The Evaluation of Evidence for Autocorrelated Data: with an Example Relating to Traces of Cocaine on Banknotes

Amy Wilson[1], Colin Aitken[1], Richard Sleeman[2], Jim Carter[3]
[1] School of Mathematics and Maxwell Institute, University of Edinburgh, Mayfield Road, Edinburgh
[2] MSA Ltd., Building 20F, Golf Course Lane, P.O. Box 77, Filton Road, Bristol
[3] Queensland Health Forensic and Scientific Service, 39 Kessels Road, Coopers Plains, Australia
Corresponding author: Amy Wilson, email: A.L.Wilson-4@sms.ed.ac.uk

Abstract

Much research in recent years for evidence evaluation in forensic science has focussed on methods for determining the likelihood ratio in various scenarios. The likelihood of the evidence is calculated under each of two propositions, that proposed by the prosecution and that proposed by the defence. The value of the evidence is given by the ratio of the likelihoods associated with these two propositions. The aim of this research is to evaluate this likelihood ratio under two scenarios. The first is when the evidence consists of continuous autocorrelated data. The second, an extension to this, is when the observed data are also believed to be driven by an underlying latent Markov chain. Four models have been developed to take these attributes into account: an autoregressive model of order one, a hidden Markov model with autocorrelation between adjacent data points and a nonparametric model with two different bandwidth selection methods. Application of these methods will be illustrated with an example where the data relate to traces of cocaine on banknotes as measured by the log peak area for the ion count for cocaine product ion *m/z* 105 in a mass spectrometer. The likelihood ratios using these four models will be calculated for these data, and the results compared.

Keywords: Forensic evidence evaluation, Hidden Markov model, likelihood ratio.

1 Introduction

Denote a set of evidential data by $\mathbf{z} = (z_1, z_2, \ldots z_n)$, and two propositions by H_C and H_B. Two scenarios will be considered. The first is when the data $\mathbf{z}$ are autocorrelated, and the second is when, in addition, the data are also driven by an underlying latent Markov chain. The aim of this paper is to evaluate the likelihood ratio given by $f(\mathbf{z} \mid H_C)/f(\mathbf{z} \mid H_B)$ for these two scenarios.

In the context of evidence evaluation, the likelihood ratio was introduced in Lindley (1977) and is used to assign a value to evidence. If we let the prosecution proposition be H_C and the defence proposition be H_B than a likelihood ratio larger than one would imply that the evidence is more likely under the prosecution proposition, and a likelihood ratio of less than one would imply that the evidence is more likely under the defence proposition. By combining the likelihood ratio with the prior on the two propositions (which should come from the decision maker, such as the jury or the judge), posterior odds of the probability of the prosecution proposition over the defence proposition, given the evidence, can be obtained. This stems from the odds form of Bayes' theorem:

$$\frac{f(H_C \mid \mathbf{z})}{f(H_B \mid \mathbf{z})} = \frac{f(\mathbf{z} \mid H_C)}{f(\mathbf{z} \mid H_B)} \times \frac{f(H_C)}{f(H_B)}$$

Four models are described here to estimate the function f in the likelihood ratio. No assumption of independence is made between adjacent datapoints. The first model is an autoregressive model with lag one. The second is a hidden Markov model, and the third and fourth are non-parametric models, each with a different bandwidth. The models will be tested using data relating to traces of the cocaine product ion *m/z 105* on sets of banknotes. The two propositions will be, H_C, that the banknotes are associated with crime involving cocaine, and H_B, that the banknotes are from general circulation. The estimation of f will be done with reference to two sets of training data: banknotes associated with crime involving cocaine and banknotes from general circulation.

