Mats Rudemo, March 8, 2001

Solutions for problems in Examination in Statistical Image Analysis, March 7, 2001

Problem 1. *Suppose that you have a set 10 images of haddocks and 10 images of whitings, eight of these images are shown in Figure 1 (in the problem set). The object is to construct by use of these data a method for discrimination between the two species.*

a) *Suggest two features that could be used for discrimination. Sketch how you could implement the computation of these features from the images, e.g. by use of a program system like matlab. Describe in words how you may implement the computations (without giving programs).*

Choose, for instance, Y_1 as the size of the fish part A of one of the 20 two-dimensional images and Y_2 as the compactness, that is,

$$Y_2 = 4\pi \frac{\text{area}(A)}{(\text{perimeter}(A))^2}. \tag{1}$$

To compute Y_1 and Y_2 we need first to find the fish segment in an image. This can, for instance, be done by use of an edge detector (a filter that finds large gradients) to find the contour. After the contour is found one needs to identify the part of the image that is inside (as by filling holes, see below). Another possibility may be to use thresholding plus perhaps some morphological operations such as dilation and erosion. After that one needs to fill out holes (as in Figure 16 in the notes). There may be problems with the lower left parts of the fishes that do not give a clear contrast to the background. Perhaps some manual intervention is needed.

Suppose that the fish segment has been found and that we have a binary image of the fish as a black segment A against a white background. Then the area of A may be defined as the number of black pixels and the perimeter as the number of black pixels with a white neighbour.

b) *Give statistical models for the data for the two features and for the 10 plus 10 fishes in the data set. Consider both a model leading to linear discrimination and a model leading to quadratic discrimination. Sketch plots showing the difference between the corresponding two discrimination methods. How would you estimate parameters in the models?*

Put $X = (Y_1, Y_2)^T$ and assume that X has a two-dimensional normal distribution with density function

$$f_X(x) = \frac{1}{2\pi(\det C)^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)), \tag{2}$$

with $\mu = \mu_i$ and $C = C_i$ for the two classes, say with $i = 1$ for haddock and $i = 2$ for whiting. Let π_1 and π_2 be the prior probabilities of the two species. (We may either choose $\pi_1 = \pi_2 = 0.5$ or let π_1 and π_2 be the relative frequencies of the two species.) If we assume $C_1 = C_2 = C$ we get linear discrimination and we choose $i = 1$, that is 'haddock', if

$$(\mu_1 - \mu_2)^T C^{-1}(x - \frac{1}{2}(\mu_1 + \mu_2)) > \ln \frac{\pi_2}{\pi_1}. \tag{3}$$

Without the assumption $C_1 = C_2 = C$ we choose 'haddock' if

$$\frac{1}{2}x^T(C_2^{-1} - C_1^{-1})x + (\mu_1^T C_1^{-1} - \mu_2^T C_2^{-1})x + \frac{1}{2}(\mu_2^T C_2^{-1}\mu_2 - \mu_1^T C_1^{-1}\mu_1) > \ln \frac{\pi_2(\det C_1)^{1/2}}{\pi_1(\det C_2)^{1/2}} \tag{4}$$

which is quadratic discrimination.

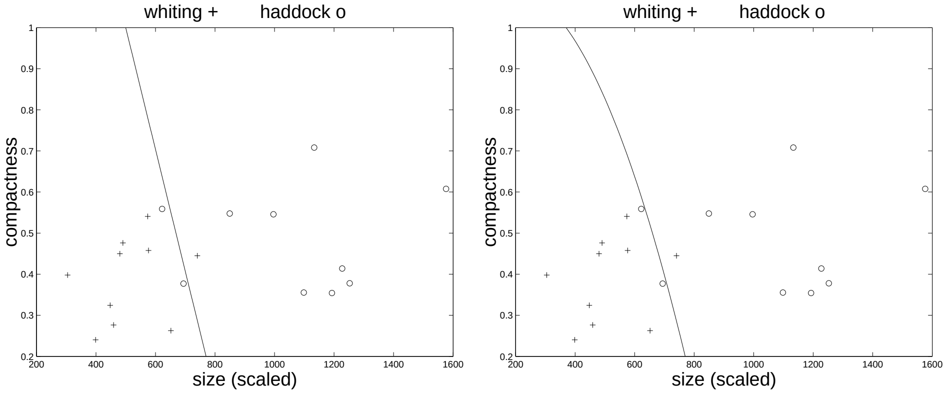

Figure 1: Sketch of how scatter plots with curves for discrimination may look. Linear discrimination to the left and quadratic discrimination to the right.

