Two approaches to mutation detection based on functional data

Ruth M Pfeiffer[1], Efstathia Bura[2], Amelia Smith[1], Joni L Rutter[1]

[1]National Cancer Institute, Division of Cancer Epidemiology and Genetics, 6120 Executive Blvd, Bethesda, MD, 20892-7244

[2] Department of Statistics, George Washington University, Washington DC, 22052

Summary

A new technique, denaturing high performance liquid chromatography (dHPLC), allows for detection of any heterozygous sequence variation in a gene without prior knowledge of the precise location of the sequence change. The results of a dHPLC analysis are recorded in real-time in the form of a chromatogram that is sequence specific.

In this paper we present methods to classify an individual based on the observed chromatogram as a homozygous wild-type, or a carrier of a specific variant for the given DNA segment by comparison to representative chromatograms that are obtained from the training set of individuals with known variant status. The first approach consists of finding a parsimonious parametric model and then classifying each newly observed curve based on comparing the most discriminating characteristic, the main mode, based the closest fit to the main mode of the training curves. The second approach consists of finding empirical estimates of the modes of each chromatogram, and using a bootstrap to test for equality with the corresponding estimates of the training curves. We apply both methods to data on the breast cancer susceptibility gene *BRCA1*.

Some key words: Functional data; Classification; Nonlinear regression; Superpopulation model; Booststrap; Mutation detection.

1. INTRODUCTION

Historically, DNA polymorphisms and mutations were detected by routine sequencing efforts; however, this approach is costly and slow. A relatively new technique, denaturing high performance liquid chromatography (dHPLC), allows for detection of any heterozygous sequence variation in a gene without prior knowledge of the precise location of the sequence change. In dHPLC analysis, a specific region of DNA ranging from 200 to 700 base-pairs is amplified by PCR and the product is heated to 95 degrees and then slowly cooled to room temperature, allowing the DNA to re-anneal less stringently. If a mutation or polymorphism is present, the annealing will form heteroduplexes (mixture of one wild-type strand with one mutant strand) as well as homoduplexes (annealing of two wild-type strands and annealing of two mutant strands). The heteroduplexes are less stable and will pass through the dHPLC system faster. The results are recorded in real-time in the form of a chromatogram. For details see [1].

The retention times of the heteroduplexes and homoduplexes are sequence specific. They depend on the mismatched base-pairs, the nearest neighbor sequences of the mismatched base-pair [2], and the hydrogen bonding between the non Watson-Crick paired bases [3]. Thus, based on the observed chromatogram a person can be classified as a homozygous wild-type, or a carrier of a specific variant for the given DNA segment. Currently the variant of each chromatogram has to be determined by eye based on comparison to curves from a training sample with known variant status.

In this paper we develop a robust characterization of the chromatograms and classify observed curves by carrier status based on comparison to representative chromatograms that are obtained from the training set of individuals with known variant status. The curve data are described in Section 2. We present two approaches to classification in Sections 3 and 4, respectively: the first approach, motivated by a traditional way to analyze chromatograms (see for example [4]) consists of finding a parsimonious parametric model that captures the main features of the training curves well and then classifying each newly observed curve based on comparing the most discriminating characteristic, the main mode, based the closest fit to the main mode of the training curves. The second approach consists of finding empirical estimates of the modes of each chromatogram, and using a bootstrap to test for equality with the corresponding estimates of the training curves. In Section 5 we apply both methods to data on the breast cancer susceptibility gene *BRCA1*. We conclude in Section 6 with a discussion of our work and future directions of research.

Our work was motivated by ongoing research in the Laboratory of Population Genetics of the National Cancer Institute on genes that influence the risk of breast cancer. Mutations in breast cancer susceptibility genes, *BRCA1* and *BRCA2*, have been shown to predispose women to early onset breast cancer and other malignancies. While three mutations, the so called "founder mutations" have been extensively studied in Ashkenazi Jewish populations, little is known about other variants in those two genes. An ongoing project seeks to complete the analysis of these two genes in a population based series of breast cancer cases from the Radiologic Technologists Health Study. This analysis will yield estimates of mutation prevalence and will allow future analyses of breast cancer risk factors in this cohort to be stratified on *BRCA1/2* status.

2. DATA DESCRIPTION

The data consist of chromatograms each corresponding to a subject whose PCR amplified DNA segment, known as an amplicon, was analyzed via dHPLC, as explained in the introduction. Following an approach discussed in [5], we consider each sample curve to be an independent realization of a stochastic process $Y^{\nu}(t)$, described by the following model,

$$Y^{\nu}(t) = \mu^{\nu}(t) + \epsilon(t), \quad 0 \leq t \leq T. \tag{1}$$

Y denotes the absorbance, t the retention time of the chromatogram, and $\mathrm{E}(Y^{\nu}(t)) = \mu^{\nu}(t)$. The mean μ^{ν} corresponds to the νth variant ($\nu \geq 1$) observed in the analyzed DNA segment. The instrument error is denoted by ϵ. The sample consists of m curves, each a realization of the process $Y^{\nu}(t)$ for some ν, observed at times $t_1, t_2, \ldots, t_{n_i}$, $i = 1, 2, \ldots, m$. The observation $y_{ij} = y_i(t_j)$ denotes the jth point of the ith curve, $i = 1, \ldots m$, $j = 1, \ldots, n_i$. All chromatograms are based on equidistant time points and on the same number of observed points; that is, $n_i = n$ for all $i = 1, 2, \ldots, m$. Our goal is to classify each person in the sample as the carrier of variant ν based on observation of (1) and information available from a training dataset. The total number of variants that can be observed for the analyzed DNA region is not known a priori, as is the case in classical classification analysis, and thus the number of classes obtained from the training data is merely a lower bound. It is conceivable that a person carries a variant that was not observed in the training data. In this case, it would be appropriate to reject the hypothesis that the person belongs to any previously observed class.