2 Models

A crime has been committed. Part of the evidence is a sample of autocorrelated data $\mathbf{z} = (z_1, \ldots z_n)$. To calculate the likelihood ratio for this data, the parameters used in the function f in the likelihood ratio must be estimated for two sets of data: that associated with H_C and that associated with H_B. The set of data associated with H_C is denoted by $\mathbf{y} = (y_{ij}, i = 1, \ldots m_C, j = 1, \ldots n_{C_i})$. In this set, there are m_C different autocorrelated samples, each containing n_{C_i} datapoints. The set of data associated with H_B is denoted by $\mathbf{x} = (x_{ij}, i = 1, \ldots m_B, j = 1, \ldots n_{B_i})$, with m_B different autocorrelated samples, each containing n_{B_i} datapoints. In the following sections, models are described for a general sample of data $\mathbf{w} = (w_1, \ldots w_{n_D})$. To apply the models, $\mathbf{w}$ should be replaced by $\mathbf{x_i}$ and D by B_i if the parameters for the model for sample i conditional on the proposition H_B are being estimated, and $\mathbf{w}$ should be replaced by $\mathbf{y_i}$ and D by C_i if the parameters for the model for sample i, conditional on the proposition H_C are being estimated.

2.1 Autoregressive Model

An autoregressive model $AR(1)$ specifies the following relationship amongst the variables:

$$w_t - \mu = \alpha\,(w_{t-1} - \mu) + \epsilon_t$$

where $t = 2, \ldots, n_D;\ \epsilon_t \sim N(0, \sigma^2)$ and $w_1 \sim N(\mu, \sigma^2)$.

The parameters required to estimate the function f in the likelihood ratio for each of the two propositions are therefore $\theta = (\mu, \sigma, \alpha)$. The likelihood of the training data can be used in conjunction with prior distributions to determine posterior distributions for the model parameters. Posterior distributions can be obtained for each of the n_{B_i} samples in $\mathbf{x}$ (denote these parameters by θ_{B_i}) and for each of the n_{C_i} samples in $\mathbf{y}$ (denote these parameters by θ_{C_i}).

2.2 Hidden Markov Model

In a hidden Markov model, each observed data point is associated with an unobserved state. The states form a Markov chain and determine the probability density function of the data point. A Markov switching model is used here, which also allows for dependence between adjacent datapoints, so that autocorrelation between adjacent datapoints is modelled.

Denote the latent states of the training data associated with H_C by $\mathbf{S_C} = \{S_{C_{ij}}; i = 1, \ldots m_C, j = 1, \ldots n_{C_i}\}$. Each datapoint y_{ij} in $\mathbf{y}$ is associated with a latent state $S_{C_{ij}}$. Similarly, denote the latent states of the training data associated with H_B by $\mathbf{S_B} = \{S_{B_{ij}}; i = 1, \ldots m_B, j = 1, \ldots n_{B_i}\}$, so that each datapoint x_{ij} in $\mathbf{x}$ is associated with a latent state

$S_{B_{ij}}$. Four states are used here, so $S_{C_{ij}}$ and $S_{B_{ij}}$ can take values in $[1, 2, 3, 4]$, but this model could be extended to allow a different number of states. The states associated with the general sample of data $\mathbf{w}$ are denoted $S_D = (S_1, \ldots S_{n_D})$. The states are used to allow for two different mean and variance levels. Four states are required so that the mean level of the previous datapoint is also encoded in the state of the current datapoint. Let 0 and 1 denote the two sets of mean and variance levels. The four hidden states of the model are defined as, in the format (level of previous datapoint, level of current datapoint): state 1 (0,0), state 2 (0,1), state 3 (1,0), and state 4 (1,1). The transition matrix, giving the probabilities of moving between these states, is:

$$\mathbf{P} = \begin{pmatrix} 1-p_{01} & p_{01} & 0 & 0 \\ 0 & 0 & p_{10} & 1-p_{10} \\ 1-p_{01} & p_{01} & 0 & 0 \\ 0 & 0 & p_{10} & 1-p_{10} \end{pmatrix}$$

It is assumed that $\mathbf{w}$ come from a hidden Markov model given by:

$$w_t - \mu_{S_t} = \alpha(w_{t-1} - \mu_{S_{t-1}}) + \epsilon_{S_t}$$

where $\epsilon_{S_t} \sim N(0, \sigma^2_{S_t})$ for $t \in (1, 2, \ldots n_D)$, $w_1 \sim N(\mu_{S_1}, \sigma^2_{S_1})$ and the subscript S_t indicates that the parameter value for the current datapoint of state S_t should be used.

