whose values violate the Hardy–Weinberg equilibrium.

In the framework of identifying the HapMap populations, the HWE-restricted multinomial model outperforms the original multinomial model; see Fig. 2 for results for marker sets of $T = 20$ and $T = 50$ markers. The results for other sizes of marker sets are similar (data not shown). The box plot shows the distribution of the difference between the number of erroneously classified individuals in the HWEmultinomial model and the multinomial model.

We also used the t test on the difference between the number of erroneously classified individuals in the HWEmultinomial model and the multinomial model. For each number of markers T and number of components K we tested 200 data sets, over which the behavior of the difference is systematic. The difference is statistically significant for most combinations of T and K: in chromosome 1, the P values given by the t tests are $\leq 9 \times 10^{-4}$. The only exceptions

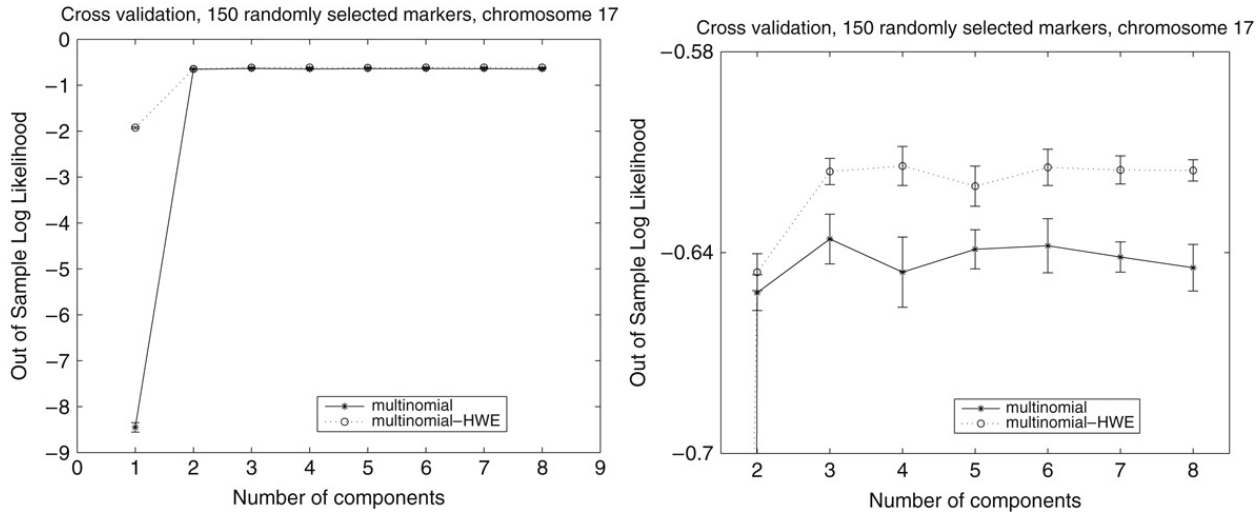


Fig. 3. Left: Out of sample log likelihood in a data set of 150 randomly selected markers of HapMap chr 17. The error bars show one standard error in both directions. Horizontal axis: number of mixture components. Right: the same, zoomed in.

are the cases $T = 150$ and $T = 500$ for which the smallest K are not enough to yield statistical significance. A general observation is that the larger the K, the smaller the P values. Similarly, in chromosome 17, at $T = 20, 50$ or 150, the P values of the t tests are ≤ 0.0452 at all K. In the most difficult setting having $T = 500$, at $K = 3$ the two models perform identically but at $K \geq 4$ the HWEmultinomial model is statistically significantly more accurate.