The most general approach to classifying the amplicons would be by comparison of their means μ^{ν}, $\nu = 1, 2, \ldots$. This is difficult, not in the least because it is an infinite dimensional problem. Procedures for comparing mean curves rely on the global shape of the curves, see [6], [7] and [8]. These methods are sensitive to contaminations that may influence certain regions of the curve, such as the tails, more heavily than others.

Motivated by the physical process generating the chromatograms, we classify the curves based on more specific features, instead of estimating and testing for the whole mean function μ^ν. The main characteristic of traditional multicomponent chromatograms is a random collection of peaks allocated along the t-axis. Each peak corresponds to a single component and the area under the peak is the proportion of the component in the analyzed compound (see, for example, [9]). Analogously, the dHLPC curve is characterized by a single peak if a person is homozygous for the specific part of the gene and may have up to four peaks for individuals who are heterozygous for a single nucleotide mutation or polymorphism (see [1] for details). The peaks corresponding to the variant occur earlier than the peaks for the wild-type curves, as the heteroduplexes corresponding to the variant DNA are less stable. In principle, one should thus be able to distinguish wild-type and variant carriers by the number of peaks observed in their chromatograms. The drawback of this approach is that the peaks have to be fairly wide apart to be detected. Sequence specific characteristics can cause the heterozygous peaks to be so close together that the result is a unimodal curve, with the location of the mode shifted to the left of the wild-type modes. In consequence, the most discriminating features of the chromatograms are the number of observed modes and their location.

The basic classification procedure we develop is carried out as follows: first, the location and heights of the modes for each μ^ν, are estimated based on curves derived from individuals with known variant status determined by sequencing. Next, the location and heights of the modes are estimated for each newly observed curve and compared to the estimates obtained from the training set. The absolute heights of the modes are subject to variation, as the area under the curve, and thus the amplitude of the modes is proportional to the amount of DNA used in the analysis, which may be difficult to control precisely. Information on the relative height of the modes compared to the highest one, and the location of modes are not much influenced by variations in DNA conentration. The length of the analyzed DNA segments and the fact that variants are relatively rare within a segment imply that the number of different variants that are expected to be seen in each amplicon is small. In particular, in the *BRCA* genes this number does not exceed two or three for most amplicons. Nonetheless, if a new chromatogram cannot be categorized into any of the previously observed classes, sequencing the DNA is recommended to identify the true variant status and perhaps discover a new variant thereby.

3. CLASSIFICATION OF THE CHROMATOGRAMS: A PARAMETRIC APPROACH

3.1. The individual curve model

We first estimate the location of the peaks and their associated heights based on a parametric mixture model that follows an approach presented in [4]. The observed chromatogram (or thermogram in the context of [4]) y is modeled as

$$y^{\nu}(t; \boldsymbol{\omega}^{\nu}) = \phi(t; \boldsymbol{\omega}^{\nu}) + e^{\nu}(t), \tag{2}$$

where $\boldsymbol{\omega}^{\nu}$ is a set of parameters specific to curve category ν. The error term is assumed to be independent across time and normally distributed with mean zero and variance σ^2. The mean function $\phi^{\nu} = \mathrm{E}(Y^{\nu})$ is modeled by a mixture of several curves, each corresponding to a peak in the chromatogram

$$\phi^{\nu}(t;\boldsymbol{\omega}^{\nu}) = \sum_{i=1}^{K^{\nu}} \varphi(t, \boldsymbol{\theta}_i^{\nu}).$$

Several choices for φ are suggested in [4], the linear-Gaussian model, the Gaussian-exponential model, and the Gaussian-Gaussian model. In our problem the Gaussian-Gaussian model and, in a slight modification, the exponential-Gaussian model seem to be the best choices for a peak model based on comparison of the residual sums of squares of the models. We thus parameterize φ as

$$\varphi(t, \boldsymbol{\theta}) = \begin{cases} he^{\frac{-\lambda^2(t-s)^2}{2}} & \text{if } t \leq s \\ he^{\frac{-\gamma^2(t-s)^2}{2}} & \text{otherwise,} \end{cases}$$

for the Gaussian-Gaussian model, or

$$\varphi(t, \boldsymbol{\theta}) = \begin{cases} he^{-\lambda(t-s)} & \text{if } t \leq s \\ he^{\frac{-\gamma^2(t-s)^2}{2}} & \text{otherwise,} \end{cases}$$

for the exponential-Gaussian model with $\boldsymbol{\theta} = (h, \lambda, s, \gamma)$. The parameters have the following interpretation: h denotes the height of the peak, s the location and λ, γ the "thickness" of the ascending and descending parts, respectively. Notice that this model is not a mixture of probability density functions, but a nonlinear regression model. For a fixed K^{ν} we have $\boldsymbol{\omega}^{\nu} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_{K^{\nu}}')'$. The elements of $\boldsymbol{\omega}^{\nu}$ are estimated by minimizing the sum of squares

$$SSE(\boldsymbol{\omega}^{\nu}) = \sum_{j=1}^{n} (y_j - \phi(t_j, \boldsymbol{\omega}^{\nu}))^2. \tag{3}$$

