Mathematical Statistics
Stockholm University

Classification of SNP genotypes by a Gaussian mixture model in competitive enzymatic assays

Hedvig Norlén, Erik Pettersson, Afshin Ahmadian,
Joakim Lundeberg and Rolf Sundberg

Research Report 2008:3

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2008:3**,
http://www.math.su.se/matstat

Classification of SNP genotypes by a Gaussian mixture model in competitive enzymatic assays

Hedvig Norlén, Erik Pettersson, Afshin Ahmadian, Joakim Lundeberg and Rolf Sundberg*

February 2008

Abstract

We present statistical models and methods for classification of bi-allelic SNP genotypes when data represent two signal intensities, one signal x from a primer matching one of the alleles, and the other signal y matching the other allele. One such technique is protease-mediated allele-specific extension (PrASE), and the study is at the same time a case study on PrASE data. Most information for classification is contained in the variate $\log(x/y)$, for which we derive a special 3-component mixture model from molecular principles. We describe inference in this mixture model, but we also discuss other topics such as assessing the number of components, the information available in the orthogonal variation, detection of overall outlying individuals, and supplementary use of the Hardy–Weinberg law.

Key words: EM algorithm, genotype calling, PrASE, single nucleotide polymorphism

*Postal address: Norlén: Mathematical Statistics, Stockholm University, Pettersson, Ahmadian, Lundeberg: Gene technology, Royal Institute of Technology (KTH), Sundberg: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: rolfs@math.su.se (corresponding author). Website: www.math.su.se/∼rolfs

1 Introduction

Much recent effort in genomics has focused on analysis of the bi-allelic base variations called single nucleotide polymorphisms (SNPs). These variations are distributed across the genome with a frequency as high as one in every 1000 base pairs (bp) (Hapmap, 2005). Occasionally it has been possible to associate (link) a certain SNP with a specific disease, but in order to find clues to more complex diseases or phenotypic variation, combinations of several SNPs must be investigated. In such cases the investigations will require characterization of a large number of SNPs in a large number of individuals. This task requires rapid and automatic genotyping techniques, and the present work aims at developing the statistical basis for an algorithm for automated assignment of each individual to its genotype status (genotype calling), that is for each SNP position characterization of the individual as homozygous with one of the two base variants, or as heterozygous.

The present study is specifically aimed at data obtained by a competitive enzymatic assay called protease-mediated allele-specific extension (PrASE), in which 3'-terminus allele-specific primers match one each of the two alleles and mismatch the other (Hultin *et al.*, 2005). However, similar methodology has been found adequate also under other data-generating mechanisms, in particular by Carvalho *et al.* (2007).

In the PrASE assay, with fluorescent-labelled nucleotides detectable by a scanner, a homozygous template should generate a relatively strong signal for one of the allele-specific primers and a relatively weak signal for the other. Accordingly, a heterozygous sample generates moderate signals for both primers. More precisely, the data for this study were collected after specific amplification with Tri-nucleotide Threading (TnT) technology (Pettersson *et al.*, 2006) and genotyping with PrASE.

Several other approaches to genotype classification have been reported recently, for use with different genotyping methods (Lovmar *et al.*, 2005; Hardenbol *et al.*, 2005; Callegaro *et al.*, 2006; Moorhead *et al.*, 2006; Xiao *et al.*, 2007; Plagnol *et al.*, 2007; Carvalho *et al.*, 2007). These papers range from relatively primitive undertakings, represented by nonparametric bivariate clustering without explicit statistical models, to more sophisticated univariate or bivariate parametric mixture models exploiting more or less of the structure that could be expected in data.

The aim of the present paper is to demonstrate that such structures exist and can be incorporated in the statistical analysis. From molecular principles we will derive a special bivariate Gaussian mixture model for suitably transformed PrASE data, and the model will be seen to fit well to typical such data. A similar type of model with its associated classification method has recently been proposed for Affymetrix SNP array (Genechip) data by Carvalho *et al.* (2007). From a molecular point of view, the types of data are different. Also, the latter paper suggests the model without particular motivation and without considering more than a univariate marginal part of the model. Another difference is in the type of applications in mind. Affymetrix genome-wide SNP arrays are pre-made arrays of many thousands of SNPs, whereas PrASE is more intended for tailor-made choice of SNPs in investigations of a moderate size. The Infinium assay from Illumina, for which a genotype calling algorithm has been proposed by Teo et al. (2007), also, like Affymetrix, relies on a different molecular mechanism

and is targeting thousands or many more of fixed SNPs. Furthermore, the model behind their genotype calling method can be criticized for being unnatural, see Section 2.

The outline of this report is as follows. First the experimental situation will be described. Next, the statistical model is developed in detail (Sections 3–5). The methods used for model fitting are discussed in Sections 6 and 7. with results presented and discussed in Section 8. Sections 9–10 bring up some further aspects (outliers, and Hardy–Weinberg equilibrium), before the conclusions are summarized in Section 11.

2 Experimental design and data

The PrASE genotyping technique is a further development of the allele-specific extension (ASE) assays, where fluorescently labeled products are distinguished by their signature tags. For a detailed description of the PrASE technique, see Hultin *et al.* (2005). The signals acquired in this approach are not strictly binary, as ideal data would be, mainly due to imperfect primer synthesis and to differences in hybridization properties of the signature tags. These differences can be attributed to sequence context and to the quality of the synthetic primers. Hence, each SNP will have some unique characteristics, and a flexible algorithm that can handle such differences is desired.

In this study 75 SNPs were identified and selected from genes linked to various types of human cancers. Primers for parallel tri-nucleotide thread (TnT) amplification were designed by a custom-made software tool (Pettersson *et al.*, 2006). Samples from 96 individuals were amplified and analyzed by PrASE in an array format. The fluorescence detector registered the signal intensities x and y from the primer extension of the first and second allele, respectively. The left hand side of Figure 1 shows how SNP genotype data of PrASE typ appear in these original variables for a well-behaved SNP. Often x and y are replaced by two other characterics, the *allelic fraction* $AF = x/(x+y)$, $0 \leq AF \leq 1$, and the *total fluorescent signal intensity* (on log-scale) $t = \log_{10}(x+y)$, also called the *strength*. An equivalent substitute for AF often used is the *contrast* $c = (x-y)/(x+y) = 2(AF) - 1$. The right hand side scatter-plot of Figure 1 shows t versus AF for the same SNP. Many more such diagrams are found in Hultin *et al.* (2005) and Pettersson *et al.* (2006), and similar types of diagrams are found for example in Lovmar *et al.* (2005) and Hardenbol *et al.* (2005).

When SNP genotype data are represented as in one of the diagrams of Fig. 1, it is understandable that scientists have used cluster analysis in a nonparametric way. It is obvious that (x, y)-data do not at all have Gaussian component distributions, and there are also problems with Gaussian distributions for the allelic fraction AF or the contrast c, since these quantities are restricted to bounded intervals. Some algorithms are based on Gaussian distributions for AF or c, even though the homozygotes are found close to the boundaries (Moorhead *et al.*, 2006; Plagnol *et al.*, 2007). Others have used various artificial constructions. In Hardenbol *et al.* (2005), AF is modelled by 'Gaussian distributions with non-Gaussian tails', and Teo *et al.* (2007) use truncated t-distributions. Furthermore, the three components in Figure 1 (right) clearly do not have the same variance or any other simple feature relating them, so typically the three components are modelled without any parameters in common.

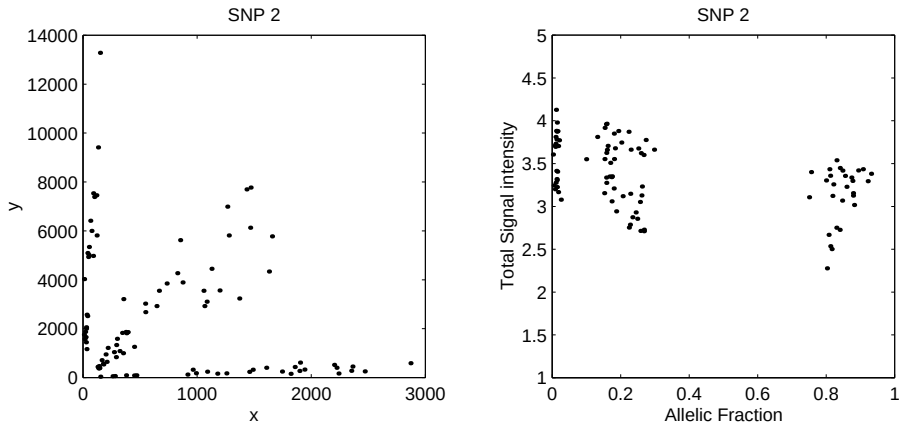

Figure 1: *Two common plot types, illustrated on SNP number 2.*
Left hand side: Plot of signals x and y.
Right hand side: Plot of total signal intensity against allelic fraction (AF)

Figure 2 shows a scatter-plot of $\log y$ versus $\log x$ for the PrASE data. The first impression now is that the three point clusters are located as three parallel and quite similar bands, whose positions differ essentially along the 135° direction. Differences along this direction can be expressed in terms of AF,

$$\log x - \log y = \log\{AF/(1-AF)\} \tag{1}$$

or *vice versa* (with log read as $\log_e$),

$$AF = e^{\log x - \log y}/(1 + e^{\log x - \log y}) \tag{2}$$

In the sequel, these data will be represented by the variables $\log\ x$ and $\log\ y$, and a technically and biologically motivated model for what is seen in Figure 2 and other diagrams of the same type will be derived.