3.3.2. HWE-restricted multinomial versus structure

As a comparison, we also clustered the data using the *structure* program (Pritchard et al., 2000; Falush et al., 2003). To allow faithful comparisons to the multinomial models, we did not use the admixture nor the correlated allele frequency nor linked loci options, although they are available in the current version of the program. (We would like to emphasize that our aim in this study was to see the effect of complexity control in the clustering accuracy of a multinomial model, and not to devise a widely applicable method such as *structure*.) The *structure* program uses MCMC simulations to fit a probabilistic model to the data, and again the results are dependent on a successful initialization. We initialized the simulation 10 times at each data set and chose the in-sample likelihood optimal parameters among those 10 initializations. The lengths of both the burn-in and actual simulation were 40 000 steps.

In the case of chromosome 1, for $T = 50$ markers, the *structure* program was able to find the populations of HapMap data better than the HWEmultinomial model, at $K = 3$ populations, over 100 data sets. At $K = 4$ and $K = 5$ the HWEmultinomial model was slightly better, but not statistically significantly. At chromosome 17, for $T = 20$ markers, over 90 data sets, the *structure* program was statistically significantly better (P value of t test 8×10^{-4}) at $K = 3$ populations, but at $K = 4$ and $K = 5$ the HWEmultinomial model was better (P values 0.0217 and 0.0260). As a conclusion one can say that the HWEmultinomial model is comparable to the *structure* program in recognizing populations in HapMap data. In addition, the HWEmultinomial model is much simpler to implement and faster to estimate than *structure*: the time required for estimating the HWEmultinomial model by an EM algorithm is less than 1% of the time required for the MCMC simulations in *structure*.

3.3.3. Choosing the number of markers

A further question is the choice of the number of markers T. Irrespective of the method, it seems that the larger the T the better, in terms of clustering accuracy. However, increasing T will increase the number of parameters in the model, making the estimation slower and prone to local minima. At $T = 20$, the amount of erroneously clustered individuals is about 10%–20% at all models; at $T = 50$ it decreases to 1% or 2%, and at $T = 150$ all models mostly perform flawlessly. Among these, $T = 50$ is a suitable compromise.

3.4. Out-of-sample log likelihood

The out-of-sample (that is, test data) log likelihood (OSLL) measures how well the model is able to generalize to unseen observations. Typically models that tend to overfit have poor out-of-sample likelihood.

We used 10-fold cross validation on each data set. That is, we split the observations into 10 parts such that 9/10 of the observations were used to estimate the model. The likelihood of the remaining 1/10 observations given the model was then computed. This was repeated at each part of the data, to get the standard error. At each part, the model was estimated only once, and we did not select any in-sample optimal parameter values as was done in Section 3.3 at the clustering experiment. An example of one data set of 150 randomly selected markers in chromosome 17 is given in Fig. 3.

