Chapter 3

Integrating Out Hidden Variables: Yuille, Coughlan, Kersten, Schrater

"Condition on what you know and marginalize over what you don't care about."

Unpublished sayings of Lao Tzu (Translation by Dr. J.M. Coughlan).

"We must integrate all Republicans under a big tent, or else we will split the (expletive deleted) vote".

Unpublished tapes of Richard Nixon. (Restored by Dr. A.L. Yuille).

Why integrate out hidden variables? When is marginalization necessary? Consider, for example, splitting the vote in an election. Suppose the Republicans and Democrats both have a "favourite" candidate and a "challenger". Both the Republicans are slightly less probable to be elected than the "favourite" Democrat. So, if the task is to estimate the most probable candidate then you would get the Democrat. But, if you want to estimate the most probable party voted for you would get Republican. In this case it would be a good idea for the Republicans to integrate their members under a big tent behind a single candidate (who would get the votes of all the Republican groups). See figure (3.1).

DAN QUESTION – Are you on the boundary as an example??

3.1 Hidden Variables: An Example and Overview

The previous chapters have described how we can perform classification and estimation. So far, however, we have only dealt with comparatively simple models. For example, we assumed that the observations x are directly related to the states s by a conditional distribution $P(x|s)$.

In more realistic cases there may be other variables involved. For example, suppose we wish to have a model for recognizing objects under variable illumination conditions (but fixed viewpoint and pose). In this case, the data x is the image of one of the objects $s_1, ..., s_N$. But the images of the objects will depend on the (unknown) lighting conditions which are characterized by parameters s (labelling, for example, different lighting config-

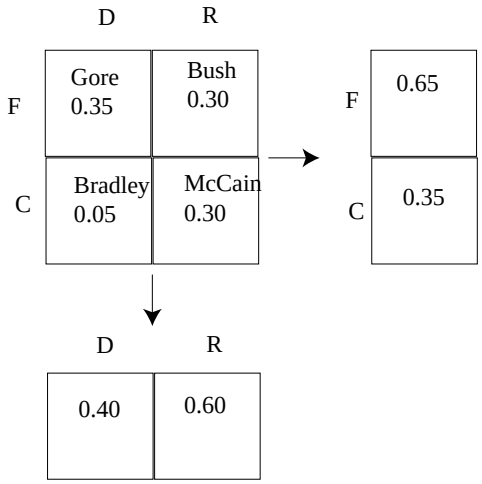

Figure 3.1 The Democrats (D) and the Republicans (R) both have a favourite (F) and a challenger (C) as candidates. The probabilities of election of these candidates – Gore, Bradley, Bush, McCain – is specified in the box (top left) and the favourite candidate is Gore (in this example!). But marginalizing over candidates makes the Republicans the more popular party. The best strategy for the Republicans is to integrate their members into a big tent behind a single candidate.

urations). The conditional distribution for the data must be of form $P(x|s,h)$ and we may have a prior probability on the lighting conditions and objects $P(s,h) = P(s)P(h)$ (where s takes values $s_1, ..., s_N$). (We assume that the prior probability of the lighting conditions is independent of the objects).

Bayes theorem allows us to determine

$$P(s,h|x) = \frac{P(x|s,h)P(s)P(h)}{P(x)}. \tag{3.1}$$

At this stage we are faced with a choice. Are we interested purely in recognition? Or do we also want to estimate the lighting conditions h? Or, perhaps, we might simply want to estimate the lighting conditions and not care about the object recognition.

Firstly, suppose only care about recognition. It this case we need to marginalize over the variables h[1]. This marginalization will take the form of integration, if the variables h are continuous, or summation if they are discrete. This gives:

$$P(s|x) = \int dh P(s,h|x) \quad or \quad P(s|x) = \sum_h P(s,h|x). \tag{3.2}$$

For our specific example, marginalizing over the light sources means that we are effectively looking for properties of the image which are relatively invariant to the lighting

[1]Note that if a loss function is *independent* of a variable then decision theory says that one should marginalize over the variable.

conditions. From theoretical studies (Belhumeur et al) it is known that there are, in general, no image properties of an object which are fully invariant to lighting (except for some extremely simple objects). Nevertheless, there are *probabilistic regularities* about how image properties of objects behave under different lighting. Such properties are those which, in principle, are captured by $P(s|x)$.

We should now do standard Bayes decision theory on the marginalized distribution $P(s|x)$. Indeed we *can apply basic decision theory to any problem, where the state vectors have arbitrarily many hidden variables, provided we condition on the measurements and marginalize out the variables that we are not interested in.* In particular, as we will see in a later chapter!! we can apply decision theory to hidden markov models for speech where the number of hidden variables is enormous.

But this general statement "marginalize over the variables you are not interested in" is more easily said than done. In many cases, this marginalization is impossible to perform either analytically or even by computer (in an acceptable time). In such cases, we will need to fall back on approximation techniques which will be discussed later in this chapter. We should also add that, although the *best strategy* requires marginalization, it may be unnecessary and approximations may be very effective in realistic situations. In addition, as we will show later, there are situations involving *symmetry* where the information required to make decisions can be extracted without needing to performing the marginalization!

Secondly, suppose we want to estimate both the object and the lighting conditions. In this case, we should start with equation (3.1) and attempt to estimate both the variables s, h simultaneously. This estimation requires specifying a decision function. If we allow all lighting conditions then the variable h is continuous. We should therefore be wary of trying to use a MAP estimator because, from the previous chapter, this involves rewarding our decision *only if our estimation is perfectly accurate.* But it seems highly unlikely that light source configurations can be estimated to high precision (either by a computer algorithm or by human observers). Thus a somewhat "tolerant" loss functions will be needed when estimating the light source h. The set of objects we are considering is discrete so we could use a Kronecker delta loss function for the s variable.

Thirdly, suppose we attempt to estimate the light source directly. There are, of course, several algorithms for estimating the light source directly can be found in the computer vision literature. Such algorithms, however, rely on restrictive assumptions about the images being viewed. Moreover, none have been rigourously evaluated to determine their degree of precision[2]. From our decision theoretic perspective, to estimate the light source requires marginalizing over the objects $\{s_i\}$ to obtain the distribution $P(h|x) = \sum_i P(h, s_i|x)$ and estimating h from this. Whether this is feasible, or not, depends on whether the distribution $P(h|x)$ is sharp enough to give a reliable estimator for the light source h, see

[2]Including one by the first author of this book.

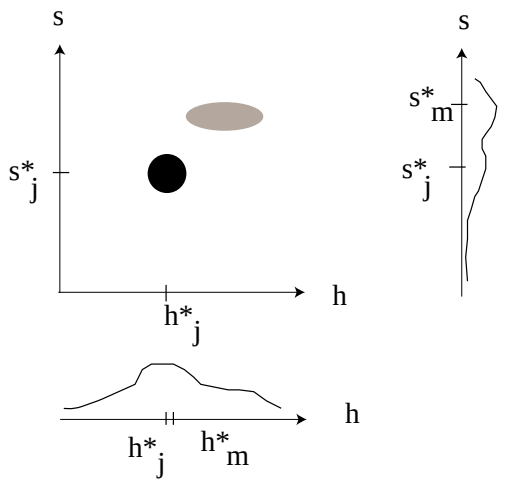

Figure 3.2 The joint probability density function of h, S are zero except in the circle and ellipse. The probability density is greatest in the circle (dark shading) and is smaller in the ellipse (light shading). In this example, estimating the most probable s and h *independently* from their marginal distributions gives a different answer than estimating both of them simultaneously from the joint distribution. This is because the elongation of the ellipse (major axis along h and minor axis along s) means that the marginal distribution for s is determined mostly by the density in the ellipsoid. By contrast, the marginal distribution for h is determined mainly by the circle. If the distributions for our object recognition under variable lighting are of this type, then trying to first estimate the lighting (i.e. h) in order to use it to then estimate s would give poor results. The best strategy would be to estimate s directly by marginalzing out the lighting.

figure (3.2). It may be, of course, that $P(h|x)$ has multiple peaks. For example, it has been shown (Belhumeur, Kriegman, Yuille) that certain objects cannot be distinguished from each other *when viewed under a range of different lighting conditions* if the light source directions are unknown (i.e. if we see object s_i under *any* lighting h_i then we can determine a light source h_j such that object s_j, viewed under lighting h_j, looks identical to object s_i viewed under s_i.). See figure (3.3). We will return to this example in more detail in a few pages.

On pragmatic grounds, it is important to determine when the problems can be broken down into well-defined, and solvable, subproblems. A simple way to solve object recognition under variable lighting would be to first estimate the light source direction, independent of the object, and then proceed to estimate the object itself. Bayes decision theory tells us that this is *not the optimal procedure*. However, it may nevertheless be sufficiently accurate in any given application. It does throw away information. For example, it is theoretically possible that there are many object and lighting pairs which give rise to similar images. In such a case, the posterior distribution for the lighting direction may have many peaks and the estimation of lighting will be ambiguous. This type of problem may arise theoretically but, in practice, it may be irrelevant. So it should be emphasized

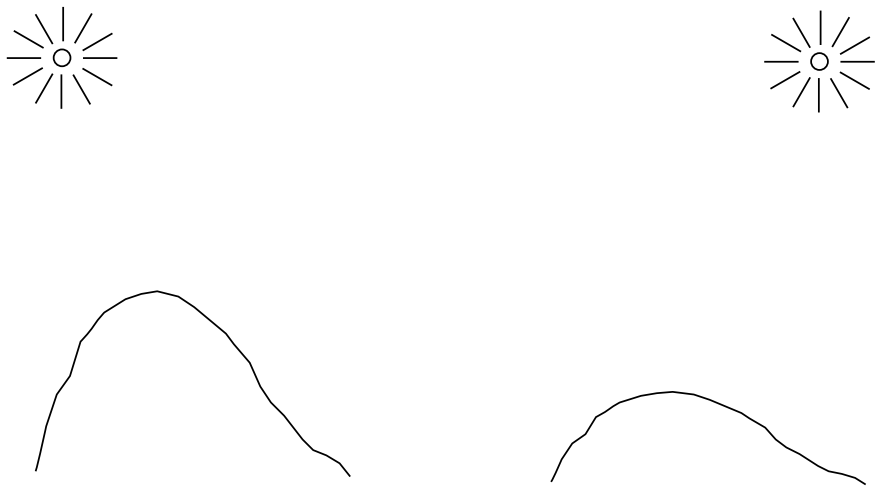

Figure 3.3 If the lighting conditions are unknown then it is impossible to distinguish between two objects related by a GBR (generalized bas relief ambiguity). Because for any image of the first object, under one illumination condition, we can always find a corresponding illumination condition which makes the second object appear identical (i.e. generate the identical image).

that for some problems *there may well be short-cuts which give the optimal solution, or close to it, by using simplified models.* However, to determine if such short-cuts are reliable we would have to know the full probability distributions and determine when the short cuts are justified. (See later chapter of the book).

These three tasks are the key concepts that we will discuss in this chapter. The standard problem is specified by a distribution $P(x, h, s)$ where x is the observations. We can then choose to either estimate s directly from $P(s|x)$, h directly from $P(h|x)$ or h, s jointly from $P(h, s|x)$.

3.2 Marginalizing over Continuous Hidden Variables

If the hidden variables are continuous, then marginalizing over them corresponds to integration. If this integration can be done, either analytically or by computer, then the problem reduces to standard decision theory on the marginal distribution $P(s|x)$. But there are also some cases where we may want to estimate the hidden variables alone, by estimating them from $P(h|x)$, or by jointly estimating them with the state s (i.e. by estimating from $P(h, s|x)$.)

In general, however, the integrations required for marginalization cannot be performed and so approximations are necessary. One of most important approximation techniques is Laplaces's method and its relatives such as the saddle point and stationary phase approximations. For other situations, it may be possible to extract the relevant information to solve the decision problem without needing to explicitly do the integral.

Now suppose that we are trying to integrate out the hidden variable h to obtain

$P(s|x) = \int dh P(s,h|x)$. Contributions to this integral come from all values of h for which $P(s,h|x)$ is non-zero. In particular, it is quite possible that there are subregions of H space (the space of the hidden variables) where $P(s,h|X)$ is comparatively small but which nevertheless *make a big contribution to the integral because of the size of the subregion.* In physicist's terminology, we have to consider the *phase space* of the h variable. The ambiguity in the original problem gets removed because of phase space considerations.

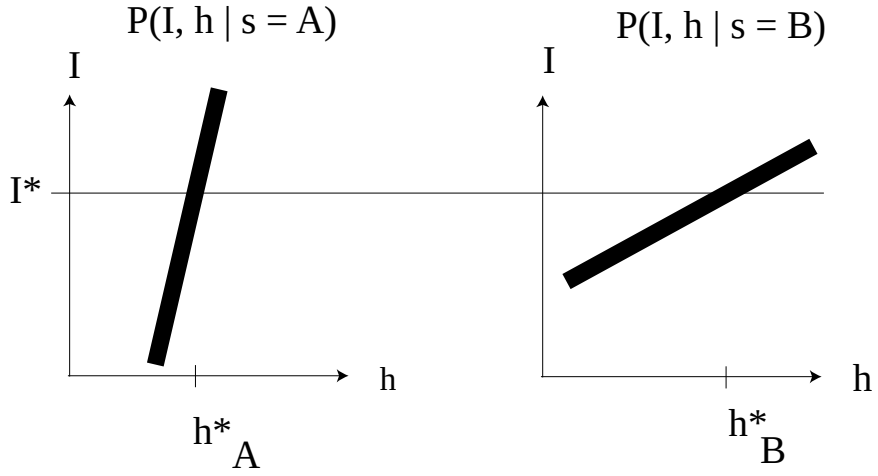

Figure 3.4 The probability density for I, h is zero except on the shaded rectangular bar (where it is constant). Because of the different amount of phase space (i.e. the different angles of the rectangles) we find that although $P(I, h_A^*|s = A) = P(I, h_B^*|s = B)$ we nevertheless have $P(I|s = A) < P(I|s = B)$.

More intuitively, suppose an image x is consistent with an object s_i and a very specific lighting condition h_i. But small changes in the lighting conditions $h_i \mapsto h_i + \Delta h$ cause big changes in the image which affect the probability that the viewed object is indeed s_i. In other words: $P(s_i|x, h_i)$ is large but $P(s_i|x, (h_i + \Delta h))$ is small for small changes Δh in the lighting. This is called a *non-generic* case (Freeman) in the sense that it is requires an *accidental alignment* of the light source to obtain the observed image for object s_i. It is better to seek interpretations of the data which are insensitive to small changes in the lighting direction. Such interpretations are called *generic*. It transpires (Freeman) that the Bayesian procedure of integrating out the hidden variables captures this concept of genericity and the precise mechanism is through phase space.

To get some insight into this we now proceed to work out an example (developed by Freeman and Brainard) which illustrates this point.

3.3 Phase Space and Integrating Hidden Variables

To understand the effect of phase space when integrating out variables, consider the following abstract example (from Brainard and Freeman).