To estimate parameters we let X_{im}, denote the mth (two-dimensional) observation vector in the ith class. Here $i = 1, 2$, corresponding to 'haddock' and 'whiting', and $m = 1, \ldots, n$, with $n = 10$. Then use the estimates

$$\hat{\mu}_i = \frac{1}{n}\sum_{m=1}^{n} X_{im}, \quad i = 1, 2. \tag{5}$$

If we make no assumption on equality of the covariance matrices we use the covariance matrix estimates

$$\hat{C}_i = \frac{1}{n-1}\sum_{m=1}^{n}(X_{im} - \hat{\mu}_i)(X_{im} - \hat{\mu}_i)^T, \quad i = 1, 2. \tag{6}$$

If we assume equality of the covariance matrices we use instead the estimate

$$\hat{C} = \frac{1}{2}(\hat{C}_1 + \hat{C}_2) \tag{7}$$

for the common covariance matrix C.

c) *Suggest 10 features for the discrimination and one or (preferably) several methods for selecting 3 features out of these in an optimal way.*

Let, for instance $Y_3, \ldots, Y_{10}$ be eight moment variables of the type considered on page 21 in Mattias Andersson's master thesis.

We want to select 3 variables among $Y_1, \ldots, Y_{10}$ so that the error-rate becomes small.

One possibility is to consider all possible $(10{\cdot}9{\cdot}8)/(3{\cdot}2{\cdot}1) = 120$ combinations of three variables and choose the combination that gives the smallest error-rate estimate.

Another possibility is to use forward selection, which in the present case takes the following form:

- Start by chosing the variable that considered alone gives the smallest error-rate estimate.
- Then choose a second variable that together with the first chosen variable gives the smallest error-rate estimate.
- Finally choose a third variable that together with the first two variables chosen gives the smallest error-rate estimate.

For the error-rate estimate we can either use the resubstitution error-rate estimate or the cross-validation error-rate estimate. The cross-validation error-rate estimate is in general more accurate but requires more computation with repeated parameter estimates successively leaving out one observation vector.

Problem 2. *Figure 2 (in the problem set) shows two microscope images with nerve cells viewed against a background with a net of squares. The images are taken with an interval of 1 minute and 15 seconds. The cells can move and they can divide. The black rectangle surrounds a cell that has moved slightly to the left in the time interval between the image acquisitions.*

a) *Regard first one of the images. Discuss how one can construct an image analysis method to count the nerve cells and how to estimate the positions of the cell centres (with a suitable definition of cell centre). Discuss briefly how one can test the hypothesis that the positions of the cell centres are completely randomly distributed, that is distributed as points from a Poisson process. Regard the images and discuss what types of deviations from a completely random distribution you may expect.*

First we try to construct a binary image with each cell as a connected segment of black pixels. This task is similar to the task in the first part of Problem 1. One could for instance try to find the contour of cells characterised by a pronounced shift from medium grey to white pixels. Or one could try thresholding to find "white" connected parts with each hole in such a white connected part corresponding to one cell. Perhaps one could use the fact that most of the cells seem to be convex. Some manual intervention may be needed.

Suppose that we have found the cells so that each cell is a connected black segment. Then we count the number of cells by counting the number of segments.

Let C denote the set of pixels for one cell, and let n_C denote the number of such pixels. For simplicity we identify a pixel with its corresponding pixel centre (s,t), and we define the cell centre $(\overline{s},\overline{t})$ by

$$\overline{s} = \frac{1}{n_C} \sum_{(s,t)\in C} s, \qquad \overline{t} = \frac{1}{n_C} \sum_{(s,t)\in C} t. \tag{8}$$

Regard now the point process X of cell centres on the set $A \subset \mathbb{R}^2$, where A is area shown in one of the images of Figure 2 (in the problem set). The intensity of X we estimate by

$$\hat{\lambda} = \frac{X(A)}{|A|}. \tag{9}$$

Let further $\hat{K}$ denote a K-function estimate, for instance the estimator on page 52 in the notes,

$$\hat{K}(r) = \frac{1}{\hat{\lambda}^2 |A|} \sum_{x\in X} \sum_{y\in X} \frac{1\{0 < \|y-x\| < r\}}{w(x,y)}. \tag{10}$$

We can then plot the square-root of $\hat{K}(r)$ against r. Deviations from the straight line $\sqrt{\pi}r$ indicates deviations from a completely random distribution.