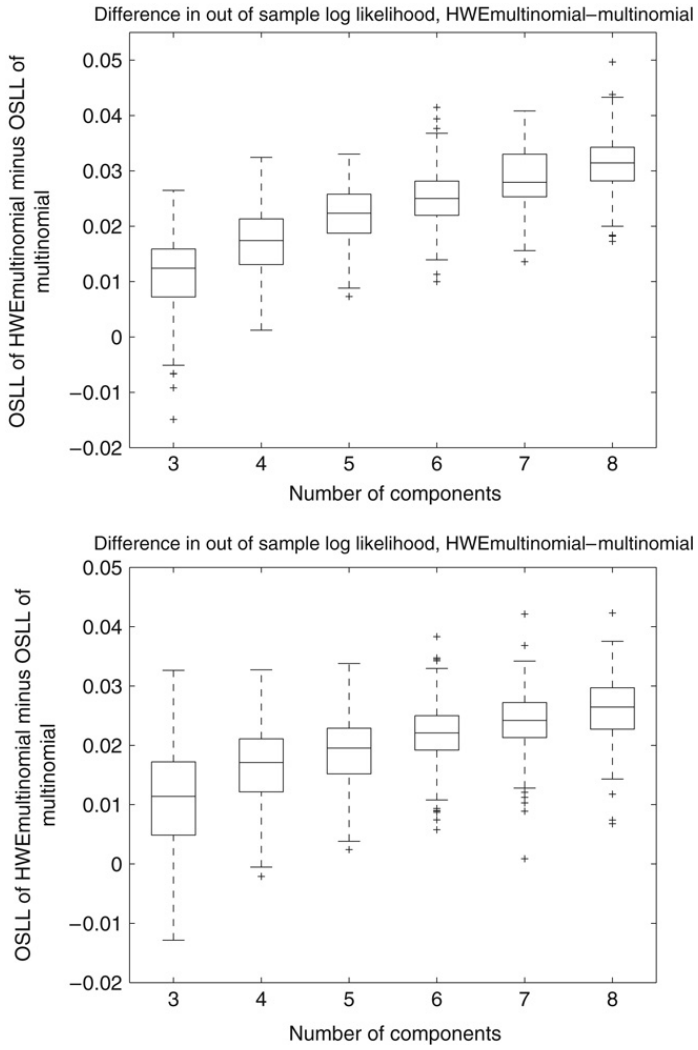


Fig. 4. Difference in the out of sample log likelihoods (OSLL): OSLL of HWEmultinomial minus OSLL of multinomial. The difference is statistically significant at all K. Results over 200 data sets, each containing 150 (top) or 500 (bottom) randomly selected markers of HapMap chr 1 (top) or chr 17 (bottom). Horizontal axis: K, number of mixture components.

The OSLL of the HWE-restricted model is statistically significantly better than the OSLL of the original multinomial model. Box plots of the results over 200 data sets with the number of markers $T = 150$ and $T = 500$ are shown in Fig. 4. Results for other values of T are similar. We also conducted a t test on the difference between the OSLL's and saw that the difference is statistically significant: in both chromosomes 1 and 17, at $T = 20$, 50, 150 and 500, at all numbers of components $K = 3, \ldots, 8$, the P values are extremely small (the largest being 3×10^{-62}.) A general observation is that the P values get smaller as K increases or T decreases. Of course, the P values in a t test are highly dependent on the number of samples, which in our case is quite large, resulting from 10-fold cross validation on 200 data sets.

In terms of the out of sample likelihood it is particularly interesting to see the behavior of the models at a very large number of markers T: the number of parameters of the model is then large compared to the number of observations, and overfitting can be a problem. We see that the HWEmultinomial model outperforms the original multinomial model also at the largest T.

It is not straightforward to compare the out-of-sample likelihoods of the multinomial models and the *structure* program (Pritchard et al., 2000; Falush et al., 2003), as the models are quite different; we have thus chosen not to report out-of-sample likelihoods of *structure*.

3.5. Perlegen data

We also ran experiments on the Perlegen data (Hinds et al., 2005) containing the genotypes of 71 individuals, given at http://genome.perlegen.com/browser/download.html. The OSLL of the HWE-restricted multinomial model was clearly better than the OSLL of the original model, the difference being statistically significant. This again shows that the restricted model generalizes better to unseen data. In terms of clustering the data into populations — the Perlegen data contain 3 ethnic groups — we did not get statistically significant results, due to the small size of the data.

4. Conclusions

We have studied the use of the Hardy–Weinberg equilibrium (HWE) as a means of complexity control in mixture modeling of genotype data. We have presented a multinomial mixture model that takes into account the HWE between the frequencies of marker values. The HWE is a natural biological constraint that is known to hold at population level, assuming random mating and no selective effects. Our hypothesis was that HWE provides a way of regularization or complexity control, preventing overfitting when the number of markers is quite large compared to the number of observations. Our findings indicate that this is indeed the case: the multinomial model incorporating the HWE fits better to unseen data than an ordinary multinomial model. Interestingly, our model is also able to identify the ethnic groups of HapMap data more accurately than the ordinary model.

We have compared our model with one of the state-of-the art methods for identifying populations, namely the *structure* program (Pritchard et al., 2000) and found the methods comparable with each other in HapMap data. In addition, the proposed model is simpler, and therefore significantly faster to estimate than *structure*. A topic of an interesting follow-up study would be to compare the two methods on a data set containing many more populations than the HapMap data.