The parameters required are therefore $\theta = (\mu_0, \mu_1, \sigma^2_0, \sigma^2_1, \alpha, p_{01}, p_{10})$. The likelihood of the training data can be used in conjunction with prior distributions to determine posterior distributions for the model parameters conditional on each of the two propositions. Posterior distributions can be obtained for each of the n_{B_i} samples in $\mathbf{x}$ (denote these parameters by θ_{B_i}) and for each of the n_{C_i} samples in $\mathbf{y}$ (denote these parameters by θ_{C_i}). A Metropolis Hastings sampler can be used to obtain these posterior distributions. The calculation of the likelihood, which is required in the Metropolis Hastings sampler, can be done using the forward algorithm, as discussed in Rabiner (1989).

2.3 Nonparametric Model

The parametric models assume a Normal distribution for the error terms, an assumption which is dispensed with for the nonparametric models. As before, a general notation is used. To distinguish between different samples, the notation is now $\mathbf{w}_i = (w_{i1}, \ldots, w_{n_{D_i}}); i = 1, \ldots, m_D$ where m_D is the number of samples in the set and n_{D_i} is the number of datapoints in the i-th sample. The joint density function of $\mathbf{w}_i$ may be written as:

$$f_{D_i}(w_{i1}, w_{i2}, \ldots w_{in_{D_i}}) \quad = \quad f_{D_i}(w_{i1}) f_{D_i}(w_{i2}|w_{i1}) \ldots f_{D_i}(w_{in_{D_i}}|w_{i,n_{D_i}-1})$$

allowing for autocorrelation of lag one. The conditional density function $f_{D_i}(w_{it} \mid w_{i,t-1})$ for each $i \in (1, 2, \ldots m_D)$ can be estimated nonparametrically by:

$$\hat{f}_{D_i}(w_{it}|w_{i,t-1}) = \frac{\hat{g}_{D_i}(w_{it}, w_{i,t-1})}{\hat{r}_{D_i}(w_{i,t-1})}. \tag{1}$$

The functions $\hat{g}_{D_i}$ and $\hat{r}_{D_i}$ are kernel density estimates for sample i, given by:

$$\hat{g}_{D_i}(w_{it}, w_{i,t-1}) = \frac{1}{(n_i - 1)h_1 h_2} \sum_{j=2}^{j=n_i} K_1\left(\frac{w_{it} - w_{ij}}{h_1}\right) K_2\left(\frac{w_{i,t-1} - w_{i,j-1}}{h_2}\right)$$

and

$$\hat{r}_{D_i}(w_{i,t-1}) = \frac{1}{(n_i-1)h_3}\sum_{j=2}^{j=n_i} K_3\left(\frac{w_{i,t-1}-w_{i,j-1}}{h_3}\right).$$

Here, h_1, h_2 and h_3 are bandwidths, and K_1, K_2 and K_3 are kernel functions, see Fan *et al.* (1996), Hall *et al.* (1992) and Silverman (1986) for further details. For applications of kernel density estimation in forensic science on independent observations see Aitken and Taroni (2004). The Gaussian kernel is used for all three functions K_1, K_2 and K_3. Two different bandwidth types are used. The first type is a fixed bandwidth, in which h_1, h_2 and h_3 remain constant at all values of w_{it} and $w_{i,t-1}$. The second type is an adaptive nearest neighbour bandwidth (Breiman *et al.* (1977)). This type of bandwidth will vary,depending on the amount of data close by, becoming larger as the amount of nearby data reduces.

Given $\hat{g}_{D_i}$ and $\hat{r}_{D_i}$, the conditional density function $\hat{f}_{D_i}(w_{it}|w_{i,t-1})$ in (1) can then be obtained for each of the m_B samples in $\mathbf{x}$ and for each of the m_C samples in $\mathbf{y}$.

3 Classification for a set of datapoints of unknown type

Let $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ be the datapoints from a sample for which it is wished to calculate the evidential value. The likelihood ratio associated with the propositions H_C and H_B is given by $f(\mathbf{z} \mid H_C)/f(\mathbf{z} \mid H_B)$. If this statistic is greater than one, then the evidence assigns more weight to H_C.