The observation x is determined by two unknown variables s, h by probability distri-

bution:

$$P(x|s,h) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-sh)^2/(2\sigma^2)}, \tag{3.3}$$

We assume that all values of s, h are equally probable *a priori*. This is technically an *improper prior* on the variables s, h because it is not normalized.

Suppose we want to estimate s, h simultaneously. Then at first sight the problem seems to be ambiguous. If we apply ML estimation we find that there are a whole set of equally likely estimates s^*, h^* provided $s^*h^* = x$. See figure (3.5).

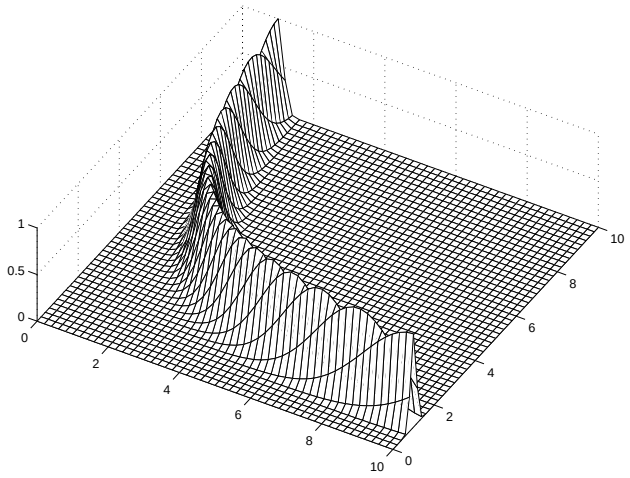

Figure 3.5 The probability distribution $P(x|s,h) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-sh)^2/(2\sigma^2)}$ as a function of s, h, shown for $x = 10.0$.

The problem, however, becomes well posed if we decide we want to estimate s only and *integrate out* the h variable. This integration can be done explicitly by observing that $P(x|s,h)$ is a Gaussian function (ignoring the normalization constant) of the variable h with mean x/s and variance $\sigma^2/(s^2)$, Integrating with respect to h gives:

$$\int dh P(x|s,h) = 1/s, \tag{3.4}$$

so the, slightly surprising, result is that the most probable value of s is 0 for *any* observation x. (Note we have not assumed a prior $P(s)$ or, to be more precise, we have assumed the *improper prior* that all values of s are equally likely).

The reason for this result is simple. For any value of x there will be a set of values of s, h which are sufficiently consistent with it to give significant contributions to the integral $\int P(x|s,h)dh$. These significant contributions lie close to the curve $sh = x$. Almost all the contributions come from places where $|sh - x| < 2\sigma$. If we fix s, then the main contributions come from the set of h such that $|h - x/s| < 2x/s$. So for small s, the

contributions to the integral come from a very large region in the space of h's. In other words, the amount of phase space of the h variables which make contributions increases. In fact, as $s \mapsto 0$, all possible values of h give contributions to the integral. Hence the best solution is $s = 0$ independent of the value of the observation. (The fact that the best solution is independent of the data is, of course, an artifact of this particular example.)

You might wonder whether this result would disappear if we remove the noise from the problem by specifying that $x = sh$ so that the observation is a *deterministic* function of s, h. In such case the probability distribution becomes zero except on the line $x = sh$. Does the phase space contribution still apply?

The answer is that phase space is still important even in the case with no noise. To see this, observe that we can model the deterministic function $x = sh$ by a probability distribution $P(x|sh) = \delta(x - sh)$ where $\delta(.)$ is the Dirac delta function. Then the result can be obtained by observing that $\int dh\delta(x - sh) = \int d\hat{h}(1/s)\delta(x - \hat{h}) = 1/s$ where we have performed the change of variables $\hat{h} = sh$. Alternatively, we can derive the same result by a limiting argument. It can be shown that the Delta function can be expressed as the limit of a Gaussian distribution as the variance of the Gaussian tends to zero. From above, we see that the integral with respect to h gives a result $1/s$ which is independent of the variance σ^2. So as we take the limit $\sigma \mapsto 0$ the result is still $1/s$.

This example is admittedly extreme but it brings out the main point. When estimating s from $P(s|x) = \int dhP(s, h|x)$ we must take into account the phase space of the h variables.

A more interesting example occurs when we put a prior distribution $P(h)$ on the hidden variables h. To make life easy for ourself, we assume that $P(h)$ is a mixture of Gaussians so that we can get a nice analytic expression. So let us select:

$$P(h) = \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(h-h_1)^2/(2\sigma^2)} + \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(h-h_2)^2/(2\sigma^2)}. \tag{3.5}$$

We now compute $P(x|s) = \int dhP(x|s, h)P(h)$ to obtain:

$$P(x|s) = \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2(s^2+1)}}e^{-(h_1 s-x)^2/(2\sigma^2(s^2+1))} + \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2(s^2+1)}}e^{-(h_2 s-x)^2/(2\sigma^2(s^2+1))}. \tag{3.6}$$

This can be plotted, see figure (3.6). Observe that the prior for h biases the value of s to be close to the solutions $s = x/h_1$ and $s = x/h_2$. But the solution for smaller value of s has a larger amount of phase space, see the previous example, and so it has a higher peak.

3.3.1 Generalized Bas Relief Ambiguity Example

We now consider a more complicated and realistic example. In the Lambertian lighting model the intensity of a surface is given by:

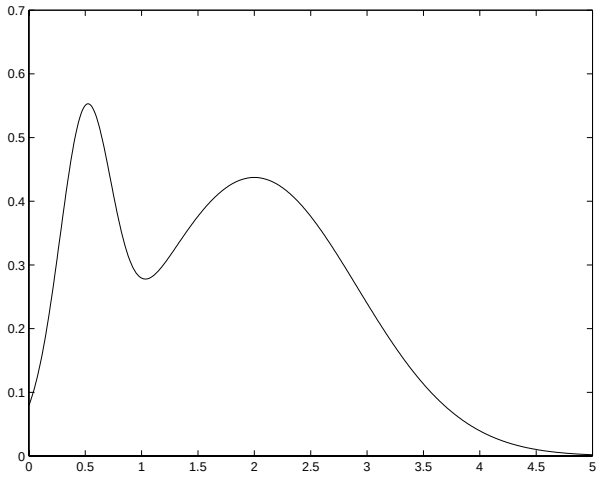

Figure 3.6 The marginal $P(x|s)$ as a function of s with $x = 1$. We set $h_1 = 0.5$ and $h_2 = 2.0$. Observe that there are two peaks but the higher one is for the smaller value of s.

$$I(x) = \vec{b}(x) \cdot \vec{s}, \tag{3.7}$$

where $\vec{b}(x) = a(x)\vec{n}(x)$, with $a(x)$ the surface albedo and $\vec{n}(x)$ the surface normal. The vector $\vec{s}$ is the light source strength and direction. This assumes a single light source and that there are no shadows, cast or attached, are present. (In the following subsection we will generalize to cases where there are attached shadows and also self-cast shadows). We let the size of the image (i.e. the number of pixels x) be N.

It has been shown (Belhumeur et al) that there is an ambiguity in these equations. So that we can perform the transformation $\vec{b}(x) \mapsto \mathbf{G}\vec{b}(x)$ and $\vec{s} \mapsto \mathbf{G}^{T,-1}\vec{s}$ where G is a 3×3 matrix represented a *generalized bas relief* (GBR) transformation (here ^{T} denotes matrix transpose and $^{-1}$ stands for matrix inverse), see figure (3.3). If two objects O_1, O_2 are related by a GBR, so that $\vec{b}_1(x) = \mathbf{G_{12}}\vec{b}_2(x)$ for some GBR $\mathbf{G}_{12}$, then for any illumination of object O_1 there will always be a corresponding illumination of object O_2 so that the images of the two objects are identical. It would therefore seem that there is no way of telling which object is present unless the illumination conditions are specified. A simple form of a GBR corresponds to scaling the object in depth to flatten it by an amount λ. Renaissance artists exploited human observer's relative insensitivity to such flattening by making "bas relief" sculpture which are flattened and hence need less material ("bas" is the French for "low").

We now analyze the effect of the phase space of generic views on this problem. As we will see, integrating out the lighting configurations will help resolve the ambiguity between the two objects.

We assume that the imaging model introduces independent Gaussian noise. The

probability models are therefore:

$$P(\{I(x)\}|O_1, \vec{s}_1) = \frac{1}{(2\pi\sigma^2)^{(N/2)}} e^{-\sum_x \{I(x) - \vec{b}_1(x)\cdot\vec{s}_1\}^2/(2\sigma^2)}, \tag{3.8}$$

$$P(\{I(x)\}|O_2, \vec{s}_2) = \frac{1}{(2\pi\sigma^2)^{(N/2)}} e^{-\sum_x \{I(x) - \vec{b}_2(x)\cdot\vec{s}_2\}^2/(2\sigma^2)}. \tag{3.9}$$

To determine the evidence for each model we must integrate out over the lighting configurations $\vec{s}_1$ and $\vec{s}_2$. Each of the likelihood functions can be re-expressed in form:

$$P(\{I(x)\}|O_1, \vec{s}_1) \frac{1}{(2\pi\sigma^2)^{(N/2)}} e^{-\{\sum_{\mu,\nu=1}^3 T_1^{\mu\nu} s_1^\mu s_1^\nu - 2\sum_{\mu=1}^3 s_1^\mu \phi_1^\mu + \psi\}/(2\sigma^2)}, \tag{3.10}$$

where $\mathbf{T}_1$ is a matrix with components $T_1^{\mu\nu} = \sum_x b_1^\mu(x) b_1^\mu(x)$, $\vec{\phi}_1$ is a vector with components $\phi_1^\mu = \sum_x b_1^\mu(x) I(x)$, and $\psi = \sum_x \{I(x)\}^2$. We are, of course, using the indices μ, ν to label the three spatial components of the vectors $\vec{b}(x)$.

The likelihood function is now quadratic in the variables $\vec{s}_1^\mu$ that we wish to integrate over. This integral can therefore be done by standard methods of completing the square. The result is given by:

$$\int d\vec{s} P(\{I(x)\}|O_1, \vec{s}_1) = \frac{1}{(2\pi\sigma^2)^{(N/2)}} \frac{(2\pi\sigma^2)^{3/2}}{|\det \mathbf{T}_1|^{1/2}} e^{-\{\psi - \sum_{\mu,\nu=1}^3 T_1^{-1\ \mu\nu} \phi_1^\mu \phi_1^\nu\}/(2\sigma^2)}. \tag{3.11}$$

A similar result can be obtained for integrating out $P(\{I(x)\}|O_2, \vec{s}_2)$ with respect to $\vec{s}_2$. It yields a similar formula with T_1, ϕ_1 replaced by T_2, ϕ_2 where $T_2^{\mu\nu} = \sum_x b_2^\mu(x) b_2^\mu(x)$, $\phi_2^\mu = \sum_x b_2^\mu(x) I(x)$ (the number $\psi = \sum_x \{I(x)\}^2$ is the same for both cases). To relate these results we recall that $b_1^\mu(x) = \sum_{\nu=1}^3 G_{12}^{\mu\nu} b_2^\nu(x)\ \forall\ x$. This leads to the relations $\phi_2^\mu = \sum_{\nu=1}^3 G_{12}^{\mu\nu} \phi_1^\nu$ and $T_2^{\mu\nu} = \sum_{\rho,\tau=1}^3 G_{12}^{\mu\rho} G_{12}^{\nu\tau} T_1^{\mu\rho}$. It is then straightforward algebra to check that

$$\sum_{\mu,\nu=1}^3 T_1^{-1\ \mu\nu} \phi_1^\mu \phi_1^\nu = \sum_{\mu,\nu=1}^3 T_2^{-1\ \mu\nu} \phi_2^\mu \phi_2^\nu \quad |\det \mathbf{T}_1| = |\det \mathbf{G}_{12}|^2 |\det \mathbf{T}_2|. \tag{3.12}$$

It is also straightforward algebra (exercise for the reader) to determine that $\psi - \sum_{\mu,\nu=1}^3 T_1^{-1\ \mu\nu} \phi_1^\mu \phi_1^\nu = \min_{\vec{s}} \sum_x \{I(x) - \vec{b}_1(x)\cdot\vec{s}_1\}^2$. We define this to be $E_{min}[\{I(x)\}]$. (Similar results apply for the second model). This gives:

$$\int d\vec{s}_1 P(\{I(x)\}|O_1, \vec{s}_1) = \frac{1}{(2\pi\sigma^2)^{(N/2)}} \frac{(2\pi\sigma^2)^{3/2}}{|\det \mathbf{T}_1|^{1/2}} e^{-E_{min}[\{I(x)\}]/(2\sigma^2)}$$

$$\int d\vec{s}_2 P(\{I(x)\}|O_2, \vec{s}_2) = \frac{1}{(2\pi\sigma^2)^{(N/2)}} \frac{(2\pi\sigma^2)^{3/2}}{|\det \mathbf{G}_{12}||\det \mathbf{T}_1|^{1/2}} e^{-E_{min}[\{I(x)\}]/(2\sigma^2)}. \tag{3.13}$$

So we see that, after integration, the two hypotheses are not equally likely. The difference is the factor $|\det \mathbf{G}_{12}|$ in the denominator. This says that of two hypotheses $\vec{b}_1(x), \vec{b}_2(x)$ related by $\vec{b}_2(x) = \mathbf{G}\vec{b}_1(x)$ we prefer $\vec{b}_1(x)$ if $|\det \mathbf{G}_{12}| > 1$ and $\vec{b}_2(x)$ otherwise. Now the determinant of a GBR is given by λ where the transformation scales the z-axis by λ. So if O_2 is enlarged in the z direction (i.e. $\lambda > 1$ relative to O_1, then we prefer O_1. So of the two possible hypotheses we prefer the most flattened one! (Why? Well, intuitively if the object is flat then its appearance is based on its albedo and is largely independent of lighting conditions – so it is very stable under lighting changes).

3.3.2 Symmetry and Generic Views

What happens if we have cast or attached shadows?[3] And multiple light sources? We now show that the basic results of the previous subsection, namely the dependence of the determinant of the GBR, remain unchanged.

For this more realistic, and important, case it will be impossible to perform the integrals to marginalize out the lighting conditions. Instead, we present a new method which takes advantage of the *symmetry* of the problem to deduce the results for the two object classification *without needing to evaluate the integral.*

Symmetry has been present in the examples in the previous two subsections. In both cases there is an ambiguity between the variables s, h which, in mathematical terms, is a symmetry. For the abstract example where $P(x|s,h) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-sh)^2/(2\sigma^2)}$ we have the symmetry $h \mapsto \lambda h$ and $s \mapsto (1/\lambda)s$ for any λ. For the GBR example we have the transformation $\vec{b}(x) \mapsto \mathbf{G}\vec{b}(x)$ and $\vec{s} \mapsto \mathbf{G}^{T,-1}\vec{s}$ where $\mathbf{G}$ is a member of the GBR group.

The key point is that when evaluating the evidence of two models related by a symmetry transformation then the relative evidence *depends only on the symmetry transformation itself.* To put it another way, the symmetry of the problem is broken by the phase space contribution.

