Mats Rudemo, April 5, 2005

Solutions for problems in Examination in Statistical Image Analysis, March 15, 2005

Problem 1. *In a two-colour microarray experiment images were obtained separately for two colour channels: red cy5 (here corresponding to wild-type Arabidopsis) and green cy3 (corresponding to one transgenic Arabidopsis line). Figure 1 below shows to the left the signal intensity for the red channel in one part of the array with 9 spots and to the right a detail with the central of these nine spots. The signal is registered in two bytes, and the signal thus lies between 0 and* $2^{16} - 1 = 65535$. *Consider modeling of images such as the right part of Figure 1.*

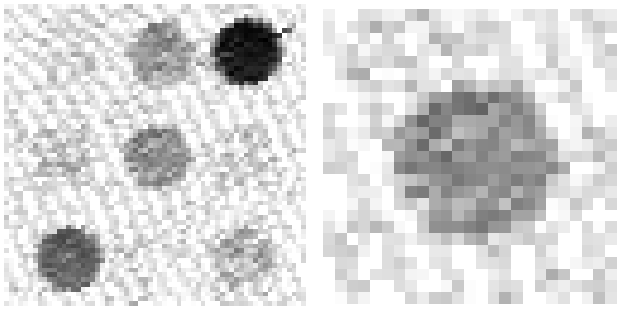

Figure 1: *Left: red channel image of nine spots in a microarray experiment. Right: a detail with* 25×25 *pixels of the left image corresponding to the central spot. In the images black corresponds to high signal intensity.*

a) *Formulate a statistical model for an image such as the right part of Figure 1. Assume that the registered intensity consists of a sum of a signal part and a noise part. The signal part is assumed to be constant (with a given spot amplitude) within a circle with a given spot centre and a given spot radius. The noise part is assumed to consist of normal variates with a constant mean and a constant variance. These noise normal variates are assumed to be independent for different pixels. The parameters corresponding to spot centre, spot radius, spot amplitude, noise mean and noise variance are assumed to different for different spots.*

Let S denote the set of spots. With each spot $s, s \in S$, we associate a set A_s of pixels, in the present case for instance a square of 20 by 20 pixels with the spot approximately in the center. We assume that no pixel belongs to more than one such set, and some pixels may not be associated with any spot. Let $Y = Y(x)$ denote the (possibly transformed) intensity at a pixel, x, with pixel centre coordinates $x = (x_1, x_2)$.

Consider a spot s and pixels $x \in A_s$. Let $c_s = (c_{s1}, c_{s2})$ be the spot centre of spot s, and let $r_s(x) = \|x - c_s\|$ be the distance from pixel x to the spot centre. Assume that

$$Y(x) = B_s \frac{1}{\pi\sigma_s^2} 1(r_s(x) \leq \sigma_s) + b_s + \epsilon(x), \qquad x \in A_s \tag{1}$$

where B_s measures the intensity of spot s, b_s is a constant representing the background, $1(P) = 1$ if P is true and $1(P) = 0$ otherwise, $\sigma_s > 0$ is the radius of the spot and $\epsilon(x)$ corresponds to zero-mean noise at x. We assume that $(\epsilon(x), x \in A_s)$ are independent and normally distribution with mean zero and constant variance σ_e^2.

b) *Suggest a method for estimating the parameters for a given spot based on data such as those shown in the right part of Figure 1.*

A suitable method is to use maximum likelihood. We disregard the possibility that some intensity values are saturated, that is, are above the upper two-byte limit $2^{16}-1$. (Note that at least in the right part of Figure 1 no intensity values are saturated.)

The 6 parameters $B_s, \sigma_s, c_{s1}, c_{s2}, b_s, \sigma_e$ may be estimated by maximizing the log likelihood function

$$L = \sum_{x \in A_s} \log\{\frac{1}{\sigma_e}\phi(\frac{Y(x) - \frac{B_s}{\pi\sigma_s^2}1(\|x - c_s\| \leq \sigma_s) - b_s}{\sigma_e})\} \tag{2}$$

where ϕ is the standardized normal density function, $\phi(y) = \frac{1}{\sqrt{2\pi}}\exp(-y^2/2)$.

The log likelihood (2) can be maximized by standard iterative maximization techniques, e.g., quasi-Newton or Nelder-Mead. Such algorithms are available for instance in the Matlab optimization toolbox.

c) *Look at the images in Figure 1. Discuss how reasonable the different assumptions for the modelling described in* **a)** *above seem.*

The assumptions are:

- signal part is assumed constant within a circle: *seems ok*
- the noise part is assumed to consist of normal variates with a constant mean and a constant variance: *constant mean and constant variance seem ok, normality difficult to judge from figure*
- the noise normal variates are assumed to be independent for different pixels: *independence does not seem so well satisfied, note the stripe pattern*

Problem 2.

Eggs of parasites of swines can be detected in fecal samples from the animals. Figure 2 shows images of eggs from seven subspecies of Eimeria *parasites. Suppose that we want to discriminate between subspecies and that we have an image analysis algorithm that finds the contour of the eggs and the distances X and Y defined in the following way. We assume that the contour of the eggs is convex. Let P_1 and P_2 be two points on the contour maximally apart. Let X be the distance between P_1 and P_2. Let L_1 be the line going through P_1 and P_2. Let P_3 be the point on L_1 midway between P_1 and P_2, and let L_2 be the line through P_3 perpendicular to L_1. Let Y be the distance between the two points on the contour where L_2 crosses the contour. Draw an image showing these points, lines and distances. Put $Z = Y/X$. We want to discriminate between parasite subspecies by use of Z only. Consider for simplicity the case with two parasite subspecies.*