3 Statistical modelling of data

The basic assumption when modelling competitive enzymatic, array-based SNP data, for a selected SNP and the corresponding pair of primers, is that the expected contributions from different sources add, and that the fluorescence signal of such a contribution is proportional to the intensity of the source. This means that the alleles of a homozygous sample will together contribute twice as much as the same but single allele in a heterozygous sample. On the other hand the latter sample will also receive an additive contribution from the other allele. We will formalize these primer-mediated contributions below.

We can imagine three types of signal, of more or less specific character, to be called match, mismatch and unspecific. Under specified experimental conditions and for a heterozygous sample, let the primer-mediated contributions to be expected for primer $i = 1, 2$ (*i.e.* x and y) be α_{ij}, $j = 0, 1, 2$, as follows:

α_{11} = match contribution from allele 1 with primer 1

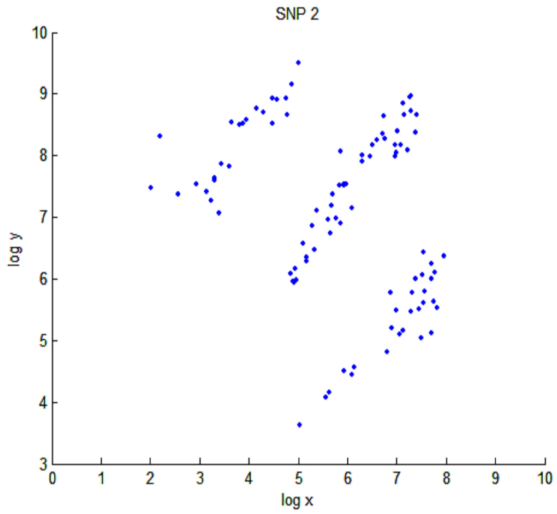

Figure 2: *Plot of* $\log_e y$ *versus* $\log_e x$ *for SNP number 2, all 96 individuals.*

α_{12} = mismatch contribution from allele 2 with primer 1
α_{10} = unspecific contribution with primer 1, not related to any of the alleles
$\alpha_{1'0}$ = unspecific contribution with antitag, unrelated to the tag of primer 1

α_{21} = mismatch contribution from allele 1 with primer 2
α_{22} = match contribution from allele 2 with primer 2
α_{20} = unspecific contribution with primer 2, not related to any of the alleles
$\alpha_{2'0}$ = unspecific contribution with antitag, unrelated to the tag of primer 2

These primer-mediated contributions from the two possible homozygous samples or a heterozygous sample are expressed in Table 1.

Table 1 Primer-mediated contributions from homozygous and heterozygous samples for different primers 1 and 2 (corresponding to x and y).

	Homozyg. (1+1)	Heterozyg. (1+2)	Homozyg. (2+2)
Primer 1 (x)	$2\alpha_{11} + \alpha_{10} + \alpha_{1'0}$	$\alpha_{11} + \alpha_{12} + \alpha_{10} + \alpha_{1'0}$	$2\alpha_{12} + \alpha_{10} + \alpha_{1'0}$
Primer 2 (y)	$2\alpha_{21} + \alpha_{20} + \alpha_{2'0}$	$\alpha_{21} + \alpha_{22} + \alpha_{20} + \alpha_{2'0}$	$2\alpha_{22} + \alpha_{20} + \alpha_{2'0}$

The aim of the primer construction is that primer 1 should be particularly sensitive to allele 1, and analogously for primer 2 and allele 2, such that $\alpha_{11} >> \alpha_{12} + \alpha_{10} + \alpha_{1'0}$, and analogously for α_{22}.

Typically the unspecific contributions are small even in comparison with the mismatches, and therefore can be neglected. This will be assumed in the sequel, but is not always satisfied. For occasional SNPs, relatively high unspecific signal contributions can be observed, blurring the picture. Due to sample complexity they cannot be predicted. They occur rarely, but when they do, they are non-systematic and cannot be modelled, in contrast to the systematic relationship between match and mismatch primers.

The two primers will typically not be equally efficient, due to differences in primer synthesis (hybridization properties towards template) and the fact that they carry different signature tags (hybridization properties towards the anti-tags on the array). This means that for example α_{11} and α_{22} will differ, and they may differ substantially. However, a primer efficiency advantage factor for primer 1 over primer 2, $\alpha_{11} > \alpha_{22}$, should be the same for α_{12} versus α_{21}, so we have reason to expect ratios α_{11}/α_{22} and α_{12}/α_{21} to be the same. It follows that we should have equality between α_{12}/α_{11} and α_{21}/α_{22}, say $\alpha_{12}/\alpha_{11} = \alpha_{21}/\alpha_{22} = \theta$, where θ is an unknown but typically small efficiency number, representing mismatch to match efficiency. Then Table 1 can be rewritten in the form of Table 2.

Table 2 Primer-mediated contributions from homozygous and heterozygous samples for different primers x and y when unspecific contributions are negligible, and $\alpha_{12}/\alpha_{11} = \alpha_{21}/\alpha_{22} = \theta$

	Homozyg. (1+1)	Heterozyg. (1+2)	Homozyg. (2+2)
Primer 1 (x)	$2\alpha_{11}$	$(1+\theta)\alpha_{11}$	$2\theta\alpha_{11}$
Primer 2 (y)	$2\theta\alpha_{22}$	$(1+\theta)\alpha_{22}$	$2\alpha_{22}$

The intensity level will vary between individuals, and this variation is often substantial. Hence we should also allow an individual intensity factor λ, being the same for the two primers but varying between individuals. Without restriction we may take λ to have geometric mean 1 over all individuals, that is average zero on log-scale. In a log–log diagram for (x, y), Table 2 implies that we can expect the homozygous samples to be found in points

$$P_\alpha + (0, \log\theta) + \log\lambda\,(1,1) \tag{3}$$

and

$$P_\alpha + (\log\theta, 0) + \log\lambda\,(1,1), \tag{4}$$

where $P_\alpha = (\log(2\alpha_{11}), \log(2\alpha_{22}))$, and the heterozygous sample in the point

$$P_\alpha + \log\{(1+\theta)/2\}(1,1) + \log\lambda\,(1,1). \tag{5}$$

These model implications are shown in Figure 3. In words this means that irrespective of the values of α_{11}, α_{22} and θ, the heterozygous sample (point B) will be located along a 45° line below P_α, and the homozygous samples (points A and C) will be located at equal distances $\gamma(\theta)$ from the perpendicular projection of P_α on a line of slope 135°. In the latter direction, the heterozygous sample will thus be located in the middle between the two homozygous samples, but moved a distance $\nu = \nu(\theta)$ along the 45° direction. Because this distance is always > 0, the expected position of a heterozygous sample is above the 135° line between the two homozygous samples (A and C).