3.1 Parametric models

The parameters for which posterior distributions were obtained in the previous section are denoted by θ_{C_i} and θ_{B_i} (for either the autoregressive or the hidden Markov model), where i denotes the sample used to obtain the posterior distributions. The likelihood $f(\mathbf{z}|H_D)$ (swapping D for C or B as appropriate), is given by:

$$\begin{aligned} f(\mathbf{z}|H_D) &= \int_{\Theta_D} f(z_1 \mid \theta_D) f(z_2 \mid z_1, \theta_D) \ldots f(z_n \mid z_{n-1}, \theta_D) f(\theta_D \mid \mathbf{w})\, \mathrm{d}\theta_D \\ &\simeq \sum_{i=1}^{i=m_D} v_i \int_{\Theta_{D_i}} f(z_1 \mid \theta_{D_i}) f(z_2 \mid z_1, \theta_{D_i}) \ldots f(z_n \mid z_{n-1}, \theta_{D_i}) f(\theta_{D_i} \mid \mathbf{w_i})\, \mathrm{d}\theta_{D_i} \end{aligned}$$

Here $\mathbf{w_i} = \mathbf{x_i}$ if $D = B$ and $\mathbf{w_i} = \mathbf{y_i}$ if $D = C$. Let the weights v_i be given by $v_i = n_{D_i} / \sum_{i=1}^{i=m_D} n_{D_i}$. Each integral can be estimated using Monte Carlo integration.

3.2 Nonparametric models

When the nonparametric models are used, the likelihood takes a slightly different form. For proposition H_D, the likelihood for $\mathbf{z}$ is given by:

$$\begin{aligned} f(z_1, z_2, \ldots z_n \mid H_D) &= f(z_1 \mid H_D) f(z_2 \mid z_1, H_D) \ldots f(z_n \mid z_{n-1}, H_D) \\ &\simeq \sum_{i=1}^{m_D} v_i \hat{f}_{D_i}(z_1 \mid H_D) \hat{f}_{D_i}(z_2 \mid z_1, H_D) \ldots \hat{f}_{D_i}(z_n \mid z_{n-1}, H_D) \end{aligned}$$

	Hidden Markov Model / AR(1)	AR(1)	Nonparametric fixed bw	Nonparametric adaptive nn
Associated with cocaine	0.373 (25/67)	0.371 (26/70)	0.271 (19/70)	0.257 (18/70)
General circulation	0.096 (18/188)	0.151 (29/192)	0.321 (62/193)	0.269 (52/193)

Table 1: Misclassification probabilities out of (.) samples

where $\hat{f}_{D_i}(z_t \mid z_{t-1})$ is the estimated conditional density of sample i, $i \in (1, \ldots, m_D)$, for datapoints $t = 2, \ldots, n_{D_i}$ and $\hat{f}_{D_i}(z_1)$ is the marginal density for datapoint 1. The method for estimating these functions was given in the previous section. v_i is a weight assigned to each sample i, with $\sum v_i = 1$. Let $v_i = n_{D_i} / \sum_{i=1}^{i=m_D} n_{D_i}$ as for the parametric models.

4 Results

Tandem mass spectrometry data, taking the form of the ion count for the cocaine product ion *m/z 105* for samples of banknotes are available. Further details can be found in Dixon (2006) and Lloyd (2009). A peak detection algorithm has been developed which converts these ion counts into a peak area, which corresponds to a measure of the amount of cocaine on each of the banknotes within a sample. To reduce the skewness of these data, the logarithms of these peak areas were taken. For a sample of banknotes brought in by law enforcement agencies, it is desired to calculate the likelihood ratio for the propositions: H_C, that a sample of banknotes is associated with criminal activity involving cocaine, and H_B, that a sample of banknotes is from general circulation. Two training sets of data were used to calculate this likelihood ratio. The first, the data associated with H_C, or $\mathbf{y}$, consisted of samples of banknotes seized from a suspect who was found guilty of a crime involving cocaine. The second set of data, $\mathbf{x}$, associated with H_B, consisted of samples of banknotes which had been taken from general circulation (see Wilson *et al* (2013)).