To illustrate this point, we extend our analysis of the GBR to include attached and cast shadows. To take into account attached shadows we write the illumination equation as $I(x) = \max\{\vec{b}_1(x) \cdot \vec{s}^1, 0\} + \max\{\vec{b}_1(x) \cdot \vec{s}^2, 0\} + ... + \max\{\vec{b}_1(x) \cdot \vec{s}^M, 0\}$, where the maximum operation removed points x where $\vec{b}(x) \cdot \vec{s} \leq 0$ (these correspond to attached shadows). To allow for cast shadows, we also set the contribution from light source $\vec{s}^i$ to be zero at a point x if the light is blocked in reaching that point. We define a cast shadow function $f_1(x; \vec{s}^i)$ which is zero if point x on object 0_1 lies in a cast shadow under lighting condition $\vec{s}^i$ and equals 1 otherwise. It is known (Belhumeur et al) that the GBR ambiguity holds even when cast and attached shadows are present.

We now obtain

[3] It is known that Leonardo da Vinci wrongly believed that shadows would not be invariant under bas relief transformations.

$$P(\{I(x)\}|O_1, \vec{s}^1, ..., \vec{s}^M) = \frac{1}{(2\pi\sigma^2)^{(N/2)}} e^{-\sum_x \{I(x) - \sum_{i=1}^M f_1(x:\vec{s}^i) \max\{\vec{b}_1(x)\cdot\vec{s}^i, 0\}\}^2/(2\sigma^2)}. \tag{3.14}$$

The evidence for model O_1 is therefore given by the integral:

$$K[\{I(x)\}] = \int d\vec{s}^1 d\vec{s}^2 ....d\vec{s}^M P(\{I(x)\}|O_1, \vec{s}^1, ..., \vec{s}^M) = \int d\vec{s}^1 d\vec{s}^2 ....d\vec{s}^M \frac{1}{(2\pi\sigma^2)^{(N/2)}} e^{-\sum_x \{I(x) - \sum_{i=1}^M f_1(x:\vec{s}^i) \max\{\vec{b}_1(x)\cdot\vec{s}^i, 0\}\}^2/(2\sigma^2)}. \tag{3.15}$$

It is impossible to calculate this integral analytically. But we do not need to! We only need to compare its value to that of the evidence for model O_2. This can be done by observing that to compute the evidence for O_2 we merely have to replace $\vec{b}_1(x)$ by $\vec{b}_2(x)$ in the exponent. These are related by a GBR $\vec{b}_2(x) = \mathbf{G}_{12}\vec{b}_1(x)$. Now we perform a change of variables (this is the clever bit) so that $\vec{\hat{s}}^i = \mathbf{G}_{12}^{T,-1}\vec{s}^i$ for all i. With this change of variables the exponent is now the same whether we are computing the integral for model O_1 or O_2! But changing the variables means that we have to introduce a Jacobian factor in the integral. The factor, of course, is simply $|\det \mathbf{G}_{12}|^M$. So, the difference in evidence between the two models is given only by this factor. Setting $M = 1$ recovers our original result. But now we have extended it to deal with cast and attached shadows and multiple light sources. Once again, there will be a tendency to favour "flatter" surfaces if possible.

What happens if we have a prior distribution on the objects? This does not alter the conclusions greatly. We integrate $P(\{I(x)\}|O_1, \{\vec{s}^i : i = 1, ...M\})P(\{\vec{b}_1(x))\})$ with respect to $\{\vec{s}^i : i = 1, ...M\}$. The point is that the prior $P(\{\vec{b}_1(x))\})$ is independent of $\{\vec{s}^i : i = 1, ...M\}$ and so can be taken outside the integral. This gives:

$$\log\{P(\{I(x)\}|O_1)P(\{\vec{b}_1(x))\})\} = -M \log|\det \mathbf{G}_{12}| + \log K[\{I(x)\}] + \log P(\{\vec{b}_1(x))\}). \tag{3.16}$$

Recall that the term $\log K[\{I(x)\}]$ is independent of the GBR. So the two important terms are the first term which encourages the object to be flat and the final term which pulls it towards the prior. This means that the interpretation is pulled towards the most probable *a priori* interpretation desired by the prior and the flat interpretation determined by the generic factor. The conclusion is that there is an overall bias towards flatter objects unless the prior is incredibly strongly peaked (i.e. almost a delta function).

Overall, we see that integrating over the lighting conditions "breaks" the GBR ambiguity. Observe, moreover, that it induces a bias towards surfaces which are flat which is against the spirit of bas relief in art where one tries to use a pattern with only small relief to substitute for a pattern with large relief. (Of course, this effect is not very strong if the scaling λ of the transformation is close to one). We suggest that the effectiveness of

bas relief is because of prior expectations on shape and albedo. Such priors can only be partially effective, however.

3.3.3 Approximating the marginals: Gaussians approximations and Laplace's Method

In most cases it is impossible to integrate out the hidden variables analytically. In such cases it is often desirable to approximate the integral. One important class of approximations falls under the mathematical heading of *asymptotic expansions*. They are only rigourously correct in certain precise limits, to be discussed later, but they nevertheless often give good approximations in other situations. In this section we will describe Laplace's method which has been called the "workhorse of asymptotic expansions" (Keener). It is closely related to other methods such as saddle point expansions and the method of stationary phase. (We note that in this section we are describing general purpose techniques only and that for certain types of problem there are more effective methods which may not even require approximations, see later chapters).

A second, closely related, technique is to approximate the integrand by one, or more, Gaussians. In some cases, as we will discuss, this gives identical results to Laplace's method. This approximation is less well justified, in general, but it is intuitive and is applicable in situations where Laplace's method is not justified.

The simplest version of Laplace's method occurs when we need to evaluate an integral of form:

$$f(\alpha) = \int_{-\infty}^{\infty} e^{\alpha h(z)} g(z) dz, \tag{3.17}$$

where we are interested in the behaviour of α for *large* α. This is known as an *asymptotic expansion* and the expansion is only provably correct in this limit although it may be a good approximation in other situations. In the case of large α it becomes legitimate to expand $h(z)$ in a Taylor series about the value z^* which maximizes $h(z)$. It can then be proven, by *Watson's Lemma* (see Keener), that one can obtain an asymptotic series expansion for $f(\alpha)$ which is valid in the limit as $\alpha \mapsto \infty$. More concretely, we expand $h(z) = h(z^* + (1/2)(z - z^*)^2 (d^2h/(dz^2))(z^*) + O((z - z^*)^3)$. The first order term in the Taylor series expansion vanishes because $(dh/(dt))(z^*) = 0$ (since z^* is a maximum). We denote $(d^2h/(dz^2))(z^*)$ by h^*_{zz} for brevity and observe that it is a negative number (because z^* is a maximum). The expansion gives:

$$\begin{aligned} f(\alpha) &\approx \int_{-\infty}^{\infty} e^{\alpha h(z^*)} e^{\alpha h^*_{zz}(t-z^*)^2} g(z^*) dt \\ &= g(z^*) e^{\alpha h(z^*)} \frac{\sqrt{\pi}}{\sqrt{|\alpha h^*_{zz}|}}. \end{aligned} \tag{3.18}$$

and other higher order terms can be obtained (see Keener).

Essentially what we are doing is approximating the integral by a Gaussian distribution (higher order terms in the expansion will go beyond this approximation). But *note we are doing this for large α so we can neglect the terms from $g(z)$*. If α is not large then we must do a series expansion in the function $g(z)$ as well. For now we treat the asymptotic (i.e. large α) case only. This requires first finding the value of z that maximizes the exponent. Then we do a quadratic expansion in the exponent. This turns the integrand into a Gaussian which we can calculate analytically.

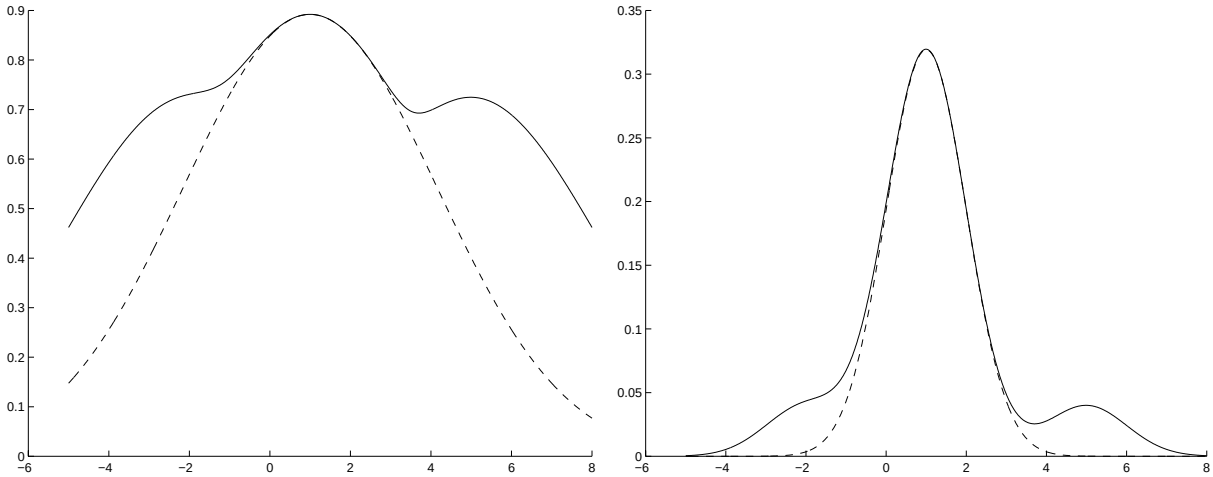

Figure 3.7 A bad Laplace approximation (left) and a better one (right).

The expansion becomes precise in the limit as $\alpha \mapsto \infty$. For finite values of α it remains an approximation. Its usefulness will depend on the form of the function $e^{\alpha h(z)}g(z)$. In this case, we must expand the $g(z)$ term as well. If the function $e^{\alpha h(z)}g(z)$ can be well approximated by a Gaussian then Laplace's method will yield good results. But the results will be poor if, for example, the integrand has multiple peaks, see figure (3.7). In such cases, it would be best to approximate the integrand by a sum of Gaussian distributions centered about each maxima of the integrand (this, of course, can become complicated).

For this section, we will mainly be concerned with using Laplace's method to approximate integrating out hidden variables. More precisely, we will want to compute:

$$P(x|s) = \int dh P(x|h,s)P(h) = \int dh P(h) e^{\log P(x|h,s)}. \tag{3.19}$$

It may occur that the distribution $P(x|h,s)$ is of form:

$$P(\vec{x}|\vec{h},\vec{s}) = \frac{1}{(\sqrt{2\pi\sigma^2})^N} e^{-||\vec{x}-\vec{f}(\vec{h},\vec{s})||/(2\sigma^2)}, \tag{3.20}$$

where we are assuming independent Gaussian noise for all the N pixels. The function $\vec{f}(\vec{h},\vec{s})$ details how the hidden variables $\vec{h}$ and the state variables $\vec{s}$ combine to form the

image.

We can now apply Laplace's method by expanding the function $\vec{f}(\vec{h},\vec{s})$ in a Taylor series about the value $\vec{h}^*$ which maximizes $\vec{f}(\vec{h},\vec{s})$ (observe that we will get different expansions depending on the the value of $\vec{s}$ and, in particular, $\vec{h}^*$ is a function of $\vec{s}$.) The expansion is only fully justified in the limit of small σ^2 (i.e. we have set the x in Laplace's expansion equal to $1/(\sigma^2)$. The expansion can be written as:

$$\vec{f}(\vec{h},\vec{s}) \approx \vec{f}(\vec{h}^*,\vec{s}) + \frac{1}{2}(\vec{h}-\vec{h}^*)^T \vec{\mathbf{f}}^{\prime\prime,\ *}(\vec{h}-\vec{h}^*) + 0(|\vec{h}-\vec{h}^*|^3), \tag{3.21}$$

where $\vec{\mathbf{f}}^{\prime\prime,\ *}$ is the Hessian of $\vec{f}(\vec{h},\vec{s})$ with respect to the variables $\vec{h}$ and evaluated at $\vec{h}^*$.

By using Laplace's approximation, for fixed $\vec{s}$, we obtain:

$$P(\vec{s}|\vec{x}) \approx P(\vec{h}^*)e^{-||\vec{x}-\vec{f}(\vec{h}^*,\vec{s})||/(2\sigma^2)}\frac{1}{\sqrt{\det \mathbf{C}}}, \tag{3.22}$$

where $\mathbf{C}$ is the Hessian of $||\vec{x}-\vec{f}(\vec{h}^*,\vec{y})||$ with respect to $\vec{h}$ evaluated at $\vec{h}^*$. More precisely, we can write $||\vec{x}-\vec{f}(\vec{h}^*,\vec{s})|| = \sum_a\{x_a - f_a(\vec{h},\vec{s})\}$. The Hessian $\mathbf{C}$ has components given by $\partial^2/(\partial h_i \partial h_j)||\vec{x}-\vec{f}(\vec{h}^*,\vec{s})||$ which can be evaluated to be:

$$\frac{\partial^2}{\partial h_i \partial h_j}||\vec{x}-\vec{f}(\vec{h}^*,\vec{s})|| = \frac{1}{\sigma^2}\{\sum_a \frac{\partial f_a}{\partial h_i}\frac{\partial f_a}{\partial h_j} + \sum_a\{f_a(\vec{h},\vec{s})-x_a\}\frac{\partial^2 f_a}{\partial h_i \partial h_j}\}, \tag{3.23}$$

evaluated at $\vec{h}=\vec{h}^*$.

The term $\sqrt{\det \mathbf{C}}$ is called the generic viewpoint term (Freeman).

For certain vision problems there is a natural parameter x for which and it is known that we are only interested in the behaviours for large x. In these conditions the approximation becomes justified. In other cases it needs to be empirically verified by computer simulations.

What happens if we are not in the asymptotic region (i.e. low noise case for this example)? Then we can do a Gaussian approximation to the entire integrand. Suppose we have

$$\log P(x) = \log \int dh P(x|h)P(h). \tag{3.24}$$

Then we write the integral in form:

$$\log P(x) = \int dh e^{E(h;x)}, \tag{3.25}$$

where $E(h;x) = -\log P(x|h) - \log P(h)$.

The maximum of $P(x|h)P(h)$ occurs at the minimum of $E(h;x)$. Therefore, to determine h^*, we solve $(\partial/\partial h)E(h;x)(h^*) = 0$ with the constraint that the Hessian $\mathbf{H}(\mathbf{h}^*;\mathbf{x})$ is positive definite (i.e. that h^* is a true minimum of $E(h;x)$.)

We can them perform an approximation by doing a quadratic expansion about h^*. This sets:

$$E_A(h;x) = E(h^* : x) + (1/2)(h - h^*)^T \mathbf{H}(\mathbf{h}^*)(h - h^*). \tag{3.26}$$

The integral $\int dh e^{-E_A(h;x)}$ can now be performed exactly because it is a Gaussian distribution in h. This gives:

$$\int dh e^{-E_A(h;x)} = (2\pi)^{n/2} \det \mathbf{H}(h^*) e^{-E(h^*;x)}. \tag{3.27}$$

As described, the method is an approximation. It breaks down if, for example, there are multiple minima of the energy function $E(h;x)$. It also only takes into account the second order terms in the Taylor series expansion of $E(h;x)$ around h^*. Extending the method to take into account higher order terms is significantly harder.