This theoretical reasoning is supported by data, see Figure 4 for an example, where we have introduced variables u and w representing data along the 45° and 135° axes,

$$u = (\log x + \log y)/\sqrt{2} \tag{6}$$

$$w = (\log x - \log y)/\sqrt{2}. \tag{7}$$

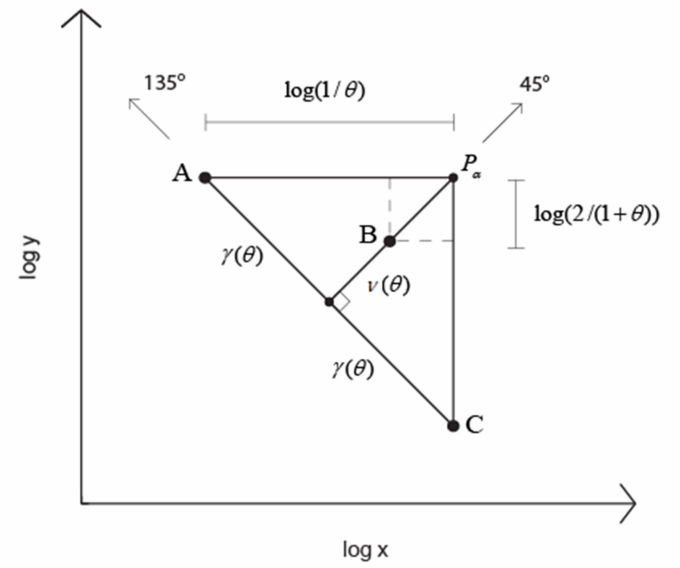

Figure 3: *Geometric illustration of the expected model implications on logarithmic scale.*

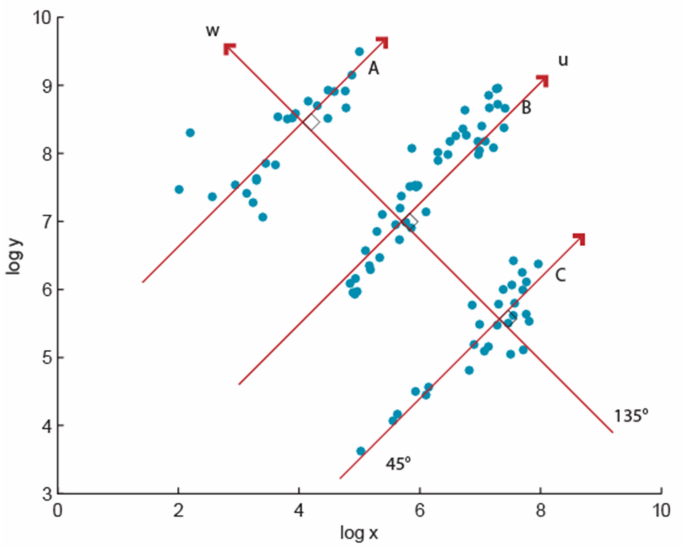

Figure 4: *Data for SNP number 2, as in Figure 2, but with u-and w-axes inserted, representing variation in the 45° and 135° directions.*

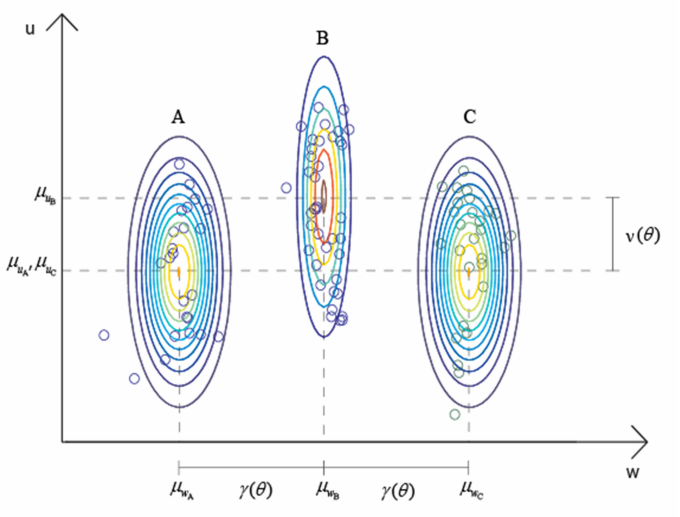

Figure 5: *Data for SNP number 2, represented by coordinates u and w. Groups A and C are the homozygous genotypes and group B the heterozygous genotype.*

Figure 4 illustrates that the means in the 135° direction (w) for the homozygous genotypes (A and C) are equidistant from the mean of the heterozygous genotype (B). The analyses to follow will be performed in the derived variables u and w. Expressed in (u, w) we assume that data derive from three bivariate normal distributions. We further assume that u and w are uncorrelated. This holds automatically in the typical case when $\log x$ and $\log y$ have the same variance. Figure 5 illustrates this, where:

- μ_{w_A} , μ_{w_B} and μ_{w_C} are the equidistant means in the w-direction, with pairwise distance

$$\gamma(\theta) = \log(1/\theta)/\sqrt{2} \quad \text{or} \quad \theta = \exp(-\sqrt{2}\gamma), \tag{8}$$

- $\mu_{u_\mathrm{A}} = \mu_{u_C}$ are the means in the u-direction for the two homozygous genotypes A and C, whereas the heterozygous genotype B has mean value $\mu_{u_\mathrm{B}} = \mu_{u_\mathrm{A}} + \nu(\theta)$ with

$$\nu(\theta) = \gamma(\theta) - \sqrt{2}\log(2/(1+\theta)) > 0. \tag{9}$$

This means that the six location parameters in the three bivariate normal distributions are reduced to three parameters.

An assumption for the technique to be efficient is that the distance γ is large, *i.e.* that the quotient θ be small. To see if this assumption is fulfilled, we anticipate the estimation of γ and show a histogram over SNPs of estimated θ-values, calculated from estimated γ-values through equation (8). This histogram is shown in Figure 6. The θ-values are reasonably small, typically ≤ 0.1.

In addition, there is extensive random variation in the data. One reason is biological variation between individuals (and between samples within individual), for example in the amount of genomic material sampled. Other variation

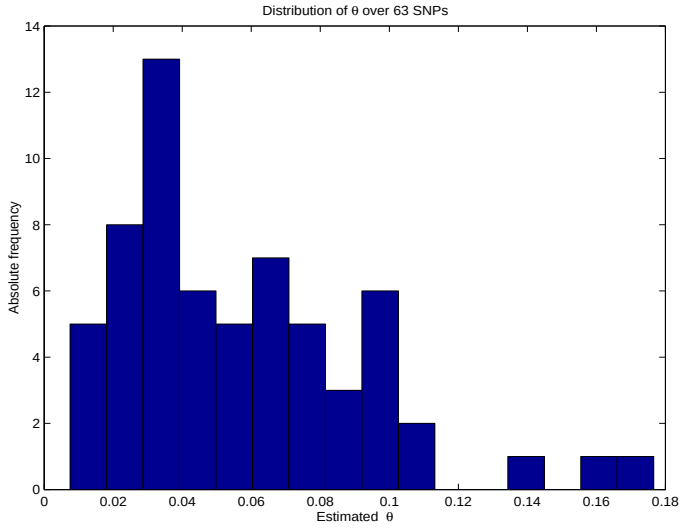

Figure 6: *Histogram of estimated θ-values, representing match–mismatch efficiency for SNPs with 3 or 2 genotypes.*

is more purely technical, for example between subarrays ("chip" effects) and between print tips, and measurement errors. Much of this variation can be expected to be multiplicative in character and represented by variation in the proportionality factor λ, common to both x and y, so after transformation of variables to u and w this will be seen exclusively in u. Variation in the amount of sampled primers and genomic material will be of this kind, but much other variation, too, for example the variation between subarrays. We must also expect occasional gross errors, which may obscure the picture for an individual or for a particular SNP. Biological variation due to pathological genetic composition will be assumed absent, but if it appears, anyhow, it will appear as outliers.

Weak signals are likely to be influenced by the background and background intensity variation, which is not necessarily multiplicative. This can perhaps largely explain why the heterozygous groups typically have a smaller variance in w than the homozygous groups. The reason should then be that for the homozygous groups one of the signals, x or y, should be small, whereas for the heterozygous group none of them is small.

The left-most cluster in Figures 4 and 5 indicates that the variance in w is somewhat larger when the measured intensity u is quite low, but for simplicity it will subsequently be assumed constant. A more sophisticated analysis would give different weight to the w-values of different observations depending on their u-values.

4 Are u-data useful for SNP classification?

We show here that intensity data u are not of much value for the separation between genotypes. However, they can be useful for the characterization of genotypes when only two genotypes are reported in the sample, see Section 10.

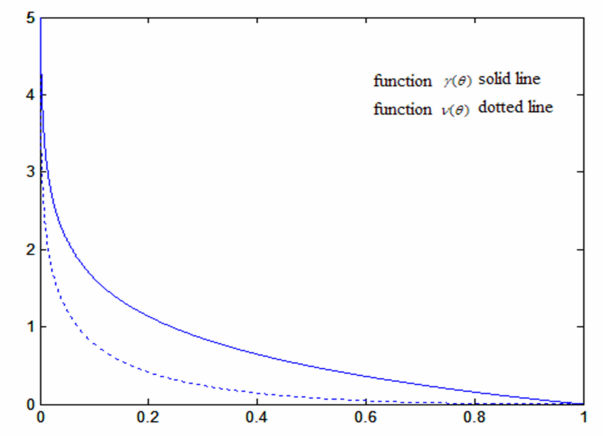

Figure 7: Plot of functions $\gamma(\theta)$ (solid line) and $\nu(\theta)$ (dotted line), $0 < \theta \leq 1$.

In Figure 3 we introduced the distances $\gamma(\theta)$ and $\nu(\theta)$ in the w and u directions, respectively, as functions of the efficiency parameter θ. given by formulae (8) and (9).