Each sample in $\mathbf{x}$ and $\mathbf{y}$ was treated as the evidential data in turn, and the four models described earlier were used to estimate the posterior distributions of the parameters (parametric models) or conditional distributions (nonparametric models) from the remaining data. Details on the prior distributions used are given in Wilson *et al* (2013). The hidden Markov model was not used for all samples. Instead, Bayes Factors were calculated for the autoregressive and hidden Markov models, and the model with the larger Bayes' Factor was used for that sample. The likelihood ratio was then calculated for the evidential data. Misclassification probabilities, where either a sample in $\mathbf{x}$ had been assigned a likelihood ratio of greater than one, or a sample in $\mathbf{y}$ had been assigned a likelihood ratio of less than one, were calculated. These misclassification probabilities are given in table 1.

Some of the samples of banknotes in the set associated with the proposition H_C were contaminated in line with general circulation. As a result, it is not expected that misclassification probabilities for the set $\mathbf{y}$ will be low, as samples with contamination in line with general circulation will be misclassified. Therefore, misclassification probabilities of general circulation samples are used to assess the models. Table 1 shows that the model which models some samples with a hidden Markov model has the smallest misclassification probability for the general circulation banknotes, at 9.6%, slightly lower than the 15% misclassification probability achieved when using the autoregressive model alone. The nonparametric

models have much larger misclassification probabilities. Analysis of the absolute values of the log likelihood ratios indicated that the nonparametric models sometimes produced large erroneous absolute log likelihood values for misclassified samples. These problems could be due to lack of data in the tails for some of the conditional distribution estimates.

5 Conclusion

The models developed in this paper give a novel method for calculating the likelihood ratio of two competing propositions when autocorrelated data are involved. Four models are described which allow for dependence between adjacent datapoints, and one of the models also allows for dependence on a latent Markov chain.

The models are tested on data based on cocaine quantities on banknotes. Previous models used on similar data have assumed independence between adjacent banknotes (Besson 2004, Jourdan *et al* 2013). The best performance was obtained when samples were modelled by either a hidden Markov model or an autoregressive model, with the model selection done using Bayes' Factors. This model had a misclassification probabilitiy of samples from general circulation of 9.6%.

6 References

Aitken,C.G.G. and Taroni,F. (2004) *Statistics and the Evaluation of Evidence for Forensic Scientists*, second edition. Chichester: John Wiley and Sons Ltd.

Besson,L. (2004) Détection des stupéfiants par IMS. Master's thesis, Univ. of Lausanne.

Breiman,L., Meisel,W. and Purcell,E. (1977) Variable kernel estimates of multivariate densities. *Technometrics*, **19**, 135-144.

Dixon,S.J., Brereton,R.G., Carter,J.F. and Sleeman,R. (2006) Determination of cocaine contamination on banknotes using tandem mass spectrometry and pattern recognition, *Analytica Chimica Acta*, **559(1)**, 54-63.

Fan,J., Yao,Q. and Tong,H. (1996) Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, **83**, 189-206.

Hall,P., Racine,J. and Li,Q. (1992) Cross-validation and the estimation of conditional probability densities, *Journal of the American Statistical Association*, **87**, 523-532.

Jourdan,T., Veitenheimer,A., Murray,C., Wagner,J. (2013) The quantitation of cocaine on U.S. currency: survey and significance of the levels of contamination, *Journal of Forensic Sciences*, in press.

Lindley,D.V. (1977) A problem in forensic science. *Biometrika*, **64**, 207-213.

Lloyd,G.R. (2009) Chemometrics and pattern recognition for the analysis of multivariate datasets, PhD thesis, Univ. of Bristol.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pp. 257-286.

Silverman,B.W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Wilson,A., Aitken,C.G.G., Sleeman,R. and Carter,J. (2013), The evaluation of evidence relating to traces of cocaine on banknotes. *Submitted for publication*