MENTION SCHRATER AND KERSTEN??

3.4 Discrete Hidden Variables

Many important problems occur when the hidden variables are discrete. They may correspond, for example, to binary variable which label different models for explaining the data (i.e. the data might be due to models A or B and we do not know which). Alternatively, they may label "outliers" in visual search tasks.

Some new techniques are required when dealing with discrete variables. It is no longer possible, for example, apply Laplace's method to approximate over them. There are, however, other approaches such as mean field theory approximations which perform similar types of approximation. Some of these methods are beyond the scope of this book and we will refer to them elsewhere. (For example, it is possible to transform discrete problems into continuous ones and then apply Laplace's method directly – see Hertz, Krogh, Palmer book. Or Yuille review article in Arbib).

The basic setup is as follows. There are state variables $s \in S$ that we wish to estimate. There are data observations $x \in X$. Finally there are hidden state variables $V \in H$ which are discrete. We have probability models $P(x, V|s)$ and we want to estimate the state s by summing out the hidden variables V.

We present this material in the following sections by treating several important cases.

3.4.1 Signal Known Exactly ModelS

In the previous chapter, we discussed the Signal Known Exactly (SKE) model. We now describe a version of it where there the signal can come in several different variants. For

example, the basic signal can be a sinusoid and the variants correspond to changes of phase. We can now ask two questions: (i) does an input signal correspond to signal or noise?, and (ii) if it is a signal, then what is its phase? Experiments can be designed in which the first task is possible and yet the second task is not (REFS FROM DAN!!).

We define a set of signal models $S_i(x) = A\cos(\alpha x + \phi_i)$, where A, α are constants and the set of $\{\phi_i : i = 1, ..., M\}$ give the M different phases that the signal can take.

As in the original SKE case, we define models for the probabilities of the observed images $I(x)$ conditioned on the signals and the noise. There is assumed to be a background intensity $B(x)$ which is spatially constant (i.e. $B(x) = B,\ \forall\ x$). The models assume additive Gaussian noise (independent at each pixel) and so we have models:

$$P(\{I(x)\}|S_i) = \prod_{x=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(I(x)-S_i(x)-B)^2/(2\sigma^2)},\ i = 1, ..., M$$
$$P(\{I(x)\}|N) = \prod_{x=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(I(x)-B)^2/(2\sigma^2)}. \tag{3.28}$$

We also define prior distributions:

$$P(N) = \frac{1}{2},\ P(S_i) = \frac{1}{2M},\ i = 1, ..., M. \tag{3.29}$$

We now compute the posterior distributions:

$$P(S_i|\{I(x)\}) = \frac{P(\{I(x)\}|S_i)P(S_i)}{P(\{I(x)\})},$$
$$P(N|\{I(x)\}) = \frac{P(\{I(x)\}|N)P(N)}{P(\{I(x)\})}. \tag{3.30}$$

Our first task was to determine whether an input is either signal or noise. In this case, we must sum over the probability that the input data is generated by each of the models S_i. We can then define a new variable S which determines whether the signal is there or not. The $\{S_i\}$ can now be considered to be hidden variables. We have:

$$P(S|\{I(x)\}) = \sum_{i=1}^{M} P(S_i|\{I(x)\}) = \frac{\sum_{i=1}^{M}\{P(\{I(x)\}|S_i)P(S_i)\}}{P(\{I(x)\})}. \tag{3.31}$$

The decision, as to whether there is a signal present, is determined by the log-likelihood ratio of $P(S|\{I(x)\})$ to $P(N|\{I(x)\})$. This can be written as:

$$\log \frac{P(S|\{I(x)\})}{P(N|\{I(x)\})} = \log\{\sum_{i=1}^{M} \frac{1}{M} \frac{P(S_i|\{I(x)\})}{P(N|\{I(x)\})}\}. \tag{3.32}$$

Conversely, if we are studying the second task of determining which specific signal (i.e. which phase) is present then we must do a different analysis. We must compare the values of $P(S_i|\{I(x)\})$ *for each* $i = 1, ..., M$ to the value of $P(N|\{I(x)\})$, and to each other.

Suppose in both cases, we are performing the MAP estimation. Then we should decide whether the input is signal or noise depending on whether:

$$\sum_{i=1}^{M} P(S_i|\{I(x)\}) \geq P(N|\{I(x)\}). \tag{3.33}$$

But to do the second task requires selecting the maximum of $M + 1$ numbers:

$$P(S_1|\{I(x)\}), P(S_2|\{I(x)\}), ..., P(S_M|\{I(x)\}), P(N|\{I(x)\}). \tag{3.34}$$

Clearly, it is quite possible that $\sum_{i=1}^{M} P(S_i|\{I(x)\}) \geq P(N|\{I(x)\})$ but $P(S_i|\{I(x)\}) < P(N|\{I(x)\})$ $\forall\ i = 1, ..., M$. In this case, the question of whether it is signal to noise can be answered. But the "evidence" for signal requires combining the evidence for different variants of the signal (i.e. different phases) and no individual S_i has enough evidence by itself to defeat the noise hypothesis.

Cases where "the whole is greater than the maximum of the parts" require than two, or more, individual signal hypotheses make non-negligible contributions to the total evidence. This means that for any $\{I(x)\}$ that is classified as being "signal" we need to have at least two i, j such that $P(S_i|\{I(x)\}) \neq 0$ and $P(S_j|\{I(x)\}) \neq 0$. This implies that there is an overlap between the individual signal responses and hence it may be hard to distinguish between them, even by a one on one experiment.

3.4.2 Robustness and Outliers

One of the simplest example of hidden variables is the need for rejecting outliers in the data. The data can consist either of simple measurements or be as complicated such as estimates of depth.

Outliers are data that do not fit the probability model which is assume to generate the data. Suppose, for example, we want to estimate the mean of a set of variables $\{x_1, ..., x_N\}$. The standard estimator is to set $\mu(x) = \frac{1}{N}\sum_{i=1}^{N} x_i$. This estimator can be derived as the ML estimator for the mean assuming that the data is generated by a Gaussian model. In other words, we assume that the data is generated by a distribution $P(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$. If the data is identically independently distributed , so that $P(\{x_1, ..., x_N\}|\mu\sigma) = \prod_{i=1}^{N} P(x_i|\mu, \sigma)$, then it is a straightforward application of ML to obtain the estimator $T(x_1, ..., x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i$.

A problem arises if some of the data samples are *outliers* which are not generated by the Gaussian distribution. Outliers could result arise because the data is contaminated in some way. Or, perhaps more commonly, because the probability model used to analyze

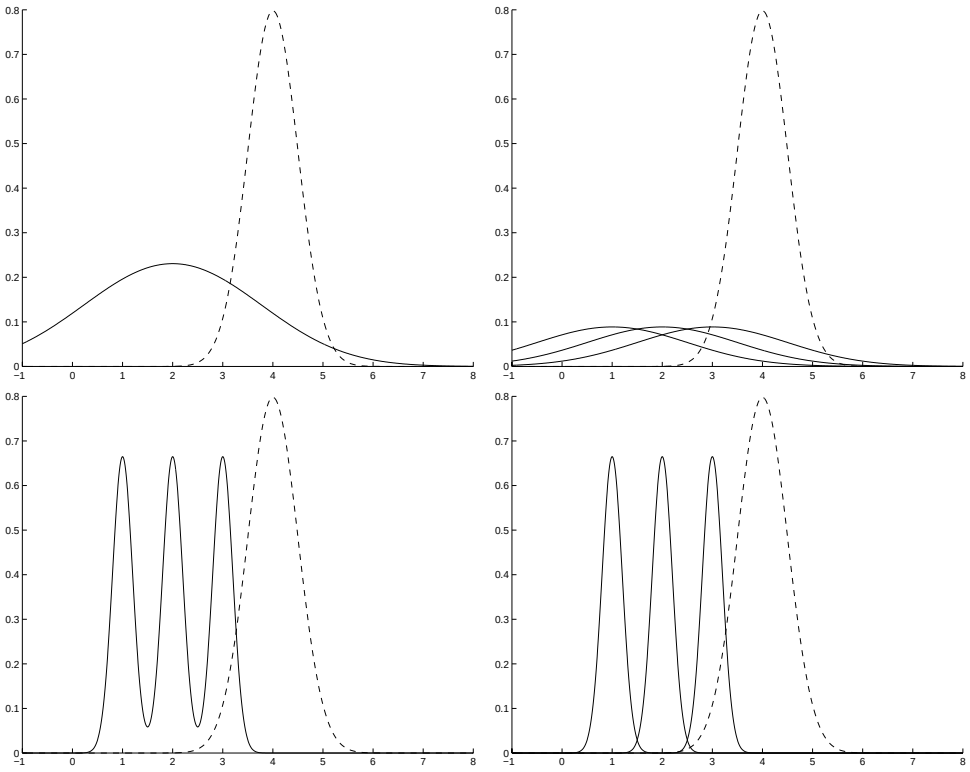

Figure 3.8 Top, in this example it is significantly more difficult (left) to distinguish noise (dashed line) from the *set of signals* (solid line) than (right) from any one signal (solid lines) (signals have $\sigma_s = 1.5$ and noise $\sigma_n = 0.5$). Bottom, in this case it is *not* significantly harder to distinguish the noise from the set of signals (left) than from the "closest" signal (right) (signals have $\sigma_s = 0.2$ and noise $\sigma_n = 0.5$).

the data is, at best, an approximation to the (unknown) true probability distribution (we discuss the issues of how well we can learn probability distributions from data in a later chapter). In both cases errors can arise from estimation because of the outliers contaminating the data.

A whole subfield of statistics, known as *Robust Statistics*, has developed to analyze this phenomena. We refer readers to Huber for theory. The type of robust statistics we describe here, including the use of hidden variables, is not standard but seems, to us, to be most appropriate for vision (and is in keeping with the spirit of this book!). See Berger.

An important application of robust methods and/or outlier detection is to the coupling of different visual cues. It sometimes happens that two visual cues, depth cues for example, give such different estimates that they are mutually inconsistent. In such a case one cue appears to "veto" the other. This can be considered to be a case of robust estimation with one of the cue values being treated as an outlier (Landy et al.). We will discuss coupling visual cues in a later chapter from a Bayesian perspective.

Another example, comes when attempting to match an object model to an image. Consider the Signal Known Exactly model described in the previous chapter. The SKE model assumes a template for the target signal and models the noise in the image as Gaussian. This assumption is fine in a laboratory environment where the stimuli are controlled (and much insight into the visual system can be obtained by studying such a model). However, the types of noise that occurs in real world stimuli are not always Gaussian. For example, in the display reading task, see figure!!, the difficulty in reading the display is due to the presence of specularities in the image. Such specularities do not satisfy the independent Gaussian noise assumption because they tend to be spatially localized. Robust techniques, however, can be used to give models that are less sensitive to specularities...

FIGURE ON DISPLAY READER – SPECULARIES AS NON-GAUSSIAN.

We now introduce some mathematics. Let us assume that the data x comes from one of two models $P_0(x|s_0), P_1(x|s_1)$. For concreteness, we can assume that these distributions are both Gaussians $P_0(x|\mu_0, \sigma_0) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(x-\mu_0)^2/(2\sigma_0^2)}$ and $P_1(x|\mu_1, \sigma_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-(x-\mu_1)^2/(2\sigma_1^2)}$.

Now let V be a binary indicator variable which takes two states $\{0, 1\}$. If $V = 0$ means that the data x is generated by the model $P_0(x)$ (corresponding to s_0) and $V = 1$ means it is generated by $P_1(x)$ (corresponding to s_1). These variables are "hidden" in the sense that they are unknown to the observer.

We can then write down a distribution:

$$P(x|V) = \{P_0(x)\}^{1-V}\{P_1(x)\}^V, \tag{3.35}$$

which implies that if $V = 1$ then the data is generated by $P_1(x)$ (i.e. $P(x|V = 1) = P_1(x)$) and $V = 0$ means that the data is generated by $P_0(x)$ (which equals $P(x|V = 0)$.

We also specify a prior distribution $P(V)$ on the hidden variables to take into account our prior knowledge of how probable it is *a priori* that the data comes from models $P_0(x)$ or $P_1(x)$. For example, we can write $P(V = 0) = 1 - \epsilon$ and $P(V = 1) = \epsilon$ for a constant ϵ. This can be expressed concisely as:

$$P(V) = (1 - \epsilon)^{1-V}(\epsilon)^V. \tag{3.36}$$

We can then write the full distribution:

$$P(x, V) = P(x|V)P(V) = \{P_1(x)\}^V\{P_0(x)\}^{1-V}(1 - \epsilon)^{1-V}(\epsilon)^{1-V}. \tag{3.37}$$

In this case, we can explicitly sum out the hidden variables analytically and compute the marginal distribution for the data x:

$$P(x) = \sum_{V=0,1} P(x,V)P(V) = (1-\epsilon)P_0(x) + \epsilon P_1(x), \tag{3.38}$$

which is a *mixture of probability distributions.*

Alternatively, however, we may wish to estimate the variables V assuming that we know the state variables s_0, s_1. In other words, we assume that the data is generated by a mixture of models and we wish to estimate which model generated the data. This requires computing $P(x, V = 0)$ and $P(x, V = 1)$. By standard decision theory, see previous chapter, we compute the log posterior ratio:

$$\log \frac{P(x,V=0)}{P(x,V=1)} = \log \frac{P(x|V=0)}{P(x|V=1)} + \log \frac{1-\epsilon}{\epsilon}, \tag{3.39}$$

and choose a threshold to make the decision (the threshold will be zero is we use MAP estimation).

One point to emphasize here is that even if we can, or would like to, sum out the hidden variables there may nevertheless be situations where we want to estimate them explicitly. ("One person's hidden variables are another person's state variables".)

We now illustrate one of the main points of robust statistics: namely how much does using the wrong model penalize us? This penalty is in terms of the accuracy of the estimates. In the example we assume that the data is generated by a mixture of Gaussian models, as above. The distributions have the same mean μ and variances σ^2 and $9\sigma^2$ respectively (i.e. $\mu_0 = \mu_1 = \mu$ and $\sigma_0 = \sigma$ and $\sigma_1 = 3\sigma$).

For Gaussian distributions it is straightforward to compute the mean and variance of the mixture distribution. The means is given by μ and the variance is $(1-\epsilon)\sigma^2 + \epsilon 9\sigma^2$. So even if ϵ is only 10% the variance estimated from the data is twice the true variance of the distribution $P_0(x|s_0)$. See Huber for a more detailed discussion of how the contamination of Gaussians degrades the performance of statistical estimators.

We now describe another robust method which, in a more complex form, appears in a number of computer vision models. Suppose we want to estimate the mean of a number of samples but we know that some samples have been contaminated. We introduce the method by writing an energy function:

$$E(\{V_i\}, s; \{x_i\}) = \sum_{i=1}^{N} V_i (x_i - s)^2/(2\sigma^2) + \sum_{i=1}^{N} \lambda(1 - V_i) \tag{3.40}$$

Here s is a continuous variable, $\{V_i\}$ are binary $\{0, 1\}$ variables, and $\{x_i\}$ are the measurement data. The constants σ^2, λ are assumed to be known. We can consider the energy to be the sum of N energy terms $E_i(V_i, s; x_i) = V_i(x_i - s)^2/(2\sigma^2) + \lambda(1 - V_i)$.