It follows that there is an unambiguous relationship between γ and ν through θ. Here θ is aimed at being a small positive number. When we have problems separating the genotypes, it is because θ is not small enough in relation to the noise levels in w and u, Figure 7 shows how the distances $\gamma(\theta)$ and $\nu(\theta)$ decrease as θ increases from 0 to 1. The conclusion is that in comparison with w, u is relatively useless for inference about the parameters. This conclusion is based on the following facts:

- The distance γ in w is always bigger than the distance ν in u, whereas in contrast the sample variances are typically substantially higher in u than in w. Hence, from u-data we have to make inference about a smaller mean value difference from samples with a higher variance.
- As θ increases, ν (dotted line) approaches zero much faster than γ (solid line), so the relative contribution of information from u-data to that of w-data is decreasing when the total information in the data for discrimination decreases.

In other words:

- When the information in u is substantial, it is not needed, because the information content in w is high enough for precise discrimination.
- When the information content in w is not very high, so additional information could help, the relative amount of information in u is too small to be of value.

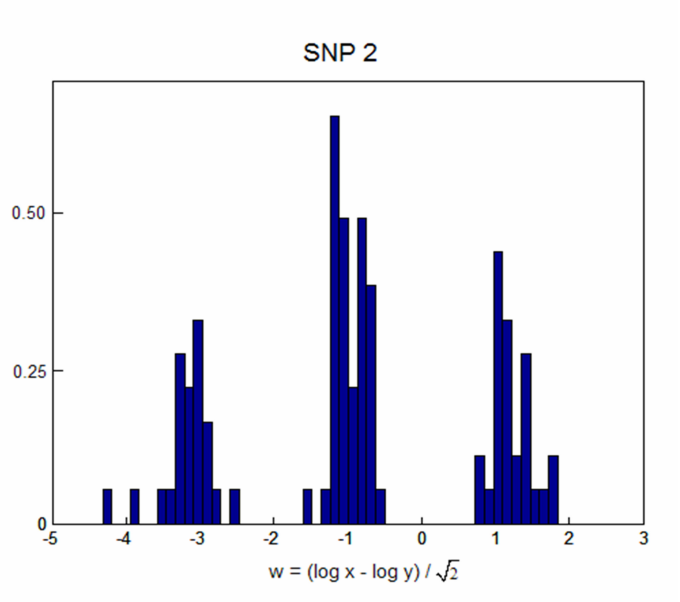

Figure 8: *Data w for SNP number 2. The middle part of the histogram represents the heterozygous genotype (group B), and it is flanked by the two homozygous types (groups A and C).*

5 Normal mixture model for w-data

In previous sections we have seen that the variability in the u direction is considerable, and that the u-data usually contribute little to the discrimination between genotypes. Therefore, the clustering analysis to be described will be based on the w-data alone.

Consider the variable $w = (\log x - \log y)/\sqrt{2}$, for a given SNP. For the purpose of illustration, a histogram for w is shown in Figure 8. We assume that the observed data $w_1, w_2, ..., w_n$ for one SNP data set come from a mixture of three univariate normal distributions with equidistant mean values $\mu_1 < \mu_2 < \mu_3$, $\mu_3 - \mu_2 = \mu_2 - \mu_1 (= \gamma)$, see Figure 5, and variances, σ_i^2 , $i = 1, 2, 3$, with $\sigma_1^2 = \sigma_3^2$. The unknown mixing proportions π_i, $i = 1, 2, 3$, satisfy $0 \leq \pi_i \leq 1$ with $\sum \pi_i = 1$. The complete parameter vector can be taken as

$$\beta = \left(\mu_1,\ \mu_2,\ \mu_3,\ \sigma_1^2,\ \sigma_2^2,\ \sigma_3^2, \pi_1,\ \pi_2,\ \pi_3\right) \tag{10}$$

or alternatively with μ_A, μ_B, etc. There are three linear restrictions as specified above, one in μ, one in σ, and one in π. The normal mixture model to be fitted has the density

$$f(w;\beta) = \sum_{i=1}^{3} \frac{\pi_i}{\sigma_i}\ \phi\left(\frac{w-\mu_i}{\sigma_i}\right) \tag{11}$$

where $\phi(.)$ is the standardized normal density function. This model is illustrated in Figure 9 (the best fitting such model).

For some data sets one of the three groups is missing, most likely one of the homozygous groups. In this case of a two-component mixture, the parameter

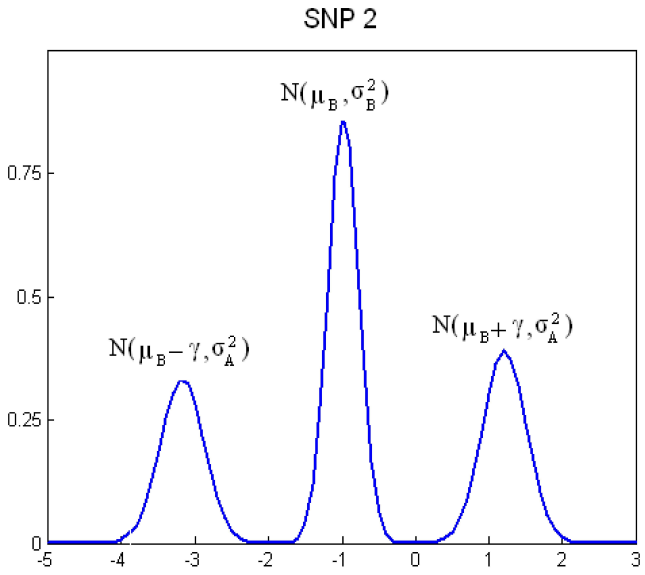

Figure 9: *Mixture model with three assumed normal densities, illustrated for SNP 2 (as fitted by MLE).*

restrictions on means and variances vanish. More rare data sets are represented by a single normal component. In this case ML estimation is standard and need not be discussed here. How to determine the appropriate number of components is discussed in Section 7.

6 Parameter estimation by the EM algorithm

We have used the maximum likelihood (ML) method for parameter estimation. The Expectation-Maximization (EM) algorithm is an iterative method for computing these estimates, when standard likelihood maximization is numerically difficult or impossible because of incomplete data problems (Dempster *et al.*, 1977; McLachlan and Thriyambakam, 1997). By incomplete data in this case we refer to the lack of membership knowledge that characterizes mixtures. With a label on each observation telling its genotype, we would have had "complete" data, and the estimation would have been much simpler. The idea is to pretend we know the complete data and adjust this knowledge in a two-step iterative method, an *expectation* E-step alternating with a *maximization* M-step.

From the mixture model we know that each observation originates from one of the $g = 2$ or $g = 3$ distribution components, and we use v as the unobserved indicator vector of dimension g, with one component 1 and all others 0, $v_{ij} = 1$ when individual j belongs to component i. Together (v, w) form the "complete" data.

The case $g = 2$ of two mixed arbitrary normal distributions is standard in the literature on mixtures, even though mixtures with a common variance parameter are perhaps more common. Our case $g = 3$ is nonstandard, however,

due to the parameter restrictions, which make the likelihood smoother and the resulting estimates more precise, but the formulae more complicated. When $g = 2$ the complete data family is a full exponential family, and the M-step is unique and extremely simple, see Sundberg (1974). When $g = 3$, our complete data model is a curved exponential family, *i.e.* the minimum sufficient statistic is of higher dimension than the parameter. This typically implies that the ML estimate does not have an explicit form, and the global likelihood maximum point need not be a unique root of the likelihood equations.

In fact, the situation has some similarities with the Behrens–Fisher problem (*i.e.* two normal samples with common mean but different variances), where it is known that the corresponding mean value estimator can be a local minimum point of the likelihood, when the data are such that they do not fit the model. It is not clear and has not been investigated if this might possibly be the case here, too. However, it is not likely to be an important defect, if the likelihood maximum is slightly underestimated when data do *not* fit the model.

