Bayesian Segmentation of Dot Pictures

STEPHEN M. OMOHUNDRO

International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, California 94704
Phone: 415-643-9153
Internet: om@icsi.berkeley.edu
Date, 1989

Abstract. We apply Bayesian decision theory to the segmentation of simple images generated from precise statistical models. In this context many of the central features of recognition processes such as competition between models, combining rules for evidence and the preference for explanation simplicity become clearly explicated.

Introduction

Statistical decision theory. The problems of model-based vision may be analyzed in the framework of statistical decision theory. By model based we mean that the perceiver knows the properties of a space M of possible models and the manner by which each model gives rise to a particular sensation out of a space S of possible sensations. The relationship between models and sensations is often statistical, with a model $m \in M$ giving rise to a probability distribution $p(s \mid m)$ on S. If we assume that the perceiver knows the prior distribution of models $\pi(m)$ and has a loss function L which measures the severity of mistakes of perception then we are abstractly in the framework of statistical decision theory. This theory may then be used to provide a yardstick against which to measure the effectiveness of different perception algorithms.

Heuristics. The truly important difficulties characteristic of vision problems are computational in nature rather than statistical. These difficulties quickly become apparent when we attempt to carry out the prescriptions of classical statistical decision theory. We find that we are asked to optimize complex functions over high dimensional spaces and to solve computational problems which are known to be NP-complete. In essence the vision task may be thought of as a search through M for the model which has the highest probability (or least expected loss) of having produced the sensation. To deal with the

computational complexity of this task, it is natural to settle for a heuristic rather than optimal search procedure. It is these heuristic procedures which ultimately give an AI flavor to the structures of visual knowledge representation and their interactions. The statistical decision theory is valuable, however, because it provides a semantically consistent platform from which to evaluate and understand the heuristic techniques.

To achieve a fundamental understanding of the processes involved in model-based vision, it is useful to gradually move from well-understood simple problems toward the complex problems whose understanding eludes us. In particular, statistical pattern recognition has been very successful at providing a rigorously valid set of tools for problems in which the model spaces are of low dimension and low complexity. We would like to understand how these essentially geometric techniques connect with the essentially symbolic techniques used with some success in artificial intelligence work.

Bayes' Theorem. There are many schools of statistical thought, but we take only a Bayesian approach in this report. In most real vision tasks the noise is small and the data is plentiful. The difficulties of the task are not those which characterize the differences between statistical schools and the different approaches should give essentially equivalent results.

Let us briefly review the use of Bayes' theorem for discrete model and sensation spaces M and S which then consist of only a finite number of points. The theorem expresses the posterior probability $p(m \mid s)$ of each model given a sensation in terms of the prior distribution $\pi(m)$ on M, and the conditional probability $p(s \mid m)$ of the sensation conditioned on the model.

$$p(m \mid s) = \frac{p(s \mid m)\,\pi(m)}{p(s)} = Np(s \mid m)\,\pi(m) \qquad \text{(EQ 1)}$$

This probability distribution gives us all of the information about the models which exists in the samples. The distribution may be used in different ways, however, to answer different questions of interest. The denominator $p(s)$ is only a normalization factor which is the same for all models and so has no effect on any decisions. We will simply denote it by N in all later formulas. The most basic task we might attempt to use the posterior distribution for is to choose a single model to explain the current data. To make this choice given the posterior distribution, we must make use of a loss function which describes the penalty for each possible model choice in the context of each possible actual model. In the discrete case, the simplest loss function assigns zero error to the correct choice and a constant error to all other choices. The prescription for model choice which arises is to simply choose the model with the highest posterior probability, making an arbitrary choice in case of ties. This choice minimizes the average number of

mistakes we make. This choice is often called the MAP (Maximum Apriori Probability) choice. In situtations with continuous parameters it is common to use the mean square error in the values of the parameters as a loss function. The optimal estimator in this case chooses the mean value for each parameter. This is sometimes called the MVE (Minimal Variance Estimator). In the asymptotic limit as the number of data samples becomes large, all of the reasonable loss functions lead to the same model choice.

We will later see several situations in which we are not concerned with choosing the best model, but rather in choosing the best value v of a property of the models. If we want to minimize the probability of making an error in v, we must compare the different values of v by summing the posterior probability of all the different models which take on the value v. This is sometimes non-intuitive because the best choice of v may not be the value it takes on the highest posterior probability model. This may happen, for example, if lots of lower probability models with a given value of v are compatible with the data, while only a few higher probability models with a different value are.

The spaces S and M which arise in vision tasks often have special structure. A point in S will usually correspond to a whole collection of data whose elements may be governed by similar laws. For example, a dot image is comprised of many dots, each of which is drawn from the same probability distribution. This gives S the structure of a product space. M will also often be a product space, with a model being composed of several simpler submodels. The generation process of a point in M will often proceed in a hierarchical fashion, first choosing the values of certain parameters m_1 according to a prior $\pi_1(m_1)$ and then choosing the rest m_2 according to a distribution $\pi_2(m_2|m_1)$ which is conditionalized on m_1. Additionally, many of the possible conditional values will vanish or be vanishingly small, allowing for computationally efficient evaluation techniques which do not need to consider every possible explanation.