Now consider the function $\hat{E}_i(s; x_i) = min_{V_i} E_i(V_i, s; x_i)$. This function is quadratic in s for $|x_i - s| \leq \sqrt{2\lambda\sigma^2}$ and takes a fixed value of λ for $|x_i - s| \geq \sqrt{2\lambda\sigma^2}$. For fixed s

we pay a "penalty" $(x_i - s)^2/(2\sigma^2)$ for $|x_i - s| \leq \sqrt{2\lambda\sigma^2}$ and a maximum penalty of λ otherwise. See figure (3.10).

As we try to minimize the total energy $E(\{V_i\}, s; \{x_i\})$ over all the variables $s, \{V_i\}$ we are faced with a tradeoff: it will usually not be possible to find adjust the s to be close to all the data points $\{x_i\}$. For some data points x_j it will be necessary to "reject them" by setting $V_j = 0$ and paying the rejection penalty λ.

To see why we call this approach "robust", we can contrast it with the alternative approach of removing the $\{V_i\}$ variables and simply setting $E[s; \{x_i\}] = \sum_{i=1}^{N}(x_i - s)^2/(2\sigma^2)$. We see that minimizing this energy with respect to s occurs by setting $s = (1/N)\sum_{i=1}^{N} x_i$, which is the empirical mean of the data points. Observe that an outlier data-point pays a penalty $(x_i - s)^2/(2\sigma^2)$ which increases quadratically. By contrast the energy function $E(\{V_i\}, s; \{x_i\})$ is far more tolerant – outlier points pay a penalty which is quadratic if they are sufficiently close to s but which reaches a maximum of λ. Thus outliers have far less effect because they can be rejected without paying exorbitant costs.

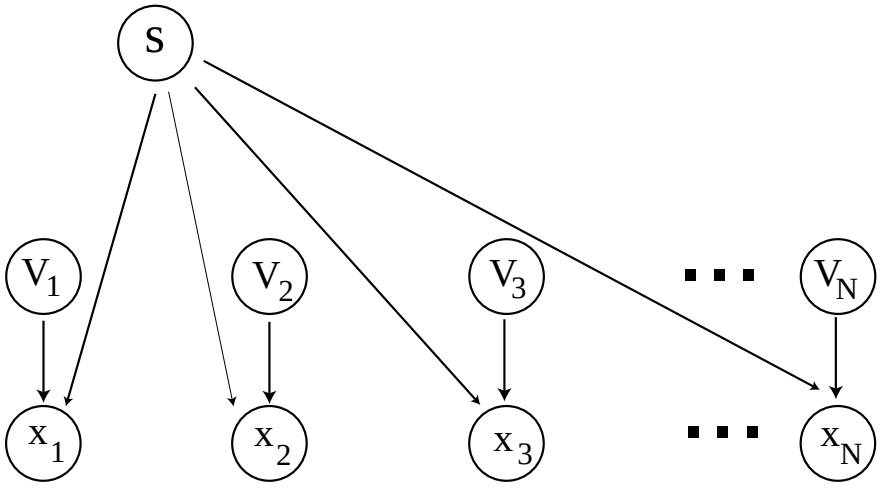

Figure 3.9 The bayes net representing the generative model behind the robust estimation of the mean.

We now put this analysis into probabilistic terms, see figure (3.9), by writing a probability distribution:

$$P(\{V_i\}, \{x_i\}|s) = \frac{1}{Z} e^{-E(\{V_i\}, s; \{x_i\})}. \tag{3.41}$$

In this formulation, the most probable states are those with lowest energy. *But there is a problem*: the distribution $P(\{V_i\}, s|\{x_i\})$ is not normalizable (i.e. to ensure that $\sum_{\{V_i\}} \int ds P(\{V_i\}, s|\{x_i\}) = 1$ would require setting Z to be infinite) and the expression should be treated as formal only. To understand this, consider deriving the "distribution" for s by summing out the $\{V_i\}$. This can be done yielding:

$$P(\{x_i\}|s) = \frac{1}{Z}\prod_{i=1}^{N}\{e^{-(s-x_i)^2/(2\sigma^2)} + e^{-\lambda}\}. \tag{3.42}$$

This can be interpreted as saying that each variable x_i is generated by a mixture of a Gaussian distribution and a uniform distribution. The trouble is that the uniform distribution cannot be normalized.

This problem can be fixed by putting the problem in a box. This means we replace the scalar λ by a function $b(s, x_i)$ chosen so that $b(s, x_i) = \lambda$ for $|x_i - s| < B$ and $b(s, x_i) = 0$ otherwise. We simply choose the "box size" B to be larger than the range of the samples $\{x_i\}$ that we get. Then, effectively, we can replace $b(s, x_i)$ by λ but still have a normalized distribution.

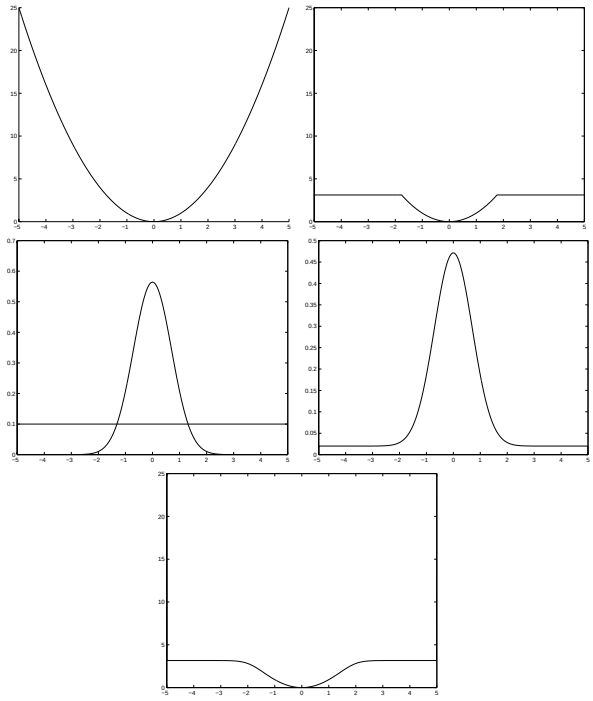

Figure 3.10 Top left, the quadratic energy function (for the Gaussian). Top right, the quadratic with a cut off corresponding to a rejection penalty λ. Middle left, the corresponding Gaussian and uniform distributions. Middle right, the mixture of the two distributions. Bottom, the effective energy function corresponding to the mixture distribution.

Certain distributions are very sensitive to contamination from outliers. The Gaussian distributions are particularly sensitive, which is unfortunate given their popularity. This can be quantified by the use of concepts from robust statistics such as influence functions (Huber). The problem arises because the "tails" of the Gaussian distribution fall off very fast. This says that data more than two standard deviations away from the mean is extremely unlikely and if such data arises, due to contamination, then it will distort

the estimates severely. It is sometimes better, in practice, to approximate a Gaussian by a *t distribution* which has tails that fall off more slowly (Ripley). This distribution is parameterized by a mean μ, a covariance $\nu\Sigma/(\nu-1)$, and two parameters ν, p. Its probability density is given by:

$$\frac{\Gamma(1/2(\nu+p))}{(\nu\pi)^{p/2}\Gamma(1/2\nu)}|\Sigma|^{-1/2}\{1+(1/\nu)(x-\mu)^T\Sigma^{-1}(x-\mu)\}^{-(1/2)(\nu+p)}. \tag{3.43}$$

3.4.3 Visual Search

We now address the visual search problem of detecting an outlier sample among a set of samples. This problem has been much studied and reported in the *visual search* literature and is sometimes called *pop-out*. Our purpose here is to treat the problem as an example of statistical inference in the presence of *hidden variables* which can correspond, for example, to binary labels for each sample which determine whether the sample is an outlier or not. We will not, in this chapter, be concerned with issues such as search strategy and reaction times (though we will say something about this in a later chapter).

Models of the type we will describe have been developed by several authors (Palmer, Verghese and Pavel) who have made explicit comparisons to experimental data. Issues that arise, both in experiments and theory, are whether there is *asymmetry* in the pop-out (i.e. whether detecting an A in a background of B's is easier than detecting a B in a background of A's), and whether *conjunctions of features* makes the pop-out task easier or harder. Another concern is how does *familiarity with the stimuli* affect performance of either expert or non-expert subjects. These issues will be discussed as we proceed.

3.4.4 Basic Bayes for Pop Out: Known Distributions

In this section we assume that the data samples are generated from one of two *known* probability distributions $P_A(.)$ or $P_B(.)$ and the Bayesian estimators are derived based on this assumption. In the next section, we will discuss situations where the probability distributions are unknown (although this situation is less clear cut).

Suppose we have a set of samples $\{x_1, ..., x_{N+1}\}$. We consider two visual tasks. The first task is to detect if there is an outlier in the samples. The second task, which assumes an outlier is present, is to determine which sample it is. The third is to detect if there is an outlier in the samples *and if so* where it is.

For the first task, we consider 2AFC where one stimulus consists of $N+1$ samples generated by $P_B(.)$. For the second stimulus, N samples of the data are generated by $P_B(.)$ and a single (unknown) sample is generated by $P_A(.)$.

This problem can be modelled using additional variables $\{V_i\}$. These are binary *indicator* variables which determine whether the data comes from $P_A(.)$ or $P_B(.)$. In other words, $V_i = 1$ if the data element x_i is generated by $P_B(.)$ and $V_i = 0$ is the element is an outlier (i.e. generated by $P_A(.)$). We do not, of course, know these $\{V_i\}$, we have to

estimate them.

$$P(x_1, ..., x_{N+1}|V_1, ..., V_{N+1}) = \prod_{i=1}^{N+1} P_B(x_i)^{V_i} P_A(x_i)^{1-V_i}. \tag{3.44}$$

The distribution of the $\{V_i\}$ will be different for the non-outlier model P_{NO} and the outlier model P_O. We have:

$$\begin{aligned} P_{NO}(\{V_i\}) &= \prod_{i=1}^{N+1} \delta_{V_i,1}, \\ P_O(\{V_i\}) &= \frac{1}{N+1} \delta_{\sum_{i=1}^{N+1} V_i, N}, \end{aligned} \tag{3.45}$$

where the $P_O(.)$ allows there to be $N+1$ outlier positions.

To evaluate the two models, we must sum out over the internal (secondary) variables $\{V_i\}$. We obtain:

$$\begin{aligned} P_{NO}(\{x_i\}) &= \prod_{i=1}^{N+1} P_B(x_i), \\ P_O(\{x_i\}) &= \frac{1}{N+1} \sum_{j=1}^{N+1} P_A(x_j) \prod_{i \neq j} P_B(x_i). \end{aligned} \tag{3.46}$$

Note that we could have derived these distributions directly without bothering with the intermediate $\{V_i\}$ variables. Why did we bother? Well, making the intermediate variables *explicit* helps by making it clear that different tasks are closely related and differ only by which variables are marginalized over. More importantly, however, for the more sophisticated models later in this book (and indeed later in this chapter) making the hidden variables explicit greatly simplifies the notation and allows us to use algorithms such as EM, see later section of this chapter.

Hence, the decision criterion for determining whether there is an outlier or not is given by the log-likelihood ratio test:

$$\log\{\frac{P_O(\{x_i\})}{P_{NO}(\{x_i\})}\} = \log\{\frac{1}{N+1} \sum_{j=1}^{N+1} \frac{P_A(x_j)}{P_B(x_j)}\}. \tag{3.47}$$

Thus, if we use MAP, we should decide that there is an outlier only if $\sum_{j=1}^{N+1} \frac{P_A(x_j)}{P_B(x_j)} > (N+1)$. As usual, we see that the effectiveness of the test depends on the log-likelihood ratio $\log P_A(x)/P_B(x)$.

Note that this is asymmetric in the A and B. The difficulty of detecting an outlier A within a background of B is *not* the same as doing the converse. To get intuition for why

this asymmetry can arise consider the simple example where the observables x can take two values only. We label these values α, β and suppose that $P_A(x = \alpha) = 1, P_A(x = \beta) = 0$ but that $P_B(x = \alpha) = 0.5, P_B(x = \beta) = 0.5$. With these distributions, detecting a sample of A in a background of B's is an almost impossible task. The sample from A will be an α but the N background samples from $P_B(.)$ will have roughly an equal number of α's and β's (typical samples will look like $\alpha, \alpha, \beta, \alpha, \beta, \beta, \alpha$ whether or not there is an outlier present). This means that the α from A will easily get lost in the background of α's from B. However, consider the opposite task of detecting a sample of B in a background of A's. Half the time, the sample from B will be a β which will be straightforward to pick out of the background of α's generated by $P_A(.)$. (All samples without an outlier will be of form $\alpha, \alpha, \alpha, \alpha, \alpha$, but samples with an outlier may be of form $\alpha, \alpha, \beta, \alpha, \alpha$). In a more technical subsection we will give further results about how asymmetry can arise.

For the second task, we assume that the data is generated by the model $P_O(\{x_i\})$. To determine the outlier, we are then asked to select the most probable configuration of the $\{V_i\}$ conditioned on the data (and, of course, with the restriction that there is only one outlier). This gives the estimate for the outlier as:

$$V_i^* = 0, \;\; i^* = \arg max_i \log \frac{P_A(x_i)}{P_B(x_i)}. \tag{3.48}$$

This again depends on the log-likelihood ratio but in a completely different form. Once again there is asymmetry in the task which can be seen by considering the previous example where the observations are α, β.

To understand the error rates for this problem we first turn it into a standard two decision classification problem. This requires deriving a probability distribution for the maximum of $\log \frac{P_A(x_i)}{P_B(x_i)}$ for N samples x_i from $P_B(.)$. This can be done using a simple identity (ref Rivest et al). We have the cumulative probability distribution (i.e. we must differentiate it to get the probability density function):

$$Pr(\max\{\log \frac{P_A(x_i)}{P_B(x_i)} : \; i = 1, ..., N\} > T | \; x \; drawn \; from \; B) = 1 - \{Pr(\frac{P_A(x)}{P_B(x)} < T | \; x \; drawn \; from \; B)\}^N. \tag{3.49}$$

We can compare this to the probability distribution for the response of the A sample:

$$Pr(\{\log \frac{P_A(x)}{P_B(x)} = y | \; x \; drawn \; from \; A) = \int dx P_A(x) \delta(\log \frac{P_A(x)}{P_B(x)} - y). \tag{3.50}$$

From these two distributions it is possible to calculate the false positive and false negative error rates as in the previous chapter.

The third task is the hardest. We now have $N + 1$ hypotheses. First, all the data is generated by $P_B(.)$, second the first element of the data is generated by $P_A(.)$ and the

rest by $P_B(.)$, third that the second element is generated by $P_A(.)$ and the rest by $P_B(.)$, and so on. We label these hypotheses as H_0, for all the data coming from $P_B(.)$. to H_i the hypothesis that the data from $P_A(.)$ is the ith element. The prior probabilities are $P(H_0) = 1/2$ and $P(H_i) = 1/(2(N+1))$ for $i = 1, ..., N+1$.