The "complete" data vector is (v, w) and the corresponding log-likelihood for the parameter vector β has a partially multinomial form

$$\begin{aligned} \log L(\beta; v, w) &= \sum_{i=1}^{3}\sum_{j=1}^{n} v_{ij} \log \pi_i + \sum_{i=1}^{3}\sum_{j=1}^{n} v_{ij} \log f_i(w; \eta_i) \\ &= \sum_{i=1}^{3}\sum_{j=1}^{n} v_{ij} \log \pi_i - (n/2)\log(2\pi) \\ &- \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{n} v_{ij} \left\{\log \sigma_i^2 + (w_j - \mu_i)^2/\sigma_i^2\right\} \end{aligned} \tag{12}$$

E-step. In this step we calculate, under the current parameter $\beta^{(k)}$ after k iterations, the conditional expected value of the complete data log-likelihood (12), given the observed data (the w-values). That is, we calculate

$$E_{\beta^{(k)}}[\log L(\beta; v, w) \,|\, w] \tag{13}$$

Note that $\log L$ is linear in the unobservable v_{ij}. This implies that the conditional expectation to be calculated is obtained by simply replacing v_{ij} by its conditional expected value given w-data, that is (after step k) by

$$v_{ij}^{(k)} = E_{\beta^{(k)}}[V_{ij} \,|\, w] = P_{\beta^{(k)}}(V_{ij} = 1 \,|\, w) \tag{14}$$

From Bayes' theorem we get

$$v_{ij}^{(k)} = \frac{\pi_i^{(k)} f_i(w_j; \eta_i^{(k)})}{\sum_{l=1}^{g} \pi_l^{(k)} f_l(w_j; \eta_l^{(k)})} = \frac{\frac{\pi_i^{(k)}}{\sigma_i} e^{-\frac{1}{2\sigma_i}(w_j - \mu_i)^2}}{\sum_{i=1}^{g} \left(\frac{\pi_i^{(k)}}{\sigma_i} e^{-\frac{1}{2\sigma_i}(w_j - \mu_i)^2} \right)} \tag{15}$$

for $i = 1, 2, 3$ and $j = 1, \ldots, n$ (dependence of μ and σ on k suppressed). Note that $v_{ij}^{(k)}$ is the posterior probability that the unit with observed value w_j belongs to the ith component of the mixture, and the ultimate aim of the

statistical analysis is to provide for each individual the best v-values.

M-step. In iteration $k+1$ we update $\beta^{(k)}$ to $\beta^{(k+1)}$ by maximizing with respect to β expression (13), that is the complete data likelihood with current conditional expected values $v_{ij}^{(k)}$ inserted for the missing v-data. The maximization is different for $g = 2$ and $g = 3$.

The case $g = 2$. We maximize (13) under the single, trivial restriction $\pi_1 + \pi_2 = 1$. We have no reason to assume $\sigma_1 = \sigma_2$, because the missing genotype is not likely to be the heterozygous one, in particular not under Hardy–Weinberg equilibrium. The explicit solution, given the v-data $v^{(k)}$ from the E-step, is for $i = 1,\ 2$,

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} v_{ij}^{(k)}/n \tag{16}$$

$$\mu_i^{(k+1)} = \sum_{j=1}^{n} v_{ij}^{(k)} w_j/(n\pi_i^{(k+1)}) \tag{17}$$

$$(\sigma_i^2)^{(k+1)} = \sum_{j=1}^{n} v_{ij}^{(k)}(w_j - \mu_i^{(k+1)})^2/(n\pi_i^{(k+1)}) \tag{18}$$

The case $g = 3$. For a mixture of three normal distributions, with three restrictions as described in Section 5, the solution may be derived and expressed by use of the method of Lagrange multipliers. We maximize

$$\begin{aligned} \log L_3^*(\beta; v, w) &= \sum_{i=1}^{3}\sum_{j=1}^{n} v_{ij} \log \pi_i + \sum_{i=1}^{3}\sum_{j=1}^{n} v_{ij} \log f_i(w; \eta_i) \\ &- \lambda_1 \left(\sum_{i=1}^{3} \pi_i - 1 \right) - \lambda_2(\sigma_1^2 - \sigma_3^2) - \lambda_3(\mu_1 + \mu_3 - 2\mu_2) \end{aligned} \tag{19}$$

The complete data estimates of π_i, $i = 1, 2, 3$ are independent of the means and variances, so they are still given by the same formula (16) as for $g = 2$, but now extended to three groups. For given π_i, the complete data estimates of σ_i^2 and μ_i satisfy the following equation system, involving the Lagrange multiplier λ_3:

$$\sigma_1^2 = \sigma_3^2 = \frac{\sum_{j=1}^{n} v_{1j}^{(k)}(w_j - \mu_1)^2 + \sum_{j=1}^{n} v_{3j}^{(k)}(w_j - \mu_3)^2}{n(\pi_1 + \pi_3)} \tag{20}$$

$$\sigma_2^2 = \sum_{j=1}^{n} v_{2j}^{(k)}(w_j - \mu_2)^2/(n\pi_2) \tag{21}$$

$$\mu_i = \frac{\sum_{j=1}^{n} v_{ij}^{(k)} w_j - \lambda_3^{(k+1)}\sigma_i^2}{n\pi_i},\ i{=}1,\ 3 \tag{22}$$

$$\mu_2 = \frac{\sum_{j=1}^{n} v_{2j}^{(k)} w_j - 2\lambda_3^{(k+1)}\sigma_2^2}{n\pi_2}, \tag{23}$$

where λ_3 satisfies

$$\lambda_3 = \frac{\sum v_{1j}^{(k)} w_j/\pi_1 - 2\sum v_{2j}^{(k)} w_j/\pi_2 + \sum v_{3j}^{(k)} w_j/\pi_3}{\sigma_1^2/\pi_1 + 4\sigma_2^2/\pi_2 + \sigma_3^2/\pi_3}. \tag{24}$$

Instead of solving this equation system iteratively, we used formula (17) extended to three groups to obtain explicit formulas (20) and (21) for the variance parameters. In that way we were able to turn the equation system into explicit formulas also for λ_3 and the mean value parameters (22) and (23). In our experience, the differences from the true roots were typically found to be negligible, but alternatively an iteration step could be applied in which the resulting mean values are used for updating the other parameters.

Starting values for the EM algorithm. Starting values for the unknown parameters are needed. These will differ depending on the number of mixture components.

With three equidistant components, natural starting values are obtained by dividing the interval $(\min(w), \max(w))$ in three subintervals of equal length, counting the number of observations in each subinterval, and calculating their mean values and standard deviations (pooling between the two outer intervals). Less data-dependent starting values ($\pi_i = 1/3$, outer means located in min and max points, etc) have also been seen to work well. In all cases of convergence problems, data appeared to have less than three true components.

Starting values for two components are more crucial, in particular if data do not clearly follow the model. They can be selected in a similar way as for $g = 3$, but see also McLachlan and Peel (2000).

Results for one of the SNP data sets. For the purpose of illustration we describe in detail the results for one SNP data set (SNP 2), see Figures 8 and 9 for data and for its 3-component model.

The 3-component EM algorithm converged in just 4 steps and with $g = 2$ it was also fast, around a dozen iterations. The ML estimates and log likelihood maximum values $\log L_g(\hat{\beta})$ are given in Table 3 below, $g = 1, 2, 3$.

Table 3 Estimated parameters and log likelihood functions for SNP 2 for three, two and one normal components (genotypes, g).

g	π_1	π_2	π_3	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_1^2$	$\log L_g(\hat{\beta})$
3	0.25	0.46	0.29	−4.5	−1.4	1.7	0.2	0.1	0.2	−107
2	0.72		0.28	−2.5		1.7	2.6		0.1	−160
1	1			−1.3			5.4			−183

Table 3 shows that the estimated likelihood function with three normal components has a much larger value, ($\log L_3$-max$= -107$), than with fewer components. Hence we can conclude that more than two components are needed. This result agrees with the impression from visual inspection of Figure 8. There are 24 estimated observations in one of the homozygous genotypes (group A), 44 in the heterozygous genotype (group B) and 28 in the other homozygous genotype (group C). The ML solution for $g = 2$ clumps together A and B. A local maximum of $\log L_2$ of the same magnitude clumps together B and C, instead.

The 3-component log-likelihood maximum value ($\log L_3$-max) should always be higher than the 2-component maximum ($\log L_2$-max), since the latter is a special case, but the relation between the deviance (*i.e.* twice $\log L_3$-max minus $\log L_2$-max) and the parameter dimensions for mixture models is not simple, For example, a standard large sample χ^2 test argument is not applicable for testing if the model can be reduced from 3 to 2 components or from 2 to 1 component (McLachlan and Peel, 2000). We require another criterion than the standard deviance criterion for assessing the number of components.

7 Criteria for assessing the number of components

We have developed an algorithm that allows the automatized estimation in parallel of one, two or three normal components (genotypes). It remains to assess the number of components. Such assessment is a frequent and important problem, but difficult and perhaps not yet completely resolved, see McLachlan and Peel (2000). There is no method that works for all real-world data. If we try to fit two components to 3-component data, we will be exposed to problems with local maxima and possible convergence problems with the iteration method. On the other hand, even with good data, one or several genotypes can be missing and the interpretation of data is not necessarily straight forward. Therefore, in ambiguous cases the automatic algorithm is supposed to send a signal that a closer inspection of data is needed.

Instead of using the deviance in this non-standard statistical test situation, we will assess the order (number of components) of the mixture model by using several different information criteria that are penalized forms of the log likelihood functions $\log L_g(\hat{\beta})$, $g = 1, 2, 3$. After trying various information criteria proposed in the literature, we selected the following four criteria for use:

1. Akaike's Information Criterion, *AIC*, in this context proposed by Bozdogan and Sclove (1984)
2. Bayesian Information Criterion, *BIC* (Schwarz, 1978)
3. Classification Likelihood Criterion, *CLC* (Biernacki and Govaert, 1997)
4. Integrated Classification Likelihood Criterion, *ICL* (Biernacki and Celeux, 2000)

For each of them the selected number of components corresponds to the lowest value of the criterion function. Notationally $\log L_g$ will here denote the estimated log likelihood maximum with g normal components, and d will denote the total number of parameters in the model.