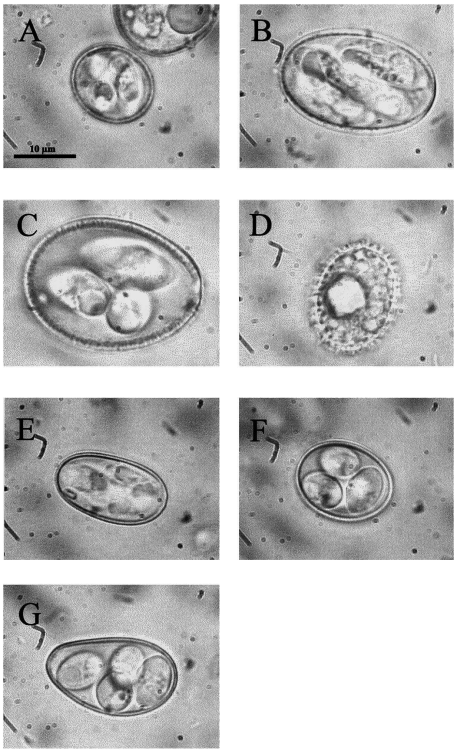

Fig. 1. Oocysts of group 1 (A: *E. perminuta*, B: *E. polita*, C: *E. scabra*), group 2 (D: *E. spinosa*) and group 3 (E: *E. debliecki*, F: *E. suis*, G: *E. porci*) *Eimeria* spp.

Figure 2: *Figure from Daugschies et al. (1999) Differentiation between porcine* Eimeria *spp. by morphological algorithms, Veterinary Parasitology 81, 201–210, showing egg shapes for seven subspecies.*

a) *Formulate a statistical model for discrimination between the two species by use of Z.*

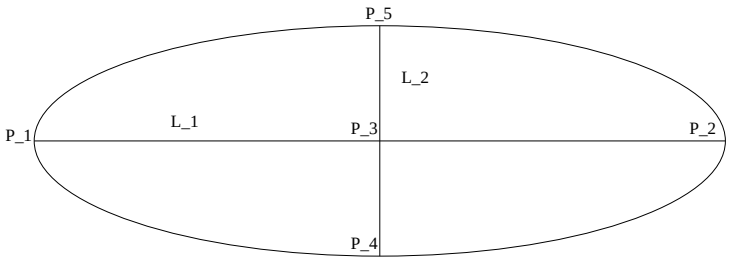

Figure 3: Drawing showing contour of egg, X is the distance between P_1 and P_2 and Y is the distance between P_4 and P_5.

Let X and Y be the distances described in the legend of Figure 3, put $Z = Y/X$, and let ω_1 and ω_2 denote the classes corresponding to the two subspecies. Let π_i denote the prior probability of class ω_i, $i = 1, 2$, and let f_i be the probability density of Z for an observation from class ω_i.

The problem of deciding if an object comes from class ω_1 or ω_2 is to be based on observation of the corresponding feature variable Z. To find discrimination we need further specification corresponding to how costly it is to make different kinds of errors, that is the cost of choosing class ω_1 when ω_2 is true and vice versa. Let us assume that these cost are equal, and more specifically, that we want to minimize the probability of misclassification.

From the course notes we know that the probability of misclassification is minimized if we use the following rule:

$$\text{prefer class } \omega_i \text{ to } \omega_j \text{ if } \pi_i f_i(z) > \pi_j f_j(z), \tag{3}$$

when $Z = z$ is observed. Assume further that Z is $\mathrm{N}(\mu_i, \sigma_i^2)$ in class ω_i, $i = 1, 2$.

Let us first assume that we have equal variances in the two classes. Then it follows from (3) that we minimize the probability of misclassification if we prefer class ω_i to ω_j if

$$(\mu_i - \mu_j)\sigma^{-2}(Z - \frac{1}{2}(\mu_i + \mu_j)) > \ln \frac{\pi_j}{\pi_i}. \tag{4}$$

which gives linear discrimination.

Let us now find a corresponding rule without the assumption of equal variances. It follows that we shall prefer class ω_i to ω_j if

$$\frac{1}{2}(\sigma_j^{-2} - \sigma_i^{-2})Z^2 + (\mu_i\sigma_i^{-2} - \mu_j\sigma_j^{-2})Z + \frac{1}{2}(\sigma_j^{-2}\mu_j^2 - \sigma_i^{-2}\mu_i^2) > \ln \frac{\pi_j\sigma_i}{\pi_i\sigma_j}. \tag{5}$$

We see that the border between the two regions where we should or should not prefer ω_i to ω_j is given by a quadratic function (quadratic discrimination).

b) *Suppose that we have images of n_1 eggs of variety 1, and n_2 eggs of variety 2. Give formulas for estimation of the parameters in the model in* **a)**.

We now have a training set with n_i objects from class ω_i, $i = 1, 2$. From both classes we assume that we have obtained independent random samples of objects. We assume further that the vector Z is normally distributed with expectation vector μ_i and variance σ_i^2 in class ω_i. Let the observations be denoted Z_{im}, $m = 1, \ldots, n_i$, $i = 1, 2$. Then it is natural to estimate the expectation in class ω_i by

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{m=1}^{n_i} Z_{im}, \quad i = 1, 2. \tag{6}$$

If we make no assumption on equality of the variances we use the variance estimates

$$s_i^2 = \frac{1}{n_i - 1} \sum_{m=1}^{n_i} (Z_{im} - \hat{\mu}_i)^2, \quad i = 1, 2, \tag{7}$$

but if we assume variance equality we use the estimate

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{8}$$

for the common variance.

For the prior probabilities we use the estimates $\hat{\pi}_i = n_i/(n_1 + n_2)$, $i = 1, 2$.