SEMIPARAMETRIC DENSITY ESTIMATION WITH APPLICATIONS TO DNA PROFILING

R. Cao[1], A. Carracedo[2] and E. Valverde[3]
[1]Mathematics Department, Faculty of Computer Science, 15071 La Coruña, Spain.
[2]Institute of Legal Medicine, Faculty of Medicine, 15705 Santiago de Compostela, Spain.
[3]Department of Molecular Biology, Pharmagen Corp., Calera 3, 28760 Madrid, Spain.

1 Introduction

The development of new methods for deeper statistical inference concerning VNTR systems in DNA profiling is the main purpose of this study. Most of the already existing methods treat the observed fragment size as a continuous variable (this is unavoidable since there is a continuous error in the measurement) but they discretize back the distribution of the fragment sizes (see for instance, Gill et al. (1991), Berry (1991) or Eriksen, Bertelsen and Svensmark (1992)).

The idea behind the study we present here is to handle the continuous distribution in a proper manner to answer the questions we are willing to (as estimation of the distribution of the true fragment size or probability of match, for instance). This is done without any parametric assumption, except on the part concerning the error, which is assumed to be normal. In other words, a semiparametric model for the probability density function is stated, estimated and finally used to find answers to questions like the ones given above.

2 The statistical model

Let us denote by X the variable of interest, namely the true number of Kbp of the fragment size. Of course, we can approximate this variable (which, in fact, is discrete) by a continuous version of it with probability density function denoted by f.

The most common approach to the problem of computing the true value of X, consists of running an electrophoretic experiment in which a migrated distance, W, is observed. In the basis of this observation, an approximation, Y, to the true value of X, is computed (typically $W = \log Y$, up to a certain constant).

If we denote by Z the *theoretical* migrated distance -this means the migrated distance that should be observed under ideal conditions and with no measurement errors-, we allow a general logarithmic-type relationship between X and Z:

$$Z = \log g(X),$$

where g is a link smooth function (without any parametric assumption on it) which enables to incorporate not only linear dependence between X and Y.

Taking into account all the relationships stated above, we end up with the following semiparametric model:

$$Y = g(X)e^{\varepsilon}, \tag{1}$$

where ε is a random error term, assumed to be normal with variance σ^2 and independent of the true fragment length.

Two important features of the model are that the conditional mean and variance of the observed fragment length, given the true fragment length, are linear and quadratic functions (respectively) of the link function $g(x)$:

$$m(x) = E\,(Y|_{X=x}) = g(x)e^{\sigma^2/2} \tag{2}$$

$$v(x) = Var\,(Y|_{X=x}) = m(x)^2(e^{\sigma^2} - 1). \tag{3}$$

In the particular case of g being linear, the conditional mean and standard deviation of the observed fragment length grow linearly with the true fragment length. This was really an evidence in our practical experimental databases.

3 Estimation of the density function

The most natural graphical way of representing the ditribution of a continuous random variable is to plot its density function. In our case, we want to estimate f, the density of X. Unfortunately, we face some problems to do this:

1. No data for the variable of interest, X, are avaliable.
2. If we try to solve the previous point by using a sample of Y, we see in (1) that we need some information about g.
3. Even in the ideal case of g being known, we ignore the quantity σ -a measurement of the experimental error-, and we have to estimate it.

It is clear then that we have to carry out two estimations: estimation of σ^2 and functional estimation of g.

To do this, it is usefull to realize of the fact that, expression (3) implies:

$$Var(Y/m(X)) = e^{\sigma^2} - 1.$$

By using a nonparametric estimator, $\hat{m}$, of the regression function m (see for instance the kernel estimator of Nadaraya(1964) and Watson(1964)) and replacing the expectation by the empirical mean, we only need to deal with a preliminary sample of both variables X and Y (this means a control database for which the true fragment lengths are known) to estimate the standard deviation σ.

Once this step is performed, the function g is estimated, in view of (2), by rescaling the nonparametric regression estimator:

$$\hat{g}(x) = \frac{\hat{m}(x)}{e^{\hat{\sigma}^2/2}}.$$

With the observed sample of fragment lengths $(Y_1, Y_2, \ldots, Y_n)$, the kernel estimator (see Parzen (1962)) can be used to construct a nonparametric estimator of the density of W:

$$\hat{f}_h^W(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - W_i}{h}\right),$$

where $W_i = \log Y_i$, K is the kernel function and h is the bandwidth parameter.