The performance of these methods has been evaluated in many mixture model studies. *AIC* tends to overestimate the correct number of components in a mixture model (Koehler and Murphee, 1988). *CLC* works well when the mixing proportions are about equal (Biernacki *et al.*, 1999) but tends to overestimate the correct number of components when there is no restriction on the mixing proportions. In a simulation study by McLachlan and Peel (2000), *ICL* selected

the true number of components in all the simulated data sets and performed better than *AIC*, *BIC* and *CLC*.

Table 4 below shows the results from application of the different information criteria on the same data set as in the previous sections (SNP 2). The table gives the concordant answer that the minimum value of the four information criteria is the mixture with three components, figures given in boldface. We may draw the conclusion that this SNP has three genotypes, as is also clear from visual inspection.

Table 4. Results from five information criteria for SNP 2 for three, two and one normal components (genotypes, g). The number of selected components is shown in bold-face.

g	AIC_g	BIC_g	CLC_g	ICL_g
3	**226**	**241**	**214**	**241**
2	330	342	328	351
1	371	376	367	376

8 Model evaluation and results

At the outset there were data for 75 SNPs. The experiment partially failed for 5 SNPs, which were excluded. The remaining 70 SNP data sets were analysed with our algorithm and the number of mixture components were assessed with the four information criteria.

Experience from application of the algorithm. The algorithm programmed fits g-component models to data for $g = 1, 2, 3$, and for each such model calculates the information criterion functions described above. In some cases the EM algorithm failed to converge, but instead produced a signal saying this. For 62 of the 70 data sets the algorithm gave parameter estimates for all three model types. For two of the other data sets, the 2-component model failed, because data looked like a single sample except for a single outlier. With the outlier as a component of its own, the likelihood goes to infinity when the variance parameter for this component goes to zero. For the remaining six data sets, all seen to be clear-cut examples of 2-component models, the EM algorithm failed for the 3-component model.

Another experience was that the iterations for obtaining estimates for the 3-component model usually converged very fast, typically in less than ten iteration steps. However, when data followed a 1-component model the algorithm required hundreds or many more iteration steps for estimating parameters in the 2- and 3-component models.

Easily interpretable results. As much as 65 of the 70 cases gave concordant results, *i.e.* all four information criteria selected the same number of components:

- **51 cases with three components (genotypes)**,
  e.g. SNP 2 and SNP 62 in Figure 9.1, a and b, respectively.

- **12 cases with two components**,

  e.g. SNP 10 and SNP 56 in Figure 9.1, c and d, respectively.

- **2 cases with one component**,

  SNP 19 and SNP 72 in Figure 9.1, e and f, respectively. These results are reasonable in view of the data in Figure 9.1 and with the other 59 data sets.

Method *ICL* tends to discriminate better than the other three methods. The value for this criterion deviates more between the different models. This has to do with the penalty term that contains both an entropy term and a term involving the number of parameters and the number of observations. Methods *AIC* and *BIC* tend to discriminate less than the other two methods and the *CLC* method gives results somewhere in between, see *e.g.* Table 5 for one of the 1-component sets (SNP 72). If one of these methods is to be chosen we recommend the *ICL* method, in line with McLachlan and Peel (2000).

Table 5. Results from the four information criteria with SNP 72 for three, two and one normal components (genotypes, g). The number of selected components is shown in bold-face.

g	AIC_g	BIC_g	CLC_g	ICL_g
3	114	130	127	154
2	114	127	179	202
1	**109**	**114**	**105**	**114**

Ambiguous results. The results for the five data sets giving conflicting number of components by the four criteria are summarized in Table 6, including a SNP with pathological likelihood. Histograms are presented in Figure 11. Here are some comments.

Table 6. The criterion values for the five data sets, for which conflicting results were obtained, see Figure 11. For each criterion, the number of selected components is marked by a bold-face criterion value.

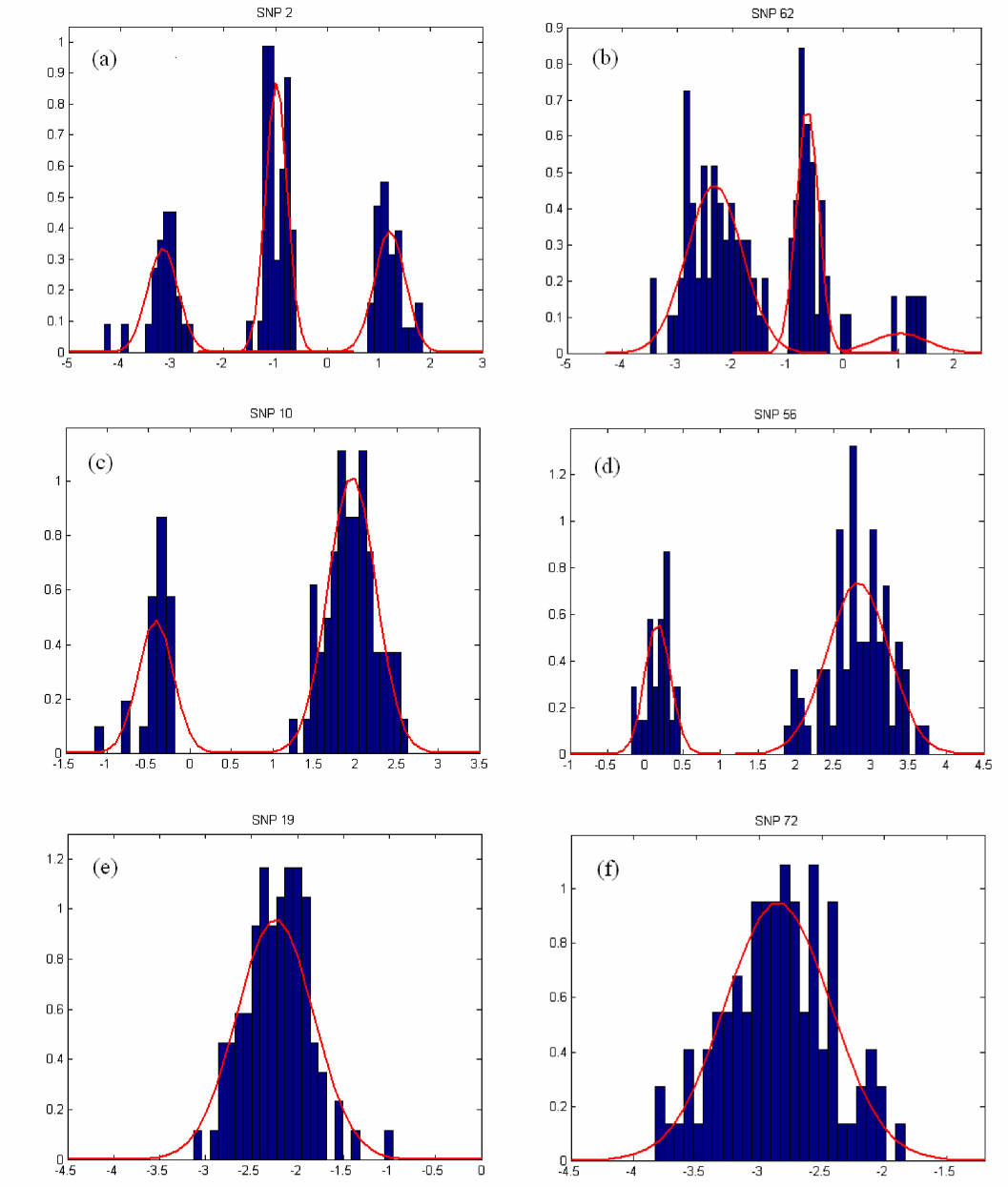

Figure 10: *Histogram of data set SNP 2 (a), SNP 62 (b), SNP 10 (c), SNP 56 (d), SNP 19 (e), SNP 72 (f)with their corresponding density functions.*

SNP	g	*AIC*	*BIC*	*CLC*	*ICL*
11	3	**96**	**111**	154	181
	2	104	117	**112**	135
	1	125	130	121	**130**
29	3	**79**	94	**81**	108
	2	74	**87**	85	**108**
	1	248	253	244	253
30	3	356	371	**350**	377
	2	**356**	**368**	350	**373**
	1	371	376	367	376
52	3	169	185	**179**	206
	2	**168**	**181**	186	209
	1	189	194	185	**194**
60	3	119	135	118	145
	2	$-\infty$	$-\infty$	$-\infty$	$-\infty$
	1	134	139	130	139

For SNP 11, the *AIC* and *BIC* methods choose three components, the *CLC* method prefers $g = 2$, while the *ICL* method prefers a single component. The log likelihood is in fact of the same order of magnitude for the three models ($\log L_3(\hat{\beta}) = -42$, $\log L_2(\hat{\beta}) = -47$ and $\log L_1(\hat{\beta}) = -61$). The penalty term for *CLC* and *ICL* involves the entropy and this tends to reduce the number of components. After looking at Figure 11(a) (SNP 11), we chose to follow the proposal from the *ICL* method that there is only a single component.