Under the assumption that $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$, the resulting nonlinear least squares estimator $\hat{\boldsymbol{\omega}}$ of $\boldsymbol{\omega}$ is also the MLE and

$$\sqrt{n}(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}) \xrightarrow{\mathcal{D}} N(0, \sigma^2 S^{-1}). \tag{4}$$

S is the uniform asymptotic limit of the matrix

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\partial\phi(t_j;\boldsymbol{\omega})}{\partial\boldsymbol{\omega}}\frac{\partial\phi(t_j;\boldsymbol{\omega})}{\partial\boldsymbol{\omega}'}=\frac{1}{n}\mathbf{F}'(\boldsymbol{\omega})\mathbf{F}(\boldsymbol{\omega}),$$

where $\mathbf{F}(\boldsymbol{\omega})=(F_{ij}(\boldsymbol{\omega}))=(\partial\phi(t_j;\boldsymbol{\omega})/\partial\omega_l)$ is a $n\times 4K^{\nu}$ matrix of partial derivatives. For large n and under appropriate regularity conditions we have approximately following [10]:

$$\hat{\boldsymbol{\omega}}^{\nu}-\boldsymbol{\omega}^{\nu}\sim N(0,\sigma^2(\mathbf{F}'(\boldsymbol{\omega}^{\nu})\mathbf{F}(\boldsymbol{\omega}^{\nu}))^{-1}).$$

The covariance matrix of the $4K^{\nu}$ nonlinear least squares estimates $\hat{\boldsymbol{\omega}}^{\nu}$, $\mathrm{Cov}(\hat{\boldsymbol{\omega}}^{\nu})$ can then be approximated by

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\omega}}^{\nu})=\hat{\sigma}^2(\mathbf{F}'(\hat{\boldsymbol{\omega}}^{\nu})\mathbf{F}(\hat{\boldsymbol{\omega}}^{\nu}))^{-1} \tag{5}$$

where $(\mathbf{F}'(\boldsymbol{\omega}^{\nu})\mathbf{F}(\boldsymbol{\omega}^{\nu}))^{-1}$ is evaluated at $\hat{\boldsymbol{\omega}}$. The unbiased estimate $\hat{\sigma}^2=(n-4K^{\nu})^{-1}\sum(y_j-\phi(t_j;\hat{\boldsymbol{\omega}}^{\nu}))^2$ is used.

To determine the number of components K^{ν} in the nonparametric regression function, we fit a series of models with increasing number of components and select the most plausible model by computing the Bayes Information Criterion (BIC). As we expect at most four peaks in each chromatogram, we restrict our models to $K^{\nu}\leq 4$ components, to avoid overfitting. Alternatively, one could use a likelihood ratio test statistic for model selection (see [4]).

3.2. The superpopulation model

There are two kinds of variability in the chromatograms. The first kind is variability specific to the individual observation y_{ij}. This kind of variability represents the measurement error and is accounted for in model (2) through the error term ϵ. The second kind of variability is related to the whole chromatogram and can be thought to represent curve specific variation that arises from variation in genotyping such as temperaturem DNA concentration, or the presence of contaminants.

To account for chromatogram specific variation we assume that, for a given variant ν, the parameter $\boldsymbol{\omega}_i^\nu$ in model (2) for the ith curve is the realization from a superpopulation model $\boldsymbol{\omega}_i^\nu \sim N(\boldsymbol{\omega}^\nu, \Sigma_0)$, with a variant specific mean parameter $\boldsymbol{\omega}^\nu$ and covariance function Σ_0. Under this model the estimates $\hat{\boldsymbol{\omega}}_i^\nu$ have the conditional distribution $\hat{\boldsymbol{\omega}}_i^\nu | \boldsymbol{\omega}_i^\nu \sim N(\boldsymbol{\omega}_i^\nu, \Sigma_i^\nu)$, where $\Sigma_i^\nu = \sigma^2(\mathbf{F}'(\boldsymbol{\omega}_i^\nu)\mathbf{F}(\boldsymbol{\omega}_i^\nu))^{-1})$. Unconditionally, we get that

$$\hat{\boldsymbol{\omega}_i^\nu} \sim N(\boldsymbol{\omega}^\nu, \Sigma_0 + \Sigma_i^\nu). \tag{6}$$

To estimate the superpopulation parameters $\boldsymbol{\omega}^\nu$ and Σ_0 based on observations $\hat{\boldsymbol{\omega}_i^\nu}, \Sigma_i^\nu, \quad i = 1, \dots, k^\nu$ from the training set, several different approaches such as REML, MLE or empirical Bayes can be considered. We use the sample mean, $\hat{\boldsymbol{\omega}}^\nu = \sum_{i=1}^{k^\nu} \boldsymbol{\omega}_i^\nu / k^\nu$, and estimate $\hat{\Sigma}_0$ by

$$\hat{\Sigma}_0 = \hat{\Sigma}_\omega^{k^\nu} - \frac{1}{k^\nu} \sum_{i=1}^{k^\nu} \Sigma_i^\nu,$$

where $\hat{\Sigma}_\omega^{k^\nu}$ denotes the sample variance of $\hat{\boldsymbol{\omega}}^\nu$ based on all k^ν curves in the training set that fall into class ν.