We wanted to concentrate on the effect of using HWE as a constraint in a multinomial mixture model, and we thus made some simplifying assumptions. Our model is a single-cause model: we assume that each individual originates from one population (one mixture component) only. In the case of admixture populations this does not hold, but instead, the genotype of an individual consists of material from several populations. Probabilistic methods that take this into account include the *structure* model by Pritchard et al. (2000) and the models by Dawson and Belkhir (2001), Bertorelle and Excoffier (1998), Wang (2003), Corander et al. (2004) and Anderson and Thompson (2002). Various multiple-cause latent variable models for multinomial data such as PLSA (Hofmann, 2001), LDA (Blei et al., 2003) and MPCA (Buntine, 2002) could also be used or extended to handle this kind of data.

Another simplifying assumption in our proposed model is that it does not take linkage disequilibrium, or dependence between neighboring markers into account. A possible future direction would be to incorporate this, too. In the present work, we chose the markers randomly along the chromosome, assuming that the distance between randomly chosen markers is quite large and the effect of linkage disequilibrium is thus small. However, it might be possible to include the dependency structure of the markers in the model.

Acknowledgements

The authors are grateful to the International HapMap consortium for providing the HapMap data, and to Jussi Kollin for his help in preprocessing the data. Dr. Ata Kabán's help with the Matlab implementations of the models is appreciated. The comments given by the anonymous reviewers have significantly helped us in enhancing the manuscript. This work was supported by Academy of Finland grant 118653 (ALGODAN).

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihod principle. In: Petrox, B., Csaki, F. (Eds.), Second International Symposium on Information Theory. pp. 267–281.

Anderson, E., Thompson, E., 2002. A model-based method for identifying species hybrids using multilocus genetic data. Genetics 160, 1217–1229.

Barrett, J.C., Fry, B., Maller, J., Daly, M.J., 2005. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics 21 (2), 263–265.

Bertorelle, G., Excoffier, L., 1998. Inferring admixture proportions from molecular data. Molecular Biology and Evolution 15 (10), 1298–1311.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Blekas, K., Fotiadis, D.I., Likas, A., 2003. Greedy mixture learning for multiple motif discovery in biological sequences. Bioinformatics 19 (5), 807–817.

Blekas, K., Likas, A., 2004. Incremental mixture learning for clustering discrete data. In: Methods and Applicatios of Artificial Intelligence. Third Hellenic Conference on Artificial Intelligence, SETN 2004. In: Lecture Notes in Artificial Intelligence (LNAI), vol. 3025. Springer, pp. 210–219.

Buntine, W., 2002. Variational extensions to EM and multinomial PCA. In: Machine Learning: ECML 2002. In: Lecture Notes in Artificial Intelligence (LNAI), vol. 2430. Springer-Verlag, pp. 23–34.

Corander, J., Waldmann, P., Sillanpää, M.J., 2003. Bayesian analysis of genetic differentiation between populations. Genetics 163, 367–374.

Corander, J., Waldmann, P., Marttinen, P., Sillanpää, M.J., 2004. Baps 2: Enhanced possibilities for the analysis of genetic population structure. Bioinformatics 20 (15), 2363–2369.

Cornuet, J.-M., Piry, S., Luikart, G., Estoup, A., Solignac, M., 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics 153 (4), 1989–2000.

Dawson, K.J., Belkhir, K., 2001. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genetics Research 78, 59–77.

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

Excoffier, L., Estoup, A., Cornuet, J.-M., 2005. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. Genetics 169 (3), 1727–1738.

Falush, D., Stephens, M., Pritchard, J.K., 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164 (4), 1567–1587.

Falush, D., Stephens, M., Pritchard, J.K., 2007. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. Molecular Ecology Notes 7 (4), 574–578.