We now specify the likelihood functions:

$$P(x_1, ..., x_{N+1}|H_0) = \prod_{i=1}^{N+1} P_B(x_i),$$
$$P(x_1, ..., x_{N+1}|H_j) = P_A(x_j) \prod_{i \neq j=1}^{N+1} P_B(x_i). \tag{3.51}$$

In this case, we simply need to pick the largest of the following set of numbers (corresponding to the votes for model H_0 and H_J, $j = 1, ..., N+1$ respectively.

$$\{\log P(x_1, ..., x_{N+1}|H_0) + \log(1/2), \log P(x_1, ..., x_{N+1}|H_j) + \log(1/(2(N+1))), \;\; j = 1, ..., N+1.\} \tag{3.52}$$

From the form of the distributions this reduces to picking the maximum of

$$\{1, \log \frac{P_A(x_j)}{(N+1)P_B(x_j)}, \quad j = 1, ..., N+1.\} \tag{3.53}$$

This is clearly the hardest task. It is quite possible that an outlier is present but that $\log \frac{P_A(x_j)}{(N+1)P_B(x_j)} < 1$ for all $j = 1, ..., N+1$. If the choice was to determine whether an outlier is present (without knowing where its position is) we would simply have to determine that $\sum_{j=1}^{N+1} \log \frac{P_A(x_j)}{(N+1)P_B(x_j)} < 1$ which makes it less likely to make mistakes. Once again, our example with α, β make it clear that this task is also asymmetric.

In this task, misclassifications can occur in several ways. Suppose there is no outlier present. Then the errors can arise with probability $Pr(\max\{\log \frac{P_A(x_j)}{P_B(x_j)}\} > \log(N+1)|\ \{x_j\}\ from\ B)$ which, using the argument above, we can express as $1 - \{Pr(\log \frac{P_A(x)}{P_B(x)} > \log(N+1)|\ x\ drawn\ from\ B)\}^{N+1}$. This form of error rate can be small. Alternatively, errors can arise if the $(N+1)^{th}$ sample is drawn from A. This case can be misclassified in two ways. Either the stimulus is classified as no outlier being present, with probability $Pr(\log \frac{P_A(x)}{P_B(x)} < \log(N+1)|\ x\ drawn\ from\ A)\{Pr(\log \frac{P_A(x)}{P_B(x)} < \log(N+1)|\ x\ drawn\ from\ B)\}^N$. Alternatively, it can be misclassified as having the outlier occurring in an incorrect position. This requires $Pr(\max\{\log \frac{P_A(x_j)}{P_B(x_j)}\} > \max(\log(N+1), \log \frac{P_A(x)}{P_B(x)})|\ \{x_j\}\ from\ B,\ x\ drawn\ from\ A)$.

3.4.5 More complex models of visual search

We now generalize the class of models we can apply these theories to. This should make you appreciate the usefulness of making the binary indicator variables V *explicit*. In later chapters, we will show that even more sophisticated visual tasks can be formulated in this way.

The same procedures can be generalized to situations where the number of outliers is either a fixed number which differs from 1 or even a random variable specified by a probability distribution. We can, for example, consider the outlier task with

$$P(x_1, ..., x_{N+1}|V_1, ..., V_{N+1}) = \prod_{i=1}^{N+1} P_B(x_i)^{V_i} P_A(x_i)^{1-V_i}, \tag{3.54}$$

and with any prior distribution $P(\{V_i\})$ on the indicator variables.

One possibility is to assume that there are H outlier points which are equally likely. This corresponds to picking a prior distribution:

$$P(\{V_i\}) = \frac{1}{Z}\delta_{\sum_{i=1}^{N+1} V_i, H}, \tag{3.55}$$

where Z is a normalization factor (exercise, what is it?).

Another possibility is to define a probability distributions for the number of outliers. Some possibilities are

$$P(\{V_i\}) = \frac{1}{Z}e^{-\sum_{i=1}^{N+1} V_i}, \text{ or } P(\{V_i\}) = \frac{1}{Z}e^{-\lambda\{\sum_{i=1}^{N+1} V_i\}^2} \tag{3.56}$$

Yet another is to assume that it is most probable for neighbouring points to be outliers (more sophisticated models of this type will be dealt with in later sections). This can be expressed by a distribution:

$$P(\{V_i\}) = \frac{1}{Z}e^{\sum_{i=1}^{N} V_i V_{i+1}}. \tag{3.57}$$

In all cases we can similar analyses. We can determine whether an outlier is present by comparing $P_B(\{x_i\}$ with $\sum_{\{V_i\}} P(\{x_i\}|\{V_i\})P(\{V_i\}$. We can attempt to ask more precise questions by enlarging the hypothesis set to include all class of allowable configurations of $\{V_i\}$ and to evaluate their evidence $\log P_B(\{x_i\}$ and $\log P(\{x_i\}|\{V_i\})P(\{V_i\}$.

These types of models will clearly predict that performance improves if more extra features are available. The analysis above would simply be modified to generalize the scale observables x_i to be vector valued $\vec{x}_i$. It has been reported in the literature that search tasks become significantly simpler when conjunctions of features are present (Treisman). Analysis of these experiments (Palmer, Verghese, Pavel) suggest that these improvements are consistent with models of the type we have been describing.

In later chapters, we will discuss how to formulate far more complicated models within the same framework by using more powerful families of probability distributions.

3.4.6 Asymmetry in Visual Search tasks

In this section, we give some insight into the search asymmetry by claiming that some search examples can be analyzed as the number of samples becomes large. This is because the samples are i.i.d. which implies that their statistical fluctuations tend to average themselves out. This section makes use of certain results, such as the law of large numbers and large deviation theory, which will only be derived in a later chapter. At present these claims should be taken on faith.

Consider the task of determining if an outlier from $P_A(.)$ is, or is not, present in a background of stimuli from $P_B(.)$. The criteria used is the likelihood test $(1/(N+1))\sum_{i=1}^{N+1} P_A(x_i)/P_B(x_i)$. Now suppose that all the data is generated by $P_B(.)$. The claim is, for certain situations to be clarified in a later chapter!!, that the *distribution of the likelihood ratio is sharply peaked at its mean value*. The mean value can be calculated by taking the expectation with respect to $\prod_{i=1}^{N+1} P_B(x_i)$. This gives that the most probable value of the test is 1 if all the data is generated from $P_B(.)$. Conversely, suppose all but one element of the data is generated by $P_B(.)$. Then the likelihood test splits into the part corresponding to samples from B and a single sample x from $P_A(.)$. The most probable contribution from the B samples is $N/(N+1)$ (by applying the argument above). Therefore the likelihood ratio takes the value 1 (with high probability) if all data comes from B and the random value $N/(N+1) + 1/(N+1)\log P_A(x)/(P_B(x))$. So the effectiveness of the test depends on the probability that $P_A(x) > P_B(x)$ given that the data is generated by $P_A(.)$. This is just the probability that a sample from $P_A(.)$ will be misclassified as being from $P_B(.)$ using ML estimation. So in this limit the search task becomes equivalent to simply classifying a stimulus as being either A or B by ML. This is therefore usually asymmetric because the overall between two distributions is usually asymmetric, see figure (3.11).

Arguments of this type should be used with great caution. In making the argument we have assumed that the distribution of a large set of samples generated by $P_B(.)$ is infinitely tightly peaked. Strictly speaking, this will only be true in the (unrealistic) limit as the number of samples goes to infinity. To make the argument rigourous we refer to a later chapter!! where we put bounds on the probabilities of the expectations over large samples from B of differing from the mean value.

3.4.7 Manhattan Example

We now describe another example of hidden variables. This involves binary indicator variables that label different types of edges in an image. The input to the system is the set of edges extracted from an image. See figure (3.13).

Most indoor and outdoor city scenes are based on a cartesian coordinate system which

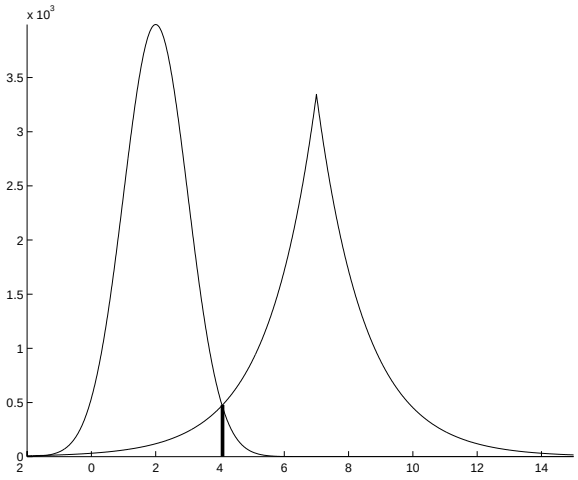

Figure 3.11 Asymmetry in error rates. The chance that a sample from $P_A(.)$ is misclassified as B is not, in general, the same as the chnace that a sample from $_B(.)$ is misclassified as A. There wil, however, be no asymmetry for the important case of equal variance Gaussians.

---

we can refer to as a Manhattan grid. This grid defines an $\vec{i}, \vec{j}, \vec{k}$ coordinate system. This gives a natural reference frame for the viewer. If the viewer can determine his/her position relative to this frame – in other words, estimate the $\vec{i}, \vec{j}$ or $\vec{k}$ directions – then it becomes significantly easier to interpret the scene. In particular, it will be a lot easier to determine the most important lines in the scene (corridor boundaries and doors, street boundaries and traffic lights) because they will typically lie in either the $\vec{i}, \vec{j}$ or $\vec{k}$ directions. Knowledge of this reference frame will also make it significantly easier and faster to outliers which are *not* aligned in this way. We define Ψ to be the compass angle. This defines the orientation of the camera with respect to the Manhattan grid: the camera points in direction $\cos\Psi\vec{i} - \sin\Psi\vec{j}$. Camera coordinates $\vec{u} = (u, v)$ are related to the Cartesian scene coordinates (x, y, z) by the equations:

$$u = \frac{f\{-x\sin\Psi - y\cos\Psi\}}{x\cos\Psi - y\sin\Psi}, \quad v = \frac{fz}{x\cos\Psi - y\sin\Psi}, \tag{3.58}$$

where f is the focal length of the camera (or eye).

At each image pixel we either have an edge (with its orientation) or no edge. The edge could result either from an $\vec{i}, \vec{j}, \vec{k}$ line or from an *un-aligned* edge. More formally, the image data at pixel $\vec{u}$ is explained by one of five models $m_{\vec{u}}$: $m_{\vec{u}} = 1, 2, 3$ means the data is generated by an edge due to an $\vec{i}, \vec{j}, \vec{k}$ line, respectively, in the scene; $m_{\vec{u}} = 4$ means the data is generated by a random edge (not due to an $\vec{i}, \vec{j}, \vec{k}$ line); and $m_{\vec{u}} = 5$ means the pixel is off-edge. The prior probability $P(m_{\vec{u}})$ of each of the edge models was estimated empirically to be $0.02, 0.02, 0.02, 0.04, 0.9$ for $m_{\vec{u}} = 1, 2, \ldots, 5$.

It is a straightforward geometry to show that an edge in the image at $\vec{u} = (u, v)$ with edge normal at $(\cos\theta, \sin\theta)$ is *consistent with an $\vec{i}$ line in the sense that it points to the*

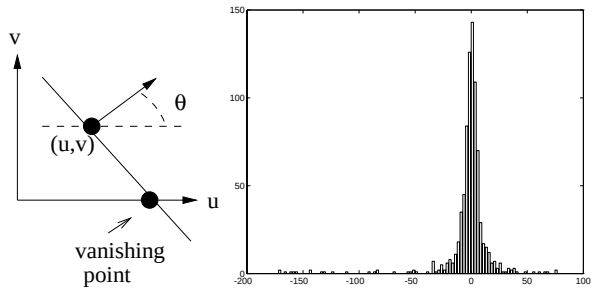

Figure 3.12 (Left). Geometry of an $\vec{i}$ line projected onto (u, v) image plane. θ is the normal orientation of the line in the image. Because our camera is assumed to point in a horizontal direction, the vanishing point lies on the u axis. (Right) Histogram of edge orientation error (displayed modulo 180°). Observe the strong peak at 0°, indicating that the image gradient direction at an edge is usually very close to the true normal orientation of the edge. This distribution is modelled using a simple box function.

vanishing point if $-v \tan\theta = u + f \tan\Psi$ (observe that this equation is unaffected by adding $\pm\pi$ to θ and so it does not depend on the polarity of the edge). We get a similar expression $v \tan\theta = -u + f \cot\Psi$ for lines in the $\vec{j}$ direction. See figure (3.12).

We assume that there is an uncertainty in estimating the edge orientation $\phi_{\vec{u}}$ at a point $\vec{u}$ described by a probability distribution $P_{ang}(\phi)$, see figure (3.12). More precisely, $P(\phi_{\vec{u}}|m_{\vec{u}}, \Psi, \vec{u})$ is given by $P_{ang}(\phi_{\vec{u}} - \theta(\Psi, m_{\vec{u}}, \vec{u}))$ if $m_{\vec{u}} = 1, 2, 3$ or $U(\phi_{\vec{u}}) = 1/(2\pi)$ if $m_{\vec{u}} = 4, 5$. Here $\theta(\Psi, m_{\vec{u}}, \vec{u}))$ is the predicted normal orientation of lines determined by the equation $-v \tan\theta = u + f \tan\Psi$ for $\vec{i}$ lines, $v \tan\theta = -u + f \cot\Psi$ for $\vec{j}$ lines, and $\theta = 0$ for $\vec{k}$ lines.

In summary, for models 1,2 and 3 the edge orientation is modeled by a distribution which is peaked about the appropriate orientation of an $\vec{i}, \vec{j}, \vec{k}$ line predicted by the compass angle at pixel location $\vec{u}$; for model 4 the edge orientation is assumed to be uniformly distributed from 0 through 2π. Places where there are no edges are automatically assigned to model 5.

Rather than decide on a particular model at each pixel, we marginalize over all five possible models (i.e. creating a mixture model):

$$P(\phi_{\vec{u}}|\Psi, \vec{u}) = \sum_{m_{\vec{u}}=1}^{5} P(\phi_{\vec{u}}|m_{\vec{u}}, \Psi, \vec{u}) P(m_{\vec{u}}) \tag{3.59}$$

In this way we can determine evidence about the camera angle Ψ at each pixel without knowing which of the five model categories the pixel belongs to.