3.3. The parametric classification algorithm

To classify a new chromatogram, we first fit model (2) to the new curve. Let $\hat{\boldsymbol{\omega}}^\nu$ denote the estimated superpopulation parameters of the training curves for mutation status ν and $\hat{\boldsymbol{\omega}}^{(new)}$ the estimated parameter vector of the new curve. The vector $\hat{\boldsymbol{\omega}}^{(new)}$ also has distribution (6). Since the chromatograms in the training set are independent of the newly observed curve, $\hat{\boldsymbol{\omega}}^\nu$ and $\hat{\boldsymbol{\omega}}^{(new)}$ are also independent.

In many amplicons, such as 1-11G in *BRAC1*, the wild-type curves appear unimodal while the variant, as for example variant one in Figure 1, is clearly bimodal. Figure 2 plots the training curves for amplicon 1-20, *BRCA1*, that include five wild-type curves, and three curves of variant one. Note that for this amplicon the wild-type as well as the variant curves appear bimodal. When we plotted the estimates for the locations for the both modes, we saw that they did not discriminate the curves. This can also be seen by looking at the curves directly: the location of the second largest mode of the variant one curves and the main mode of the wild-type curves are extremly close. When we focused on the location of the main mode, on the other hand, the wild-type chromatograms separated well from the variant curves. We thus identify mutation status by the abscissa of the maximal mode in each curve even when they appear to have more than one mode. The classification algorithm is easily extended to more complex features, which may be more appropriate for other amplicons.

Let $s^{\nu}_{\max}$, based on the superpopulation parameter $\boldsymbol{\omega}^{\nu}$, be the abscissa of the largest mode for mutation status $\nu = 1, 2, \ldots,$ and let $s^{(new)}_{\max}$ be the abscissa of the largest mode of the candidate curve whose mutation status is to be classified. For a chosen parametrization, the estimate of the abscissa of the maximal mode of the chromatogram coincides with the location component $\hat{s}_i$ of $\hat{\boldsymbol{\omega}}$ that has the largest height associated with it. Accordingly, let $\hat{s}^{\nu}_{\max}$ and $\hat{s}^{(new)}_{\max}$ be their respective estimates. The hypotheses of interest are $H^{\nu}_0 : s^{(new)}_{\max} = s^{\nu}_{\max}$ versus $H^{\nu}_1 : s^{(new)}_{\max} \neq s^{\nu}_{\max}$.

Using the independence of the two estimators and (6), a test statistic for testing H^{ν}_0 is

$$Z^{\nu} = \frac{\hat{s}^{\nu}_{\max} - \hat{s}^{(new)}_{\max}}{\sqrt{\hat{\sigma}^2(s^{\nu}_{\max}) + \hat{\sigma}^2(s^{(new)}_{\max})}}. \tag{7}$$

If $s_{\max}^{\nu}$ denotes the lth component of $\boldsymbol{\omega}^{\nu}$, $\hat{\sigma}^2(s_{\max}^{\nu})$ is given by

$$\hat{\sigma}^2(s_{\max}^{\nu}) = \frac{1}{k^{\nu}}(\hat{\Sigma}_0)_{ll} + \frac{1}{(k^{\nu})^2}\sum_{j=1}^{k^{\nu}}(\hat{\Sigma}_i)_{ll},$$

and $\hat{\sigma}^2(s_{\max}^{(new)})$ by

$$\hat{\sigma}^2(s_{\max}^{(new)}) = (\hat{\Sigma}_0)_{ll} + (\hat{\Sigma}_i^{(new)})_{ll}.$$

Under the null hypothesis and given a sufficiently large sample size, Z is asymptotically distributed as a standard normal random variable. The rejection region is defined accordingly. If k^{ν} is small, we approximate the distribution of the test statistic by a $t_{k^{\nu}-1}$ distribution. As we reduce the dimensionality of the parameter space for classification, it is not necessary to estimate all elements of Σ_0, which may be problematic if there are but few curves in the training set.

Our classification algorithm can be summarized in the following steps:

1. Find the best fitting model for each class in the training set and find the estimates of the location of the modes based on the respective model. Get the superpopulation parameter estimates.

2. Fit the parametric model to the newly observed curve and get the estimates of the location of the mode and its variance.

3. Starting with ν =wild-type, compute the test statistic Z^{ν} for H_0^{ν}. If H_0^{ν} cannot be rejected for class ν in the training set, then the new observation is classified into the same category. If H_0^{ν} is rejected, test $H_0^{\nu+1}$.

4. If all H_0^{ν}s have to be rejected, sequence the DNA to determine the true carrier status.

5. Update the training set. Classify the next curve.

In our application the training data were chosen not randomly but in the hope to cover as many variants as possible. For the *BRCA* genes this goal can be achieved by sequencing only breast cancer cases. If the training data are a completely random sample of the population, a Bayesian classification approach could be pursued. A prior probability of being a member of a given class could be estimated from the training data, with a subjectively chosen prior probability of being in none of the classes of the training set. The classification procedure would then consist of assigning the new observation to the class with the largest posterior probability of the mode, given the observed data. Even when the training set constitutes a random sample, determining the prior probability of belonging to a new class can be challenging, for example in populations that are prone to admixture. The "not seen before" class may be quite large as many different variants can be present. For the data we consider, estimation of prior probabilities is not feasible, but we still incorporate some prior knowledge about how likely membership to a given class is by ordering the H_0^νs. As the wild-type is the most common class, we start by testing whether a curve falls into this category. The other hypotheses may be tested in order of frequency of occurrence of the variant as well, so as to reflect some prior knowledge about the probability of class membership.