For SNP 29, all criteria except CLC select only two components. However, the differences between the criterion values for two and three components are not large for any method. Again the value of the entropy term has an impact. Only a single observation appears to represent one of the homozygous genotypes in this case, see Figure 11(b). Without this observation we would decide that there are two components, but the single observation is in the right place to represent the homozygous alternative. If the small middle group is heterozygous we also have reason to expect none or very few homozygous observations of the other homozygous type (cf. Hardy–Weinberg equilibrium).

For SNP 30, as for SNP 29, CLC chooses three components while the other criteria choose two components. The log likelihood has almost the same value for two and three components ($\log L_3(\hat{\beta}) = -172$, $\log L_2(\hat{\beta}) = -173$ and $\log L_1(\hat{\beta}) = -183$). By inspection of Figure 11(c) we infer that there are obvious difficulties in analyzing this data set. The data are of too low quality.

For SNP 52, *AIC* and *BIC* select two components, *CLC* three components, and the *ICL* method only one component. The data set appears to be of low quality.

For SNP 60, all criteria actually agree, but the minimum likelihood is degenerate. The natural interpretation of data is that there is one component and one additional outlying observation, where this observation perhaps represents another component. In the case $g = 2$, this interpretation corresponds to a singular solution with one component having zero variance. This yields an infinite likelihood, and all criteria have infinite values. Note that in the case of three components with a single outlying observation, such as in Figure 11(b), this situation does not occur, due to the variance restriction $\sigma_1 = \sigma_3$. When we

consider what alternatives to two components the criteria suggest, all criteria except *ICL* choose three components. The *ICL* method selects a single component. Looking at the histogram in Figure 11(e), it is reasonable to select a one- or two-component model to represent the data.

Our overall conclusion from the five cases (a) to (e) is that the *ICL* criterion selects the number of components that is best in line with our visual interpretation of the data.

9 Detection of overall outlying individuals.

Each SNP dataset contains the same $n = 96$ individuals. We may want to know if some individuals yield generally uncertain values and perhaps should be regarded as outliers. Such analyses can be made separately on u- and w-data.

The variate u measures the general intensity of the individual signal. A natural and very simple criterion to identify possibly unreliable individuals would be to look for extremely small values of the average $\bar{u}$ over all SNPs. A more sophisticated version would adjust for heterozygosity, by using an estimated ν-value, via formula (9).

In order to look for individuals with much uncertainty in w we first run the classification algorithm, as described in previous sections, to find the normal components of each SNP dataset. Then we form a quadratic distance measure ψ defined as follows. First consider a single, fixed SNP for the individual j in question and calculate a value of type

$$\psi_j = \sum_{i=1}^{g} \hat{v}_{ij} \frac{(w_j - \hat{\mu}_{ij})^2}{\hat{\sigma}_{ij}^2}, \tag{25}$$

where g is the number of normal components, $\hat{\mu}_i$ and $\hat{\sigma}_{ij}$ are mean and standard deviation, respectively, of the ith component and $\hat{v}_{ij}$ is the posterior probability that individual j with observed value w_j belongs to the ith component of the mixture, for the particular SNP. This is carried out for all K SNPs, resulting in values ψ_{jk}, $k = 1, \ldots, K$. Next we average the individual's ψ-values over all K SNPs, $\psi_j = \sum_k \psi_{jk} / K$.

Figure 12 shows a histogram of the $n = 96$ ψ-values, for the identifiable SNPs. Most individuals have small ψ-values, which means that they are typically located centrally in their normal components, whereas some individuals on average over SNPs (in quadratic mean sense) deviate more than twice the corresponding standard deviation from their component means. Since this is not only for a single SNP, but on the average over a large number of SNPs, it might be regarded as remarkable. The four individuals with highest ψ-values, in size around 6, should be further examined before their classifications are trusted.

10 Identifying the heterozygous group when one group is missing.

When the number of components has been assessed to be $g = 2$, it remains to identify which one (if any) is heterozygous. There are at least three possibilities to be tried, which are discussed below.

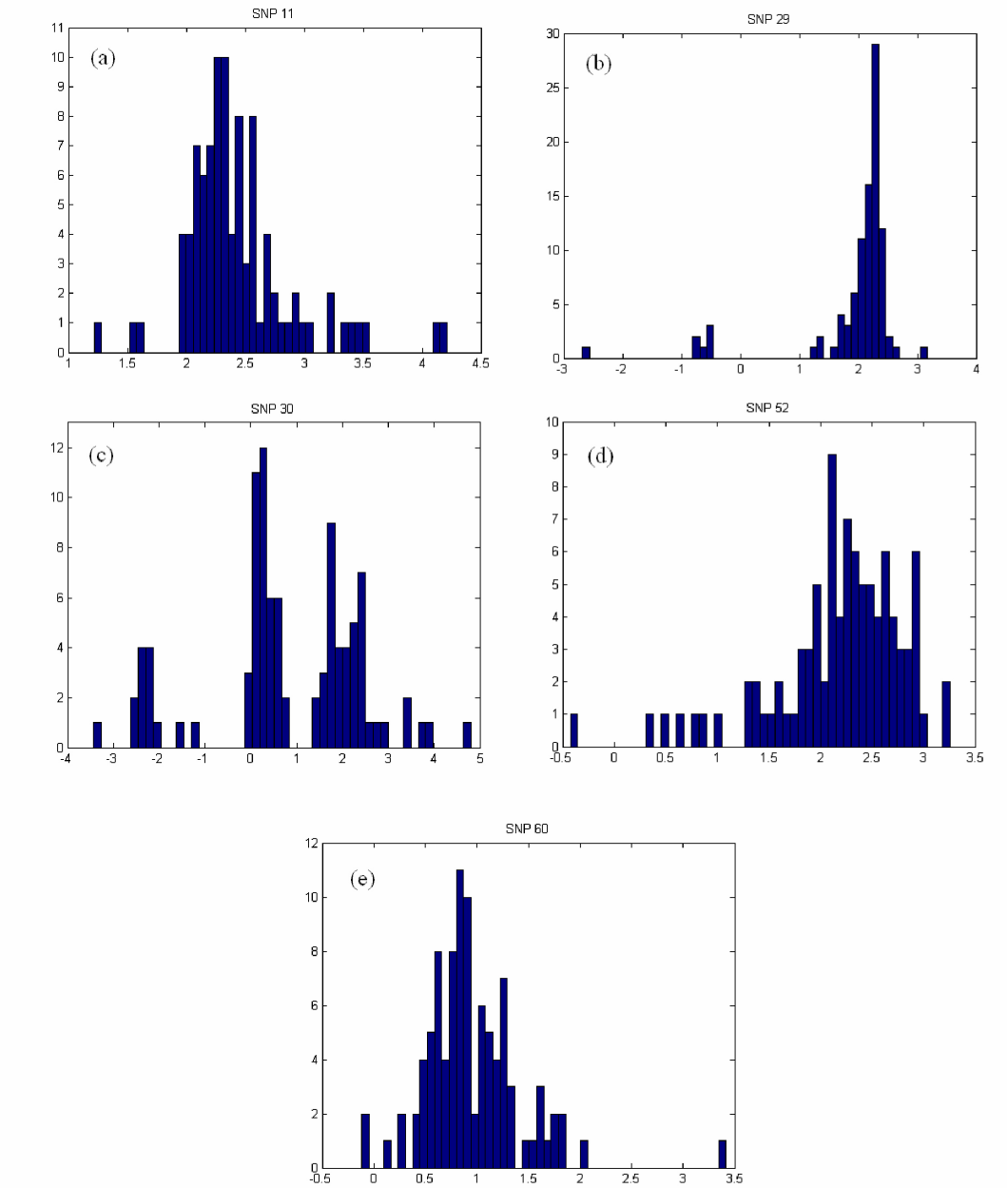

Figure 11: *Histograms for SNPs with non-concordant criteria for the number of components (SNPs 11, 29, 30, 52 and 60).*