To combine evidence over all pixels in the image, denoted by $\{\phi_{\vec{u}}\}$, we assume that the image data is conditionally independent across all pixels, given the compass direction Ψ:

$$P(\{\phi_{\vec{u}}\}|\Psi) = \prod_{\vec{u}} P(\phi_{\vec{u}}|\Psi, \vec{u}) \tag{3.60}$$

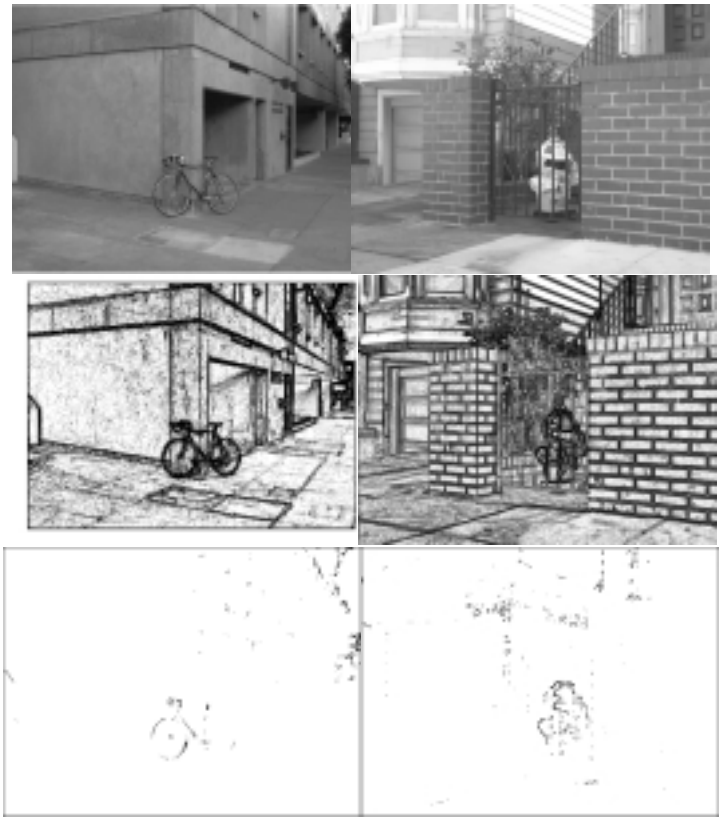

Figure 3.13 Detecting bikes (left column) and robots (right column) in urban scenes. The original image (top row) and the edge maps (centre row) – displayed as a grey-scale image where black is high and white is low. In the bottom row we show the edges assigned to model 4 (i.e. the outliers) in black. Observe that the edges of the bike and the robot are now highly salient (and make detection straightforward) because most of them are unaligned to the Manhattan grid.

Thus the posterior distribution on the compass direction is given by $\prod_{\vec{u}} P(\phi_{\vec{u}}|\Psi, \vec{u})P(\Psi)/Z$ where Z is a normalization factor and $P(\Psi)$ is a uniform prior on the compass angle.

To find the MAP (maximum a posterior) estimate, we need to maximize the log posterior term (ignoring Z, which is independent of Ψ) $\log[P(\{\phi_{\vec{u}}\}|\Psi)P(\Psi)] = \log P(\Psi) + \sum_{\vec{u}} \log[\sum_{m_{\vec{u}}} P(\phi_{\vec{u}}|m_{\vec{u}}, \Psi, \vec{u})P(m_{\vec{u}})]$. This can be computed by an algorithm which evaluates the log posterior numerically for the compass direction Ψ in the range $-45°$ to $+45°$, in increments of $1°$.

You can also integrate out the $\{m_i\}$ to get the ψ. (This is an exercise for the reader.)

3.5 The EM algorithm

The *Expectation Maximization* (EM) algorithm is a very common procedure for integrating out "hidden variables". In principle, it is very general and, as we describe below, it is guaranteed to converge to a local optimum. There is no guarantee, however, that it

will converge to the optimal solution. Indeed, as we will show, it can be reformulated as a variant of the standard steepest descent algorithm. Steepest descent algorithms are guaranteed to converge to a local minimum but will often fail to reach the global minimum unless they start out with good initial conditions.

3.5.1 Mixture of Gaussians Case

We start with a simple example. We have data $\{x_i : i = 1, ..., N\}$ which is generated by a mixture of Gaussians $P(x|\mu_a, \sigma^2)$ for $a = 1, ..., M$ (where $M < N$). Our goal is to estimate the means of the Gaussians (we assume that their variances are known). The problem is that we do not know which data is generated by which Gaussian.

We introduce an auxiliary variable $\{V_{ia}\}$ so that $V_{ia} = 1$ if data x_i is generated by model $P(x|\mu_a, \sigma^2)$ and $V_{ia} = 0$ otherwise. We impose the constraint $\sum_a V_{ia} = 1, \ \forall \ i$ to ensure that every data point is generated by exactly one model. We now write:

$$P(\{x_i\}|\{\mu_a\}) = \sum_{\{V_{ia}\}} P(\{x_i\}, \{V_{ia}\}|\{\mu_a\}). \tag{3.61}$$

Here we set

$$P(\{x_i\}, \{V_{ia}\}|\{\mu_a\}) = \frac{1}{Z} e^{-\sum_{ia} V_{ia}(x_i - \mu_a)^2/(2\sigma^2)}, \tag{3.62}$$

which assumes a uniform prior on the $\{V_{ia}\}$'s (i.e. any assignment of the data to the model is, a priori, equally likely).

First, we check that this formulations means that $P(\{x_i\}|\{\mu_a\})$ is a mixture of Gaussians. To do this we first write $P(\{x_i\}, \{V_{ia}\}|\{\mu_a\}) = \frac{1}{Z} \prod_{i=1}^N e^{-\sum_a V_{ia}(x_i - \mu_a)^2/(2\sigma^2)}$. For each i, we now sum over the variables $\{V_{ia} : a = 1, ..., M\}$ with the constraint $\sum_a V_{ia} = 1$. This yields $P(\{x_i\}|\{\mu_a\}) = \frac{1}{Z} \prod_{i=1}^N \{\sum_{a=1}^M e^{-\sum_a (x_i - \mu_a)^2/(2\sigma^2)}\}$ which shows that each x_i is generated independently from a mixture of Gaussians. (Exercise, check this by considering the special case when $M = 2$).

Now we want to estimate the most probable $\{\mu_a\}$ from $P(\{x_i\}|\{\mu_a\})$. The EM algorithm says that we can do this by iterating two steps. The first step estimates the *probability that data x_i is generated by distribution a.* More precisely, we compute $\hat{P}(\{V_{ia}\}) = P(\{V_{ia}\}|\{x_i\}, \{\mu_a\})$ using our current estimate of the $\{\mu_a\}$. Since this probability distribution factorizes over i, see equation (3.62), we can express this as the product of probability distributions $P(V_{ai}|x_i, \{\mu_a\})$ for each i. Because V_{ia} is binary valued (i.e. $V_{ia} = 0, 1$) we can represent this by the expected value of V_{ia} which we call $\bar{V}_{ia} = \sum_{V_{ia}} V_{ia} P(V_{ai}|\{x_i\}, \{\mu_a\})$.

This estimation is done by marginalization. It gives:

$$P(V_{ia} = 1|x_i, \{\mu_a\}) = \frac{e^{-(x_i - \mu_a)^2/(2\sigma^2)}}{\sum_{b=1}^M e^{-(x_i - \mu_b)^2/(2\sigma^2)}}. \tag{3.63}$$

The second stage is to make the most probable estimates of the $\{\mu_a\}$ assuming that the $P(V_{ia} = 1|x_i, \{\mu_a\})$ are fixed. More precisely, we choose the $\{\mu_a\}$ to maximize $\sum_{\{V_{ia}\}} \hat{P}(\{V_{ia}\}) \log P(\{V_{ia}\}, \{x_i\}|\{\mu_a\})$. This gives:

$$\mu_a = \frac{\sum_{i=1}^{M} \bar{V}_{ia} x_i}{\sum_{i=1}^{M} \bar{V}_{ia}}, \ \forall \ a. \tag{3.64}$$

The first stage estimates which model is most likely to have generated the data x_i (given the current values of the model parameters). The second stage, estimates the means μ_a of the models with the data weighted by the (estimated) probability that it is due to model a.

The two stages iterate and can be proven, see next subsection, to converge to a local maximum of $P(\{x_i\}|\{\mu_a\})$. (This, of course, is ML estimation but the same approach can be easily extended to deal with MAP estimation if there is a prior om the $\{\mu_a\}$). The procedure does require initialization and, frankly, whether the algorithm converges to the true maximum of $P(\{\mu_a\}|\{x_i\})$ often depends on how good the initialization is.

This example, hopefully, gives some intuition about the EM algorithm. But it is unrepresentative in several respects. Firstly, both the E and M stages can be solved for analytically. In more realistic cases one or both steps will require calculation by computer. (Imagine if we replaced $\sum_{ia} V_{ia}(x_i - \mu_a)^2$ by $\sum_{ia} V_{ia}(x_i - \mu_a)^4$ in the exponent of the probability!). Secondly, the intermediate (or hidden) variables $\{V_{ai}\}$ are binary variables of a particularly simple form. There is no need for this either, but it does drastically simplify things.

3.5.2 The general form of the EM algorithm

This subsection describes the general form of the EM algorithm. Our formulation will be for continuous hidden variables but it can be adapted directly to discrete hidden variables (just replace the integrals by summations).

Suppose the observations x are generated by state variables s and hidden variables h. We assume that the probability distributions $P(x, h|s)$ are known.

For example, x could represent the image of an object s and the hidden variables h could be the illumination conditions (or the viewpoint, or some internal state of the object).

The goal is to find the MAP estimator s^* which maximizes

$$P(s|x) = \int \frac{P(x, h|s)P(s)}{P(x)} dh. \tag{3.65}$$

The term $P(x)$ is constant (independent of s) so we drop it in order to simplify the algebra. (I.e. we search for the maximum of $P(s|x)P(x)$ with respect to s.)

The EM algorithm is guaranteed to find at least a local maximum of $P(s|x)$. It starts

by making a guess s_0 of the state variables. Then it proceeds by iterating two steps. The *E-step* estimates the distribution of h from $P(h|x, s_t)$, where s_t is the current estimate of the state. The *M-step* maximizes $\int dh P(h|x, s_t) \log\{P(h, x|s)P(s)\}$ to estimate s_{t+1}. The two steps combined give an update rule:

$$s_{t+1} = \arg\max_s \{ \int dh P(h|x, s_t) \log\{P(h, x|s)P(s)\}. \tag{3.66}$$

The EM algorithm for a mixture of Gaussians, previous subsection, can be derived from this general case by replacing the h by V, the integrals by summations, and using the specific probability distribution for the mixture case, see equation (3.62). (Details left as an exercise for the reader).

3.5.3 Why Does the EM algorithm Converge?

Now we address the issue of why the EM algorithm converges. It may seem to be a miraculous algorithm because it enables you to integrate out hidden variables that you never observe. But it is not so strange when one understands it ("A miracle is simply technology that you don't understand" W. Gates. III). It should be stressed that the EM algorithm always assumes a *probability distribution* for the hidden variables conditioned on the state variables. So knowledge about the hidden variables is built into the system from the start.

We now give a proof for convergence for the EM algorithm. The proof proceeds by showing that the EM algorithm can be simply transformed into a steepest descent/ascent algorithm of an energy function. The E-step corresponds to minimizing with respect to one set of variables (with the other variables kept fixed) while the M-step involves minimizing with respect to the second set of variables (with the first set fixed). it is clear that each step of this procedure will reduce the energy and, provided the energy is bounded below, convergence to at least a local minimum is guaranteed. (The requirement that the energy be bounded below will automatically be true unless the probability distributions are truly bizarre).

The basic idea (Hathaway, Hinton and Neal) is to define a new variable $\hat{P}(h)$ which is a probability distribution. We then define a function $F(\hat{P}, s)$ specified by:

$$F(\hat{P}, s) = \int dh \hat{P}(h) \log P(h, x|s) - \int dh \hat{P}(h) \log \hat{P}(h), \tag{3.67}$$

which can be re-expressed as the log-likelihood we wish to maximize $\log P(x|s)$ minus the Kullback-Leibler distance $D(\hat{P}(h)||P(h|x, s))$ between the estimated distribution on the hidden variables $\hat{P}(h)$ and the true distribution of h conditioned on our data x and current estimate of s.

The key result is that maximizing this "energy" function with respect to $\hat{P}$ and s is equivalent to the EM algorithm

Theorem: EM Equivalence *Alternatively maximizing $F(\hat{P}, s)$ with respect to $\hat{P}$ and s respectively (keeping the other variable fixed) is equivalent to the EM algorithm. Moreover the maximum of $F(\hat{P}, s)$ with respect to $\hat{P}$ is the evidence $\log P(x|s)$ for the state s.*

Proof. *The key point to note is that marginalizing $\log P(x|s)$ is equivalent to maximizing $\log P(x|s) - D(\hat{P}(h)||P(h|x,s))$ jointly over s and $\hat{P}(h)$, where $\hat{P}(h)$ is any distribution on h (and $D(.||.)$ is the Kullback-Leibler divergence). The non-negativity of the Kullback-Leibler divergence, combined with the fact that the divergence is zero only between identical distributions, ensures that the maximum is reached only by setting $\hat{P}(h)$ equal to the true distribution on h, i.e. $P(h|x,s)$. By expanding out the Kullback-Leibler divergence, We can rewrite $\log P(x|s) - D(\hat{P}(h)||P(h|x,s))$ as $H(\hat{P}) + \int dh \hat{h}at P(h) \log\{P(h|x,s)P(x|s)\}$, where $H(\hat{P}) = -\int dh \hat{P}(y) \log \hat{P}(h)$ is the entropy of $\hat{P}$. This can then be rewritten as $H(\hat{P}) - \int dh \hat{P}(h) \log P(h,x|s)$.*

We illustrate this proof by obtaining an energy function which corresponds to the mixture of Gaussian example discussed earlier. The hidden variables are the binary indicator variables $\{V_{ia}\}$. The probability distribution on them can be represented by their expected values $\{\bar{V}_{ia}\}$ with the constraint that $\sum_a \bar{V}_{ia} = 1, \ \forall\, i$. The entropy for this distribution can be calculated to be $-\sum_{ia} \bar{V}_{ia} \log \bar{V}_{ia}$. We therefore obtain an energy function:

$$E[\{\bar{V}_{ia}\}, \{\mu_a\}] = \sum_{ia} \bar{V}_{ia} \frac{(x_i - \mu_a)^2}{2\sigma^2} + \sum_{ia} \bar{V}_{ia} \log \bar{V}_{ia} + \sum_i \lambda_i \{\sum_a \bar{V}_{ia} - 1\}, \qquad (3.68)$$

where the $\{\lambda_i\}$ are Lagrange multipliers to impose the constraints $\sum_a \bar{V}_{ia} = 1, \ \forall\, i$.

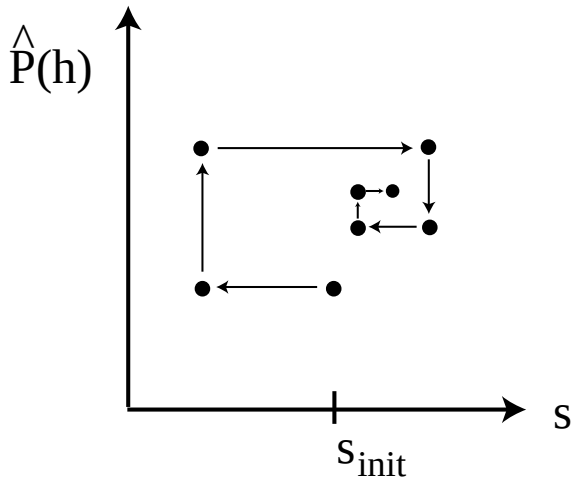

Figure 3.14 The two steps of EM treated as steepest descent, or ascent, in the two sets of variables. This formulation makes it clear that the EM algorithm will converge.