4. A BOOTSTRAP TEST FOR MUTATION STATUS

In this section we describe a simple approach for classification for the situation that requires only the main mode for discrimination. The variability specific to individual curves is fairly small, resulting in a visual impression of a roughly noise-less curve save for the discontinuities imposed by the discrete, though very fine, time grid. Instead of fitting a parametric model to the observations, or smoothing the curves, we use the raw data to obtain an empirical estimate of the main mode directly from each curve by choosing the t value that has the largest curve reading associated with it. We postulate the following model for the abscissa of the largest mode:

$$s^{\nu}_{\max,i} = s^{\nu}_{\max} + \epsilon_i, \quad i = 1, \dots, k^{\nu},$$

where ν indicates mutation status, k^{ν} denotes the number of curves in class ν in the training set, and ϵ_i is a normal random variable centered at zero with variance σ_0^2. To draw the connection to the superpopulation model from the previous section, estimates $\hat{s}^{\nu}_{\max,i}$ of $s^{\nu}_{\max,i}$ have variance $\sigma_0^2 + \sigma_i^2$, where the σ_0^2 component of the variance accounts for the between curve variation and the term σ_i^2 denotes the within curve variation.

Let $(y^{\nu}_{i1}, y^{\nu}_{i2}, \dots, y^{\nu}_{in})'$, $i = 1, \dots, k^{\nu}$ represent the curves with mutation status $\nu = 1, 2, \dots,$ and let $\hat{s}^{\nu}_{\max,i}$ denote the observed abscissa of the largest mode of the ith curve of mutation status ν. As the time grid becomes finer, i.e., as $n \to \infty$ and the number of the curves of mutation status ν becomes larger, i.e., as $k^{\nu} \to \infty$, $\bar{s}^{\nu}_{\max} = \sum_i \hat{s}^{\nu}_{\max,i}/k^{\nu} \to s^{\nu}_{\max}$.

To test the null $H_0^{\nu} : s^{(new)}_{\max} = s^{\nu}_{\max}$ at level α, we propose to find the $100(1-\alpha)\%$ level confidence interval around the difference $d^{\nu} = s^{\nu}_{\max} - s^{(new)}_{\max}$ under H_0^{ν} and reject H_0^{ν} if $\hat{s}^{(new)}_{\max}$ falls outside this confidence interval. Unfortunately the available number of curves per mutation status in the training set is often quite small, which makes the assumption of normality unreasonable and limits our ability to estimate the variance term. To find an approximate $100(1-\alpha)\%$ confidence interval for d^{ν} a two-step bootstrap approach is utilized:

1. Let k^{ν} denote the number of curves in the training set for class ν. Take a bootstrap

sample of size $k^\nu + 1$ from the curves of class ν in the training set,

$$y^\nu_{i_1\cdot}, y^\nu_{i_2,\cdot}, \dots, y^\nu_{i_{k^\nu}\cdot}, y^\nu_{i_{k^\nu+1}\cdot},$$

where i_j denotes the index of the curve and $\cdot$ stands for all points on this curve. The first k^ν curves represent the bootstrap sample from the training set, and the $(k^\nu+1)$th curve represents the "new" curve under H_0.

2. Resample points from each curve in the bootstrap sample to account for within curve variability, and get the empirical estimate for the mode. That is, take a random sample $y^{\nu\star}_{i1}, y^{\nu\star}_{i2}, \dots, y^{\nu\star}_{in}$ with replacement from $y^\nu_{i1}, y^\nu_{in}, \dots, y^\nu_{in}$. Let $\hat{s}^{\nu\star}_{\max,i}$ be the abscissa of the largest mode of the curve $y^{\nu\star}_{i1}, y^{\nu\star}_{in}, \dots, y^{\nu\star}_{in}$, for $i = 1, \dots, k^\nu+1$. The mean for this bootstrap sample is based on the first k^ν curves, $\bar{s}^{\nu\star}_{\max,b} = \sum \hat{s}^{\nu\star}_{\max,i^\star}/k^\nu$, and the mode of the "new" curve is $\hat{s}^{(new\star)}_{\max,b} = \hat{s}^{\nu\star}_{\max,(k^\nu+1)^\star}$. Calculate $d^\nu_b = \bar{s}^{\nu\star}_{\max,b} - \hat{s}^{(new\star)}_{\max,b}$.

3. Repeat steps 1 and 2 B times.

The bootstrap sample finally yields B estimates of the difference of the mean of the abscissae of the maximal modes in the training set and the abscissa of the maximal mode of the new curve. The α level cut-off point for the test statistic for testing H^ν_0 is then found from the bootstrapped empirical distribution function of the differences. To test $H^\nu_0 : s^{(new)}_{\max} = s^\nu_{\max}$ at level α, find the cut-off point $\hat{u}_{\alpha/2}$ and $\hat{u}_{\alpha/2}$ from the empirical distribution function of the bootstrapped differences $d^\nu_1 = \bar{s}^{\nu\star}_{\max,1} - \hat{s}^{(new\star)}_{\max,1}, \dots, d^\nu_B = \bar{s}^{\nu\star}_{\max,B} - \hat{s}^{(new\star)}_{\max,B}$ such that

$$P(d^\nu_b > \hat{u}_{1-\alpha/2}) = \alpha/2$$

$$P(d^\nu_b \leq \hat{u}_{\alpha/2}) = \alpha/2.$$

To classify a new observation find the location of its mode $s_{\max}^{(new)}$ and see if the observed difference between the mean of the locations of the modes of the training curves in class ν and the new mode location estimate, $\hat{d}^{\nu} = \sum_i \hat{s}^{\nu}_{\max,i}/k^{\nu} - s_{\max}^{(new)}$, falls in the interval $(\hat{u}_{\alpha/2}, \hat{u}_{1-\alpha/2})$. If it does, classify the new curve as belonging to variant status ν. Otherwise check if it falls into the confidence interval for the next class in the training set.