Using this nonparametric density estimator and the convolution structure of W, we end up with a deconvolution density estimator of the density of the random variable Z. For the particular choice of gaussian kernel K, this deconvolution estimator coincides with the previous one but with a slightly small bandwidth: $\sqrt{h^2 - \hat{\sigma}^2}$.

Finally, a simple change of variable leads to the density estimator we wanted to find:

$$\hat{f}^X(x) = \hat{f}^W_{\sqrt{h^2-\hat{\sigma}^2}}(\log \hat{g}(x)) \frac{\hat{g}'(x)}{\hat{g}(x)}.$$

The important problem of choosing the smoothing parameter h was addressed, in practice, by using the smoothed cross-validation selector (see Hall, Marron and Park (1992)), while the computational matters were optimized by means of the fast Fourier transform algorithm for density estimation.

4 Probability of match

Denote by M the event that there is a one-allele match between two fragments and denote by $w_1 = \log y_1$ and $w_2 = \log y_2$, where y_1 and y_2 are the two observed fragment lengths. In such a situation the continuous version of Bayes Rule applies to find an estimator of the posterior probability of M given the evidences (y_1 and y_2):

$$\hat{p} = \hat{P}(M|_{y_1,y_2}) = \frac{\hat{g}(w_1, w_2|_M)P(M)}{\hat{g}(w_1, w_2|_M)P(M) + \hat{g}(w_1, w_2|_{\overline{M}})P(\overline{M})}.$$

where $P(M)$ and $P(\overline{M})$ are the *priori* probabilities and the rest of the terms in the ratio admit the following expression, when the gaussian kernel is used,

$$\hat{g}(w_1, w_2|_{\overline{M}}) = \hat{f}_h^W(w_1)\hat{f}_h^W(w_2) = \frac{1}{n^2h^2 2\pi}\left[\sum_{i=1}^{n}\exp(-(w_1 - W_i)^2/2h^2)\right]\left[\sum_{i=1}^{n}\exp(-(w_2 - W_i)^2/2h^2)\right]$$

$$\hat{g}(w_1, w_2|_M) = \frac{1}{n2\pi\hat{\sigma}\sqrt{2h^2 - \hat{\sigma}^2}}\sum_{i=1}^{n}\exp\left(\frac{-h^2(w_1 - w_2)^2 + 2\hat{\sigma}^2(W_i(w_1 + w_2) - w_1w_2 - W_i^2)}{2\hat{\sigma}^2(2h^2 - \hat{\sigma}^2)}\right)$$

5 Conclusions

a) The database used to implement the model show that the function g can clearly be assumed to be linear. Further more we can accept that $g(x) = x$.
b) The semiparametric model is flexible enough to admit a wide class of density functions for the true fragment length. However, the assumptions on the model imply that the conditional standard deviation of the observed length (measured by migration) is proportional to the actual fragment length.
c) The computational algorithm used is complex but may be run very fast in a PC by using sofisticated techniques (as the fast Fourier transform algorithm for density estimation).
d) The method outperforms the majority of the existing procedures, since it makes no parametric assumptions on the marginal distribution of the fragment sizes and does not discretize the continuous estimator of the migrated distance.
e) The calculations done above for the probability of single-allele match could be extended to the pairwise match just by formulating a new model with the same structure but incorporating some correlation between the error terms (ε) of both alleles in the pair.

6 References

Berry, D.A. (1991) Inferences using DNA profiling in forensic identification and paternity cases. *Statistical Science,* vol. 6, 2, 175-205.
Eriksen, B., Bertelsen, A. and Svensmark, O. (1992). Statistical analysis of the measurement errors in the determination of fragment length in DNA-RFPL analysis. *Forensic Science International,* 52, 181-191.
Gill, P., Evett, I.W., Woodroffe, S., Lygo, J.E., Millican, E. and Webster, M. (1991). Database, quality control and interpretation of DNA profiling in the Home Office Forensic Science Service. *Electrophoresis,* 12, 204-209.
Hall, P., Marron, J.S. and Park, B. (1992). Smoothed cross validation. *Probability Theory and Related Fields,* 92, 1-20.
Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications,* 9, 141-142.
Parzen, E. (1962). On estimating a probability density function and mode. *The Annals of Mathematical Statistics,* 33, 1065-1070.
Watson, G.S. (1964). Smooth regression analysis. *Sankhyā,* Series A, 26, 359-372