Fung, H.-C., Scholz, S., Matarin, M., Simon-Sanchez, J., Hernandez, D., Britton, A., Gibbs, J.R., Langefeld, C., Stiegert, M.L., Schymick, J., Okun, M.S., Mandel, R.J., Fernandez, H.H., Foote, K.D., Rodriguez, R.L., Peckham, E., Vrieze, F.W.D., Gwinn-Hardy, K., Hardy, J.A., Singleton, A., 2006. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: First stage analysis and public release of data. The Lancet Neurology 5 (11), 911–916.

Glickman, M.E., Kao, M.-F., 2005. Apo-E genotypes and cardiovascular diseases: A sensitivity study using cross-validatory criteria. The Biometrical Journal 47, 541–553.

Hardy, G.H., 1908. Mendelian proportions in a mixed population. Science 28 (706), 49–50.
Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.B., Frazer, K.A., Cox, D.R., 2005. Whole-genome patterns of common DNA variation in three human populations. http://genome.perlegen.com/browser/download.html.
Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42, 177–196.
Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., Riley, J., Xu, C.-F., 2004. Detection of genotyping errors by Hardy-Weinberg equilibrium testing. European Journal of Human Genetics 12, 395–399.
Kocsis, I., Györffy, B., Németh, E., Vásárhelyi, B., 2004. Examination of Hardy-Weinberg equilibrium in papers of Kidney International: An underused tool. Kidney International 65, 1956–1958.
McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J., Altshuler, D.M., 2006. The International HapMap Consortium, Common deletion polymorphisms in the human genome. Nature Genetics 38, 86–92. published online: 4 December 2005.
Meilă, M., Heckerman, D., 2001. An experimental comparison of model-based clustering methods. Machine Learning 42 (1/2), 9–29.
Patist, J.P., 2006. A fast implementation of the EM algorithm for mixture of multinomials. In: Li, X., Zaiane, O., Li, Z. (Eds.), Advanced Data Mining and Applications. In: Lecture Notes in Artificial Intelligence (LNAI), vol. 4093. Springer, pp. 517–524.
Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.
Rannala, B., Mountain, J.L., 1997. Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences 94 (17), 9197–9201.
Redner, R.A., Walker, H.F., 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review 26 (2), 195–239.
Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press.
Rufo, M., Perez, C., Martin, J., 2007. Bayesian analysis of finite mixtures of multinomial and negative-multinomial distributions. In: Advances in Mixture Models. Computational Statistics & Data Analysis 51 (11), 5452–5466 (special issue).
Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6 (2), 461–464.
Smyth, P., 2000. Model selection for probabilistic clustering using cross-validated likelihood. Statistics and Computing 10 (1), 63–72.
Song, X.-Y., Lee, S.-Y., Ng, M.C.Y., So, W.-Y., Chan, J.C.N., 2007. Bayesian analysis of structural equation models with multinomial variables and an application to type 2 diabetic nephropathy. Statistics in Medicine 26 (11), 2348–2369.
Tang, H., Peng, J., Wang, P., Risch, N., 2005. Estimation of individual admixture: Analytical and study design considerations. Genetic Epidemiology 28 (4), 289–301.
The International HapMap Consortium, 2003. The international HapMap project. Nature 426, 789–796.
Wang, J., 2003. Maximum-likelihood estimation of admixture proportions from genetic data. Genetics 164, 747–765.
Weinberg, W., 1908. Über den Nachweis der Verebung beim Menschen. (On the demonstration of heredity in man). Jahreshefte des Vereins für Vaterländische Naturkunde in Württemberg 64, 368–382.
Weinberg, C.R., Morris, R.W., 2003. Invited commentary: Testing for Hardy–Weinberg disequilibrium using a genome single-nucleotide polymorphism scan based on cases only. American Journal of Epidemiology 158, 401–403.
Wigginton, J.E., Cutler, D.J., Abecasis, G.R., 2005. A note on exact tests of Hardy-Weinberg equilibrium. The American Journal of Human Genetics 76, 887–893.
Wu, B., Liu, N., Zhao, H., 2006. PSMIX: An R package for population structure inference via maximum likelihood method. BMC Bioinformatics 7, 317.