This bootstrap is a fast way of classifying the curves, as the confidence intervals for each class in the training set need to be calculated only once. The classification of a new observation only requires finding its mode.

5. DATA EXAMPLES

In this section we present classification results on amplicon 1-20 and amplicon 1-11G from *BRCA1* based on samples from the Radiologic Technologists Health Study. *BRCA1* consists of 35 amplicons, and *BRCA2* of 47 amplicons, that all have to be analyzed to get full information on the prevalence and number of different variants in those two genes. This is also true if one is interested in determining the carrier status of a single individual. An individual can be classified to be a "variant carrier", if she has a variant in any of the 82 amplicons.

The training set for amplicon 1-20, *BRCA1*, consists of five wild-type curves and two curves corresponding to a variant that we call "variant one". It does not represent a random sample from the population, but is enriched by variant one carriers, and no inference about the prevalence of this variant in the population can be drawn. Each curve for this amplicon is based on 1501 points. The curves of the training set are plotted in Figure 2. The wild-type chromatograms as well as the variant curves appear bimodal, and the location of the highest mode of the variant one curves is to the left of the highest mode of wild-type curves.

First, we use the parametric approach described in Section 3.3 to classify a dataset that contains forty nine new curves for amplicon 1-20. The DNA of every subject in this sample was sequenced as well, to determine the true carrier status. The curves in the training set are best fit with a two component Gaussian-Gaussian model. The residual sums of square for the wild-type curves are $0.7629, 0.8518, 0.5060, 2.6343$ and 1.2147, and 1.3769 and 0.7712 for the variant one curves. Table 1 shows the estimates of the maximal mode for each curve in the training set based the Gaussian-Gaussian two component model as well as the empirical estimate for the mode derived directly from the curve. Both estimates are very close, with the differences ranging from 0.0089 for the second wild-type curve, to -0.0580 for the last variant one curve.

As there are only two curves in the training set that represent variant one, the estimates of the superpopulation variance from these data will be poor. We thus assume that Σ_0 is the same for the wild-type and the variant one curves. This assumption is reasonable, if the variability between curves does not depend on the variant.

All the new curves are best fit with a two component Gaussian-Gaussian model as well. Tables 2 and 2 tabulate the results for the test statistics Z^{wt} and Z^{v1} from (7) for the parametric model. The parametric approach correctly classifies all curves. One curve is classified as a variant one curve, 46 curves are wild-type curves and two curves are classified as "neither wild-type nor variant one". Sequencing those two curves reveals the presence of a new variant, see Table 2. Figure 3 shows one of the variant two curves together with a wild-type and variant one curve for reference.

The values of the test statistic Z^{ν} for the parametric model are quite large when the curve is not a member of class ν. For example, $Z^{v1} = -18.13$ for the first curve. This is also the case when a chromatogram is not a member of any of the classes in the training set, for example $Z^{wt} = 7.97$ and $Z^{v1} = -7.344$ for the last curve in the dataset (Table 2), that is in fact a variant two curve. In our examples, as well as unreported data, there never was an ambiguous situation that would lead to classify a curve as belonging to two different classes. This supports the strategy of classifying the curves as a member of class ν if H_0^ν cannot be rejected without testing all remaining hypothesis.

For the bootstrap classification, $B = 5000$ bootstrap replications were used. The 95% bootstrap confidence interval for $d^\nu = \sum_i \hat{s}^\nu_{\max,i}/k^\nu - s^{(new)}_{\max}$, is $(-0.1156, 0.1283)$ for the wild-type curves and $(-0.0544, 0.0417)$ for the variant one curves. The bootstrap procedure also classifies all curves correctly. The values of $d^\nu = \sum_i \hat{s}^\nu_{\max,i}/k^\nu - s^{(new)}_{\max}$ for each new curve is given in Tables 2 and 3. For the bootstrap as well, the values of d^ν are rather large when the new curve is not a member of class ν.

For the second example we classify curves from amplicon 1-11G. Each curve for this amplicon is based on 2000 points. Four wild-type and two variant one curves were included in the training sets. The training curves are shown in Figure 5. The variant curves are clearly bimodal, with the higher mode on the left side, while the wild-type curves have a shoulder to the left of the higher mode.

We classify forty nine new curves. Again, the true carrier status for these curves is known from sequencing.

For the parametric procedure the curves in the training set are again best fit with a two component Gaussian-Gaussian model. Table 4 contains the estimates of the maximal mode for each curve in the training set based the Gaussian-Gaussian two component model and the coresponding empirical mode estimate. The difference between the estimates is slightly bigger than for amplicon 1-20, ranging from 0.0222 for the fourth wild-type curve to 0.0634 for the first wild-type curve.

Similarly to amplicon 1-20, all the new curves are best fit with a two component Gaussian-Gaussian model. In this example the parametric approach correctly classifies all but one curve. This curve, that is in fact a wild-type curve, is classified as "neither wild-type nor variant one". The value of the test statistic is $Z^{wt} = 3.3126$, while corresponding quantile of the t-distribution $t_{.975,3} = 3.1824$, and $Z^{v1} = -41.7490$. Even though we reject the hypothesis that this curve is a wild-type curve, the evidence for truly observing a new variant is weak.