If the estimate lies below the line inhibition is indicated, while an estimate above the line indicates clustering.

Regarding the two images in Figure 2 one might suppose that we have inhibition for very small r-values (as the cells do not seem to overlap) and perhaps clustering for slightly larger r-values.

To test formally the assumption of a purely random distribution of the cell centres a simulation type test may be used as described by Diggle (1983). This test method consists of first estimating the K-function for the observations (as discussed above) and then to find confidence limits for this function by simulating a number of Poisson processes with the same intensity and computing upper and lower envelopes for the corresponding K-function estimates.

b) *Suppose you have computed estimates of the cell centres in both images. Suggest a method for estimating the distribution of the cell centre movements. Assume for instance that the motion follows a two-dimensional normal distribution and discuss how you can estimate the parameters of this distribution.*

Let X denote the point process of cell centres in the left image of Figure 2 and let Y denote the point process of cell centres in the right image. Make a pairing of points by a "greedy search" algorithm as follows. Start growth of circles at the same rate around each of the X-points. As soon as one circle hits a Y-point stop the growth of that circle and pair the corresponding X- and Y-points. Proceed by letting the other circles grow and stop the growth of any circle as soon as a circle hits a Y-point not previously paired and pair the corresponding X- and Y-points. Stop the circle growth process when a certain size r_0 of the circles have been reached (or earlier if all X-points or all Y-points have been paired).

Let (X_i, Y_i), $i = 1, \dots, n$ denote the set of paired points and put

$$Z_i = Y_i - X_i, \qquad i = 1, \dots, n. \tag{11}$$

Assume that $(Z_1, \dots, Z_n)$ is a sample from a two-dimensional distribution $\mathrm{N}(\mu, C)$ that corresponds to the motion of cells from one image to another. Estimate the parameters by

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Z_i, \tag{12}$$

and

$$\hat{C} = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \hat{\mu})(Z_i - \hat{\mu})^T. \tag{13}$$

The assumption that $(Z_1, \ldots, Z_n)$ is a sample from a two-dimensional distribution corresponding to the motion of cells from one image to another can be critized:

- Our procedure excludes motion with $|Z| > r_0$.
- Edge effect are disregarded – one cell observed in the left image may have disappeared in the right image.
- If two cells are close in the left image, the corresponding paired cells in the right image may be swapped in the "greedy search".

However, if the time interval between the two images is small these effects are presumably neglible.

c) *Try to formulate a statistical model for the positions of cell centres in a sequence of several consecutive images, taken say with an interval of 1 minute and 15 seconds as for the two images in Figure 2. Consider both possible motion and possible division of cells, although not more than one division in the time interval between two consecutive images.*

Consider first two conscutive images.

Assume that cells divide independently of each other and that the probability that a cell divides from one image to another is p. A model may be specified as follows. Start with a set $X = \{X_1, \ldots, X_m\}$ of cell centres in one window A. For a cell centre with position X_i we generate a variable U_i that is uniformly distributed on the interval (0,1). If $U_i > p$ the cell does not divide, and we let the cell move to $X_i + Z_i$, where Z_i is $\mathrm{N}(\mu, C)$. If $U_i \leq p$ we let instead two particle be generated at positions $X_i + V_i$ and $X_i + W_i$ where V_i and W_i are independent and have a two-dimensional normal distribution $\mathrm{N}(\mu', C')$. Edge effects at the border of A have to be specified. For a simulation model one simple possibility is to use periodic boundary conditions as on page 36 in the notes. If we want to use the model for estimation of parameters another possibility is to just disregard edge effects as in **b)** above.

For a sequence of consecutive images we iterate the model above, which describes how we go from one image to the next one. For a complete model we need also to specify the starting configuration. One simple possiblity would be to start with a Poisson distribution of the points.