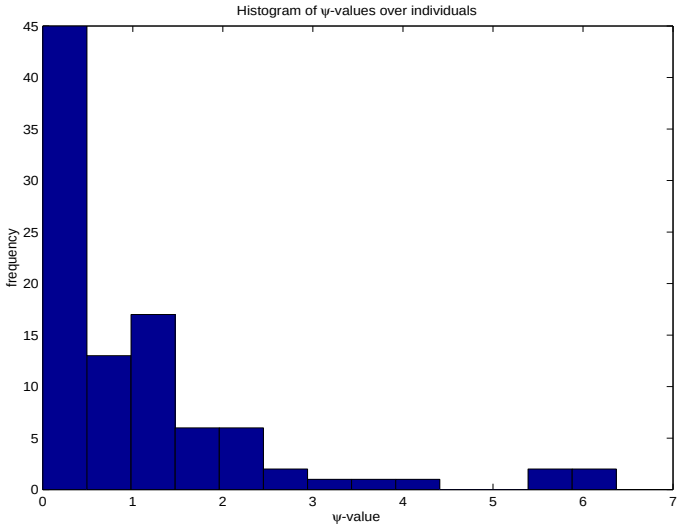

Figure 12: *Histogram of ψ-values, indicating individuals with large w-errors on the average over SNPs.*

10.1 Strong and weak signals

If we could say in advance what signals should be denoted strong, moderate and weak, we should expect (at least) one of the components to have a weak x or y signal, whereas the heterozygous component should not have any weak signal. This will perhaps yield the most conclusive answer in many cases when the clustering is clear. This criterion is not independent of the next one, however.

10.2 Inference from u-data

For identification of the two components, support can be obtained by using u-data. The component with the highest mean value in u is more likely to be heterozygous than homozygous, in particular if a two-sample t-test shows that there is a statistically significant difference between the mean values. However, note that "chip" effects and other variability make the variance in u high and the mean value comparison uncertain, so preferably the comparison between components should adjust for a chip effect in order to be efficient.

10.3 Support from Hardy–Weinberg law.

The genotype distribution (between AA, Aa and aa) of a population may but need not follow the Hardy–Weinberg law of equilibrium. The paper by Teo *et al.* (2007) states that they assume such an equilibrium — a very strong assumption. However, presuming that all individuals have been classified into two or three components (genotypes), we may check if the frequencies are consistent with Hardy–Weinberg equilibrium. For exact tests the genotype membership should be known, but for a crude check it is enough that we have approximate numbers. Assuming that the individuals form a random sample from a population, we can think of using the test results to conclude if the underlying population can be

in Hardy–Weinberg equilibrium. However, less demanding and more reasonable in the present situation is to expect only very approximate Hardy–Weinberg equilibrium, and

- in a 2- or 3-component model check for crude agreement with Hardy–Weinberg equilibrium in the classification inferred from data. Disagreement may but need not indicate errors in the classification.
- in the 2-component model infer which two genotypes are most likely present.

The latter inference is very simple as far as it concerns only estimation: The smaller of the two non-void classes is the heterozygous genotype, and in the variate w it is in the middle between the larger class and the expected position for the void class.

10.4 Illustration on SNP data sets with $g = 2$

First we reconsider the examples from Figure 10(c) and (d). For SNP 10, the larger component has substantially smaller u-values on average, due to much smaller y-values, so all criteria say that the smaller component is heterozygous, For SNP 56, the situation is very similar. One more argument speaks for the smaller component as heterozygous in this case: it has a substantially smaller variance in w than the larger group.

Finally we reconsider SNP 60, see Figure 11(e), with one observation deviating from all others. This single observation is the only one that has a very small x or y, namely $y = 0.7$. Also it has the smallest u-value of all observations, due to this small value. Hence the first two criteria above indicate that all individuals except one constitute a heterozygous component, and one individual is homozygous. However, this is very far from reasonable if there should be any tendency towards Hardy–Weinberg equilibrium. All three criteria taken together therefore indicate that the single observation represents a gross error in its very low y-value,

11 Concluding discussion

We have modelled data from a PrASE type competitive enzymatic assay, and argued theoretically for a particular bivariate Gaussian 3-component mixture model. The model is not a saturated model, but utilizes expected relationships between its component parameters. This is particularly helpful when one genotype is little represented in the sample. The model was found to describe well all typical datasets. Except when one genotype is absent, almost all information for classification is in the marginal univariate model for log-ratio-transformed intensities. For parameter estimation in this mixture model the EM algorithm typically shows a rapid and robust convergence. Our model and algorithm has proven to work well and we have constructed a program for automatic assessment of the number and character of the genotypes. We have used four different criteria for assessment of the number of genotypes present (*AIC*, *BIC*, *CLC*, *ICL)*, of which *ICL* appears to be the most reliable one. The criteria gave discordant answers in only 5 out of 70 SNPs. Such SNPs with unclear answers were left for further manual examination of the data.

Of all 70 SNPs, 12 were classified as lacking exactly one of the three genotypes. In all but one of these cases we could unambiguously characterize one of the two mixture components as heterozygous, by using information orthogonal to the log-ratio of intensities, and by crude comparison with the expected distribution under Hardy–Weinberg equilibrium.

We may draw the following conclusions from our study:

- The model described in this paper is useful for genotyping SNP data, at least for competitive enzymatic assays, but probably also for several other types of SNP data. Carvalho *et al.* (2007) have already used the same model for their analogue to the main response w with Affymetrix data (but without molecular motivation), whereas it has not yet been proposed for other types of data.

- The procedures may easily be implemented for automatic identification of the number of genotypes, using some or all of the criteria *AIC*, *BIC*, *CLC*, and *ICL*.

- The criteria for assessing the number of genotypes (components) may also be used for detecting anomalies in data or in the data structure. When the four criteria give different answers a closer inspection of the data is suggested.

- Other criteria may be used to identify outliers in the set of individuals and to identify the heterozygous group in the case of two groups.

References

Biernacki, C and Celeux, G (2000). Assessing a mixture model for clustering with the integrated completed likelihood, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22** (7), 719–725.

Biernacki, C and Govaert, G (1997). Using the classification likelihood to choose the number of clusters, *Computing Science and Statistics*, **29** (2), 451–457.

Biernacki, C and Celeux, G and Govaert, G (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model, *Pattern Recognition Letters*, **20**, 267–272.

Bozdogan, H and Sclove, SL (1984). Multi-sample cluster analysis using Akaike's information criterion, *Ann. Inst. Statist. Math.*, **36**, 163–180.

Callegaro, A, Spinelli, R, Beltrame, L, Bicciato, S *et al.* (2006). Algorithm for automatic genotype calling of single nucleotide polymorphisms using the full course of TaqMan real-time data. *Nucleic Acids Research*, **34** (7), e56.

Carvalho, B, Bengtsson, H, Speed, TP and Irizarry, RA (2007). Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data, *Biostatistics*, **8** (2), 485–499.

Dempster, A, Laird, N and Rubin, D (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. Royal Statist. Soc., Series B*, **39** (1), 1–38.

Hapmap (2005). A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Hardenbol, P, Yu, F, Belmont, J, Mackenzie, J, *et al.* (2005). Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay *Genome Research*, **15** (2), 269–275.

Hultin, E, Käller, M, Ahmadian, A and Lundeberg, J (2005). Competitive enzymatic reaction to control allele-specific extensions. *Nucleic Acids Research*, **33** (5), e48.

Koehler, A B and Murphee, E H (1988). A comparison of the Akaike and Schwarz criteria for selecting model order. *Applied Statistics*, **37**, 187–195.

Lovmar, L, Ahlford, A, Jonsson, M and Syvänen, A-C (2005). Silhouette scores for assessment of SNP genotype clusters. *BMC Genomics*, **6**: 35.

McLachlan, GJ and Peel, D (2000). *Finite Mixture Models*. Wiley Interscience,

McLachlan, GJ and Thriyambakam, K (1997). *The EM Algorithm and Extensions*. Wiley Interscience,

Moorhead, M, Hardenbol, P, Siddiiqui, F, Falkowski, M *et al.* (2006). Optimal genotype determination in highly multiplexed SNP data. *European J. Human Genetics*, **14**, 207–215.

Pettersson, E, Lindskog, M, Lundeberg, J, and Ahmadian, A (2006). Trinucleotide threading for parallel amplification of minute amounts of genomic DNA. *Nucleic Acids Research*, **34** (6), e49.

Plagnol, V, Cooper, JD, Todd, JA, and Clayton, DG (2007). A method to address differential bias in genotyping in large-scale association studies. *PLoS Genetics*, **3**, 759–767.

Schwarz, G (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Sundberg, R (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.* **1** (2), 49–58.

Teo, YY, Inouye, M, Small, KS, Gwilliam, R *et al.* (2007). A genotype calling algorithm for the Illumina Bead Array platform, *Bioinformatics*, forthcoming.

Xiao, Y, Segal, MR, Yang, YH and Yeh, R-F (2007). A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics*, **23** (12), 1459–1467.