Fourteen curves are correctly identified as not belonging to any of the classes of the training set, and their sequences determine that they are members of a new class, that we call variant two. Figure 4 shows a plot of one of the chromatograms that falls into the new class, together with a wild-type and variant one curve for reference purposes. One curve in the data set is classified as variant one, all the other curves are correctly categorized as wild-type curves.

For the bootstrap classification, $B = 5000$ bootstrap replications were used. The 95% bootstrap confidence interval for $d^\nu = \sum_i \hat{s}^\nu_{\max,i}/k^\nu - s^{(new)}_{\max}$, is $(-0.0633, 0.0816)$ for the wild-type curves and $(-0.0383, 0.0409)$ for the variant one curves. The bootstrap procedure classifies all curves correctly.

6. DISCUSSION

dHPLC is highly sensitive with sensitivity and specificity approaching 100%, has fast analysis times (less than seven minutes per run, see [1]), does not require post-PCR manipulation, and is favorable for pooling samples, which lowers its already low cost per run. In this paper we develop two methods of classifying dHPLC curves by comparing their most discriminating feature, the location of the main mode, to the same characteristic of curves with known variant status, the training set.

The first approach consists of fitting a nonlinear regression model to the curves. The choice of the nonlinear regression function is motivated by the fact that the main features of chromatograms are the number and location of the peaks of the curve. Estimates of individual curve parameters are obtained via least squares. Between curve variation is accounted for by a hierarchical model, that assumes that the individual curve parameters arise from a superpopulation. Even though we model the whole chromatogram, we further reduce the classification procedure to comparing the location of the maximal mode of a newly observed curve to the location of the mean of the modes of each class in the training data. The maximal mode was the most discriminating feature in all the chromatograms that we observed, even though some amplicons may require more complex features for classification.

For the second approach we estimate the maximal modes empirically. The cut-off values for the test of equality of the modes are found through a bootstrap procedure that accounts for within as well as between curve variability. The complexity of the bootstrap does depend on the dimensionality of the testing problem, though.

By ordering the hypotheses concerning the variants in the training set, we incorporate some prior knowledge about how likely it is that a new curve belongs to a class in the training set. Even if the training set were a random sample of the population it is still somewhat difficult to assign a prior probability to the event that a curve falls into a new class. This can be especially challenging in populations that are prone to admixture, inwhich more variants are expected to be present.

Both methods are used to classify curves for amplicons for *BRCA1* and their performance is compared. In our examples the classification based on only the largest mode works very well, as this feature is the most discriminating one for the given amplicons, even with training sets that only contain two curves for a given variant. Only one curve out of 98 curves was misclassified by the parametric approach as "not in a previously observed class", while in fact it is a wild-type curve. The bootstrap testing procedure classified all curves correctly.

Amplicons for other genes may require more complex measures for discrimination, for example the largest and the second largest mode, which has to be decided after fitting the regression model to the training data. The parametric procedure easily generalizes to this situation. As the whole curve is modeled in the parametric regression approach, only the test statistic has to be adapted to a more general classification algorithm. In principle any function of the parameters can be used for discrimination of the curves. The variance of the test statistic for a function of the parameters can be found by applying the delta method. It is desirable to keep the dimensionality of the testing problem low, as the lower dimensional the testing problem is, the more powerful the test will be. It is thus recommended to find the most parsimonious characterization that discriminates the curves well. While the parametric procedure requires not much further work to be adapted to a more complex procedure, the empirical approach becomes more involved, as different features of the curve have to be estimated separately from the curves. The bootstrap becomes more involved in this situation, thus slower and perhaps less appealing that the parametric approach.

The methods presented in this paper apply to chromatograms based on individual DNA samples. dHPLC may also be used to analyze pools of DNA. Future work will include developing methods to determine if a curve that is based on pooled DNA from several individuals contains one or more subjects who carry a variant and to classify the variant present in the pool.

Acknowledgments: We thank the investigators of the Radiologic Technologists Health Study at the National Cancer Institute for access to the data, and Mitchell Gail and Joseph Gastwirth for many helpful comments.

REFERENCES

[1] Kuklin A, Munson K, Gjerde D, Haefele R, Taylor P. Detection of single-nucleotide polymorphisms with the WAVE DNA fragment analysis system. *Genetic Testing* 1997; **1** 201–206.

[2] Ke SH, Wartell RM. Influence of nearest neighbor sequence on the stability of base pair mismatches in long DNA; determination by temperature-gradient gel electrophoresis. *Nucleic Acids Research* 1993; **21** 5137–43.

[3] Aboul-ela F, Koh D, Tinoco I, Martin FH. Base-base mismatches. Thermodynamics of double helix formation for dCA3XA3G+dCT3YT3G (X,Y=A.G.C.T). *Nucleic Acids Research* 1985; **13** 4811–4824.

[4] Robin S. A Model for Thermograms. *Biometrics* 1999; **55** 37–43.

[5] Rice JA, Silverman BW. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society B* 1991; **53**: (1) 233–243.

[6] Hall P, Hart JD. Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association* 1990; **85** 1039–1049.

[7] King E, Hart JD, Wehrly TE. Testing the equality of two regression curves using linear smoothers. *Statistics and Probability Letters* 1991; **12** 239–247.