It can be verified that minimizing $E[\{\bar{V}_{ia}\}, \{\mu_a\}]$ with respect to $\{\mu_a\}$ gives the M-

step, see equation (3.64), while minimizing with respect to $\{\bar{V}_{ia}\}$ gives the E-step (when expressed in terms of the $\{\bar{V}_{ia}\}$). See figure (3.14).

3.5.4 A Traditional Proof of Convergence of the EM algorithm.

This subsection gives a more traditional proof of convergence of the EM algorithm. The proof is, perhaps, less intuitive than the one based on the steepest descent connection.

Theorem. Convergence of EM. *Each iteration of EM satisfies $P(s_{t+1}|s) \geq P(s_t|x)$ and so, provided $P(s|x)$ is bounded above, the algorithm converges to a local maximum of $P(s|x)$.*

Proof. *By using the equality $P(h, x|s) = P(h|x, s)P(x|s)$ we can write:*

$$\int dh P(h|x, s_t) \log\{P(h, x|s)P(s)\} = \int dh P(h|x, s_t) \log \frac{P(h|x, s)}{P(h|x, s_t)} + \int dh P(h|x, s_t) \log P(h|x, s_t) + \log P(x|s) + \log P(s). \tag{3.69}$$

The second term on the right hand side is simply minus the Kullback-Leibler divergence $D(P(h|x, s_t)||P(h|x, s))$ from $P(h|x, s_t)$ to $P(h|x, s)$ and is non-positive and equals zero only if $P(h|x, s_t) = P(h|x, s)$. Therefore if we set $s = s_t$ in the equation above, we obtain

$$\int dh P(h|x, s_t) \log P(h, x|s_t) = \int dh P(h|x, s_t) \log P(h|x, s_t) + \log P(x|s_t) + \log P(s). \tag{3.70}$$

This gives:

$$\begin{aligned} \log\{P(x|s)P(s)\} - \log\{P(x|s_t)P(s_t)\} &= D(P(h|x, s_t)||P(h|x, s)) \\ &+ \int dh P(h|x, s_t) \log P(h, x|s) - \int dh P(h|x, s_t) \log P(h, x|s_t) \\ &\geq \int dh P(h|x, s_t) \log P(h, x|s) - \int dh P(h|x, s_t) \log P(h, x|s_t). \end{aligned} \tag{3.71}$$

The EM algorithm says we should select s_{t+1} to maximize $\int dh P(h|x, s_t) \log\{P(h, x|s)P(s)\}$. Hence we can be sure that $\int dh P(h|x, s_t) \log\{P(h, x|s_{t+1})P(s_{t+1})\} \geq \int dh P(h|x, s_t) \log\{P(h, x|s_t)P(s_t)\}$. Therefore if we set $s = s_{t+1}$ in the equation above, we can guarantee that the right hand side will be non-negative. Thus $\log\{P(x|s_{t+1})P(s_{t+1})\} \geq \log\{P(x|s_t)P(s_t)\}$ and the theorem is proven.

3.5.5 Another EM example: maybe too hard?

THIS EXAMPLE MAY BE TOO HARD – it hides EM under too much algebra!!

An example of the EM algorithm is for the problem of finding stop signs in an image (Yuille, Snow, Nitzberg). We have a template $\{\vec{z}_a : a = 1, ..., 8\}$ for the corner positions

of a standard stop sign viewed from front-on at a fixed distance. By the use of feature detectors (who exact form is irrelevant here) we have detected points $\{\vec{x}_i : i = 1, ..., M\}$ which are possible positions for the corners of the stop sign. Our goal is to match our template $\{\vec{z}\}$ to these data points $\{\vec{x}\}$. To allow for viewpoint variations, we assume that the images of the template is given by $\mathbf{A}\vec{z}_a + \vec{b} : a = 1, ..., 8$, where $\mathbf{A}$ is an matrix to allow for change in shape of the sign caused by the (unknown) viewpoint and $\vec{b}$ is the position of the sign. So the variables $\mathbf{A}, \vec{b}$ together are the state s that we wish to estimate. The hidden variables are binary indicator variables $\{V_{ia}\}$ which determine whether data point $\vec{x}_i$ is matched to template point $\vec{z}_a$. These matching variables must also take into account the fact that some corners $\{\vec{z}\}$ of the stop sign may be occluded by other objects and hence be invisible. Conversely, there may be data points $\{\vec{x}\}$ that do not correspond to corners of the template.

We specify a probability distribution:

$$P(\{\vec{x}_i\}, \mathbf{V}|\mathbf{A}, \vec{b}) = \frac{e^{-E[\mathbf{V}, \mathbf{A}, \vec{b}:\{\vec{x}_i\}]}}{Z}, \tag{3.72}$$

where Z is the normalization factor and

$$E[\mathbf{V}, \mathbf{A}, \vec{b} : \{\vec{x}_i\}] = \sum_{ia} V_{ia} |\mathbf{A}\vec{x}_i + \vec{b} - \vec{z}_a|^2 + \lambda \sum_i (1 - \sum_a V_{ia}). \tag{3.73}$$

The indicator matrix $\mathbf{V}$ is constrained so that, for all i, $\sum_a V_{ia} = 0$ or 1 to ensure that each data corner is matched to at most one true corner. If a data corner point is unmatched then it pays a penalty λ.

In this case, the probability distributions for the $\mathbf{V}$ can be represented by the mean values $\bar{V}_{ai}$ because the variables are binary (for distributions on binary variables the means specify the distribution precisely).

The algorithm proceeds as follows. Firstly, we initialize the variables $\{\bar{V}_{ai}\}$. Then we apply the E and the M step repetitively. The E-step involves estimating:

$$\bar{V}_{ai} = \frac{e^{-|\mathbf{A}\vec{x}_i + \vec{b} - \vec{z}_a|^2}}{e^{-\lambda} + \sum_c e^{-|\mathbf{A}\vec{x}_i + \vec{b} - \vec{z}_c|^2}}. \tag{3.74}$$

The M-step involves solving the simultaneous linear equations for $\mathbf{A}$ and $\vec{b}$:

$$\begin{aligned} \mathbf{A}(\sum_{ia} \bar{V}_{ia} \vec{x}_i) + \vec{b}(\sum_{ia} \bar{V}_{ia} - \sum_{ia} (\bar{V}_{ia} \vec{z}_a) = 0, \\ \mathbf{A}(\sum_{ia} \bar{V}_{ia} \vec{x}_i^T \vec{x}_i) + \sum_{ia} \bar{V}_{ia} \vec{x}_i^T \vec{b} - \sum_{ia} \vec{x}_i^T \vec{z}_a = 0. \end{aligned} \tag{3.75}$$

This example is again a bit simple because both E and M steps can be done analytically.

3.6 MFT Approximation and Bounding the Evidence

What happens when we have discrete variables and cannot compute the marginal distributions analytically? Is there anything like a Laplace approximation in this case? (Again, we warn that we are describing general purpose techniques in this section and for certain types of problem there are more effective methods which may not even require approximations, see later chapters).

The answer is yes, there are a set of approximations first obtained in the statistical physics literature and then applied to probability estimation tasks. In this section, we will describe one method known as the *naive mean field theory* approximation. One advantage of this derivation is that it makes it explicitly clear that the approximation gives a lower bound of quantities of interest such as the evidence. (Though natural extensions to higher order terms give approximations which may be better but which cannot be proven to be upper or lower bounds).

Suppose, for example, that we are trying to evaluate the evidence for a visual search task, see previous section, where the distribution of the data conditioned on the hidden states is given by:

$$P(x_1, ..., x_{N+1}|V_1, ..., V_{N+1}) = \prod_{i=1}^{N+1} P_B(x_i)^{V_i} P_A(x_i)^{1-V_i}, \tag{3.76}$$

and the distribution of the hidden states is:

$$P(\{V_i\}) = \frac{1}{Z} e^{\sum_{i=1}^{N} V_i V_{i+1}}. \tag{3.77}$$

The probability of the data is then given by

$$P(x_1, ..., x_{N+1}) = \sum_{\{V_i\}} \{ \prod_{i=1}^{N+1} P_B(x_i)^{V_i} P_A(x_i)^{1-V_i} \} \frac{1}{Z} e^{\sum_{i=1}^{N} V_i V_{i+1}}. \tag{3.78}$$

This summation is very difficult to perform if N is large. Mean field theory, however, gives a way to approximate it.

There are many ways to derive mean field theory approximations (refs). We choose the method that is most consistent with the spirit of this book (Jordan et al.). Suppose we want to estimate the evidence $\log P(x|s)$ for a state s when observing the data x. And suppose that we have hidden variables h and so $P(x|s) = \sum_h P(x, h|s)$. One way to approximate this is by replacing the probability distribution $P(h|x, s)$ by the "closest" element of a family of distributions $\{P_\lambda(h|x, s) : \lambda \in \Lambda\}$. (The choice of this approximation family is important and we will return to it shortly.) We measure closeness by the Kullback-Leibler divergence between $P_\lambda(h|x, s)$ and $P(h|x, s)$ to be:

$$F[\lambda] = \sum_h P_\lambda(h|x,s) \log \frac{P_\lambda(h|x,s)}{P(h|x,s)} \tag{3.79}$$

Using Bayes theorem, $(P(h|x,s) = P(x|h,s)P(h|s)/P(x|s))$, we can rewrite this as:

$$F[\lambda] = \sum_h P_\lambda(h|x,s)\{\log P_\lambda(h|x,s) + \log P(x|s) - \log P(x|h,s) - \log P(h|s)\}. \tag{3.80}$$

We can therefore write:

$$\log P(x|s) = F[\lambda] + \sum_h P_\lambda(h|x,s)\{\log P(x|h,s) + \log P(h|s) - \log P_\lambda(h|x,s)\}. \tag{3.81}$$

$F[\lambda]$ is always positive semi-definite (because it is a Kullback-Leibler divergence) and so we can turn this into an inequality $\log P(x|s) \geq \sum_h P_\lambda(h|x,s)\{\log P(x|h,s) + \log P(h) - \log P_\lambda(h|x,s)\}$.

This inequality is strengthened by selecting $\lambda^* = \arg\min_\lambda F[\lambda]$. Hence, we obtain

$$\log P(x|s) \geq \sum_h P_{\lambda^*}(h|x,s)\{\log P(x|h,s) + \log P(h|s) - \log P_{\lambda^*}(h|x,s)\}. \tag{3.82}$$

This result is only useful if it is possible to find a set of families $P_\lambda(h;x,s)$ for which it is both possible to estimate λ^* and to compute the right hand side of the inequality.

The most promising family is the set of *factorizable distributions* so that

$$P_\lambda(h;x,s) = \prod_{i=1}^{N} p_i(h_i;x,s), \tag{3.83}$$

where we denote the variable $h = (h_1, ..., h_N)$. The parameters λ correspond to the ways of specifying the distributions $p_i(h_i;x,s)$. For example, it may be assumed that each h_i is a binary variable which takes a state either 0 or 1. Then we can parameterize $p_i(h_i;x,s) = \lambda_i^{h_i}(1-\lambda_i)^{1-\lambda_i}$. The set of variables $\lambda = \lambda_1, ..., \lambda_N$ will specify the distribution $P_\lambda(h;x,s) = \prod_{i=1}^N p_i(h_i;x,s)$ uniquely. (Note that in this approximation the $\{\lambda_i\}$ will be functions of the data x and the state s which we are calculating the evidence for.)

We now return to our example from visual search. So we replace the h by $\{V_i\}$ and drop the s variable (because our example, for simplicity, only considers the evidence for a single state).

We can write $\log P(x,V)$ as

$$\log P(x|V) + \log P(V) = \sum_{i=1}^{N+1} \{V_i \log P_B(x_i) + (1 - V_i) \log P_A(x_i)\} + \frac{1}{N} \sum_{i=1}^{N} V_i V_{i+1} - \log Z. \tag{3.84}$$

We approximate by factorized distributions of form:

$$P_\lambda(\{V_i\} : x) = \prod_{i=1}^{N} \{\lambda_i^{V_i} (1 - \lambda_i)^{1-V_i}\}. \tag{3.85}$$

The mean values of the $\{V_i\}$ with respect to $P_\lambda(\{V_i\} : x)$ are given by $\hat{V}_i = \lambda_i$. Hence, we see that:

$$\sum_{\{V_i\}} P_{\lambda^*}(\{V_i\}|x)\{\log P(x|\{V_i\}) + \log P(\{V_i\}) - \log P_{\lambda^*}(\{V_i\}|x)\} = \sum_{i=1}^{N} \lambda_i \lambda_{i+1} + \sum_{i=1}^{N+1} \{\lambda_i \log P_B(x_i) + (1-\lambda_i) \log P_A($$

(3.86)

The $\{\lambda_i\}$ must be found to minimize the right hand side. The equations correspond to the well known *mean field equations* studied in statistical physics. Although there are are no algorithms known to be guaranteed to solve them (in general) there are nevertheless a set of good approximate algorithms that converge, at least, to a local minimum of these equations. Recall, that if we are only looking for a lower bound, then a local minimum will be sufficient.

In any case, suppose we have found $\{\lambda_i^*\}$ which gives a local minimum of $\sum_h P_{\lambda^*}(h|x)\{\log P(x|h) + \log P(h) - \log P_{\lambda^*}(h|x)\}$. Then the value of the bound can be computed directly by substituting in for $\{\lambda_i^*\}$. Hence, we find that:

$$\log P(x) \geq \sum_{i,j} W_{ij} \lambda_i^* \lambda_j^* + \sum_i \theta_i \lambda_i^* - \sum_i \{\lambda_i^* \log \lambda_i^* + (1 - \lambda_i^*) \log(1 - \lambda_i^*)\}, \tag{3.87}$$

where $\{\lambda_i\}$ are chosen to maximize the right-hand-side.

How does this method relate to Laplace's method described in the previous section. Although there are striking differences there are some underlying similarities. The mean field method applies only to binary, or discrete valued, variables. The model evidence is therefore defined by a sum over discrete states. It is possible, however, using techniques from analytic continuation to re-express this sum in terms of integrals over continuous variables which can be re-interpreted as the $\{\lambda_i\}$. One can then apply Laplace's method to this continuous version. The result is that the variables $\{\lambda_i^*\}$ that maximize the integrand for Laplace's method *are precisely those which minimize* $F[\lambda]$. So the mean field method correspond to doing the first part of Laplace's method but ignoring the quadratic and

higher order terms. (This analytic continuation is used a great deal in Statistical Physics but a rigourous justification for it is lacking to our knowledge).

Key points:

- Model Selection Motivation
- Continuous Variable Tasks.
- Genericity.
- Laplaces's method
- Discrete Variable Tasks
- Robustness and Outliers
- Visual Search
- Manhattan
- EM Algorithm
- MFT Approximation