[8] Fan J, Lin SK. Tests of significance when the data are curves. *Journal of the American Statistical Association* 1998; **93**: 1007–1021.

[9] Pietrogrande MC, Cavazzini A, Dondi F. Quantitative theory of the statistical degree of peak overlapping in chromatography *Reviews in Analytical Chemistry* 2000; **19** (2): 123–155.

[10] Seber GAF, Wild CJ. *Nonlinear Regression*. Wiley, New York, 1989.

Table 1: Estimates of the location of the main mode for 1-11G

Variant	parametric model	empirical estimate
wild-type	3.3025	3.2883
wild-type	3.2206	3.2117
wild-type	3.2800	3.2683
wild-type	3.3977	3.3717
wild-type	3.4036	3.3800
variant 1	2.8872	2.8300
variant 1	2.8187	2.8767

Table 2: Classification results for 1-20

true status	class(par)	class(boot)	Z^{wt}	Z^{v1}	d^{wt}	d^{v1}
wt	wt	wt	0.4526	-18.1367	0.0123	-0.4461
wt	wt	wt	0.6374	-17.8755	0.0140	-0.4444
wt	wt	wt	1.1947	-17.1118	0.0173	-0.4411
wt	wt	wt	1.2619	-17.0178	0.0257	-0.4327
wt	wt	wt	0.9270	-17.4804	0.0123	-0.4461
wt	wt	wt	0.4365	-18.1556	0.0223	-0.4361
wt	wt	wt	-0.2537	-19.1087	-0.0143	-0.4727
wt	wt	wt	1.6154	-16.5094	0.0440	-0.4144
wt	wt	wt	-0.5685	-19.5340	-0.0410	-0.4994
wt	wt	wt	0.5533	-17.9894	0.0090	-0.4494
wt	wt	wt	-0.0084	-18.7721	-0.0060	-0.4644
wt	wt	wt	-0.9676	-20.0922	-0.0277	-0.4861
wt	wt	wt	-1.3710	-20.6530	-0.0643	-0.5227
wt	wt	wt	-1.3628	-20.6232	-0.0393	-0.4977
wt	wt	wt	-0.3728	-19.2732	-0.0327	-0.4911
wt	wt	wt	-1.5803	-20.9355	-0.0477	-0.5061
wt	wt	wt	-0.8885	-19.9850	-0.0693	-0.5277
wt	wt	wt	-0.7486	-19.7914	-0.0293	-0.4877
wt	wt	wt	-1.1017	-20.2794	-0.0677	-0.5261
wt	wt	wt	-1.3844	-20.6499	-0.0743	-0.5327
wt	wt	wt	-0.6984	-19.7176	-0.0410	-0.4994
wt	wt	wt	-1.7905	-21.2184	-0.0193	-0.4777
wt	wt	wt	-1.9644	-21.4620	-0.0427	-0.5011
wt	wt	wt	-0.9658	-20.0881	-0.0577	-0.5161
wt	wt	wt	-0.7293	-19.7663	-0.0460	-0.5044
wt	wt	wt	-1.9068	-21.3874	-0.0710	-0.5294
v1	v1	v1	12.3074	-1.7806	0.4273	-0.0311
wt	wt	wt	-0.4972	-19.4452	-0.0427	-0.5011
wt	wt	wt	-0.9731	-20.0996	-0.0577	-0.5161
wt	wt	wt	-1.3281	-20.5732	-0.0577	-0.5161
wt	wt	wt	-0.9980	-20.1296	-0.0460	-0.5044

Table 2: Classification results for 1-20, continued

true status	class(par)	class(boot)	Z^{wt}	Z^{v1}	d^{wt}	d^{v1}
wt	wt	wt	-1.4942	-20.8188	-0.0710	-0.5294
wt	wt	wt	-0.4211	-19.3357	-0.0193	-0.4777
wt	wt	wt	-1.1530	-20.3479	-0.0543	-0.5127
wt	wt	wt	-0.3288	-19.2071	-0.0110	-0.4694
wt	wt	wt	-0.9557	-20.0716	-0.0343	-0.4927
wt	wt	wt	-0.3824	-19.2801	-0.0127	-0.4711
n	n	n	7.9605	-7.3443	0.2207	-0.2377
wt	wt	wt	-1.0016	-20.1370	-0.0343	-0.4927
wt	wt	wt	-1.5132	-20.8433	-0.0643	-0.5227
wt	wt	wt	-2.4054	-22.0691	-0.0877	-0.5461
wt	wt	wt	-2.4660	-22.1467	-0.0943	-0.5527
wt	wt	wt	-1.7992	-21.2372	-0.0627	-0.5211
wt	wt	wt	-1.2669	-20.5001	-0.0460	-0.5044
wt	wt	wt	-1.6169	-20.9696	-0.0660	-0.5244
wt	wt	wt	-1.5693	-20.9053	-0.0643	-0.5227
wt	wt	wt	-1.7431	-21.1315	-0.0677	-0.5261
wt	wt	wt	-1.8590	-21.2876	-0.0760	-0.5344
n	n	n	7.9605	-7.3443	0.2207	-0.2377

n stands for "not classified"
* denotes curves in the training set
** marks the misclassified curve

Table 3: Estimates of the location of the main mode for 1-11G

Variant	parametric model	empirical estimate
wild-type	7.4751	7.4117
wild-type	7.4593	7.4267
wild-type	7.4678	7.4350
wild-type	7.3722	7.3500
variant 1	5.9352	5.8850
variant 1	5.8946	5.8517