Dot pictures.

Real images have a number of artifacts which complicate analysis and mask certain essential features of the problem. In particular, the organization of images around pixels on a rectangular lattice and the vector valued nature of color image points are complications which are inessential to the processes of interest. We will therefore study a class of images which we call *dot pictures*. These consist of only a finite number n of points in the unit square $[0,1] \times [0,1]$. We do not give the dots any properties (such as color) other than their location and we assume that the locations are given precisely as a pair of real numbers. We may therefore also think of a dot picture as a point in the unit cube in the $2n$ dimensional real Euclidean space $\Re^{2n}$. There is an ambiguity

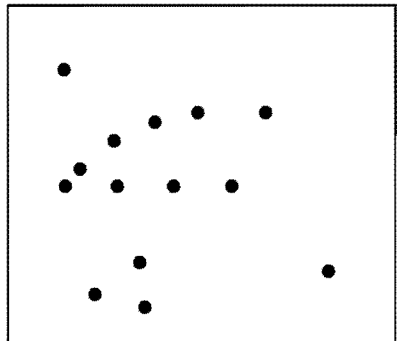

Figure 1. An example dot picture.

in this representation due to the possibility of permuting the points, but all of our models and analysis will treat the points identically and any permutation will be equivalent to any other.

You can see from figure 1 that the human perceptual system tends to impose substantial organization on a dot image, interpreting some dots as related to one another and others as being isolated. We perceive more complex entities such as lines, angles, curves, and triangles in even the simplest dot pictures. Simple experiments convince one that there is a complex set of rules which govern how we perceive such images. We are particularly attuned to relationships such as collinearity, parallelism, and equal distances. One can set up dot pictures that pit these relationships against another to determine their relative strengths and the nature of their interactions. It is natural to ask whether these interpretation mechanisms are arbitrary or whether there a setting in which they may arise from an underlying model. In this report we will discuss such a statistical basis and see how it explains the efficacy of certain of the phenomena of interest.

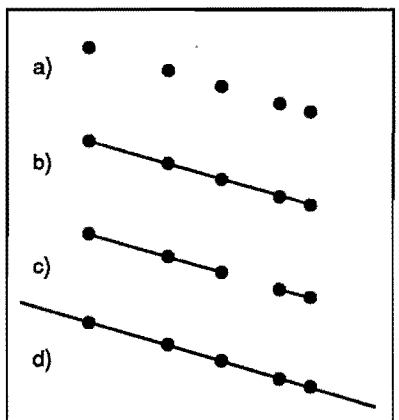

Figure 2. a) A dot picture with 5 dots. b) The preferred interpretation. c) Two lines where one seems better. d) The line is longer than necessary.

Occam's razor. An example of a phenomenon which we would like to understand has the flavor of Occam's razor (that simple explanations are to be preferred to complex ones). Figure 2a) shows 5 dots which we naturally interpret as having come from a single line segment as in b). The dots may be equally well explained by two segments as in c) or by a longer segment as in d). We prefer b) because it is in some sense the simplest explanation consistent with the data. Both c) and d) add extraneous features which don't appear to be suggested by the data. We will see in a later section that this intuitive argument may be made precise within the context of certain natural model spaces.

In the following sections, we will examine a series of successively more complex statistical models and techniques for recognizing them. We will be concerned both with the complexities of individual model objects and with the presence of multiple models. In each case we will describe a precise statistical model which we then simulate in order to measure recognition performance. The statisical models are hierarchical in nature, first choosing gross features such as the number and type of subfigure, then choosing the parameters of the subfigures, and finally choosing the dots which make up the dot image. We want the algorithms to be robust against small deviations in the dots, so we conclude with a final phase in which a small perturbation of size ε is added to the final dot locations. We are interested in the case in which ε is small and so will only study phenomena asymptotically to first order in ε. The exact form of the noise should be unimportant. We examine both Gaussian additive noise and noise which is uniform in a disk.

Discrete One-dimensional Point Models

Some of the manipulations with continuous distributions can be confusing, so we begin with a discrete version of the simplest model to highlight the essential features of later continuous developments. To further simplify the task, we will consider one-dimensional images in this section. We will assume that the interval [0,1] is quantized in steps of size α and will be interested in the limit in which α vanishes. The simplest kind of model consists of just points parameterized by their locations x. If there were no noise, then the dots generated from a point model would also have coordinate x and recognition would be trivial. With noise, the recognizer must decide how many model points there are and which dots go with which model point.

The simplest noise model is uniform over a distance ε centered on the model point location. We will assume that α vanishes more quickly than ε so that in the limit there are a large number of space points ε/α within the noise region, even though the noise region itself vanishes.

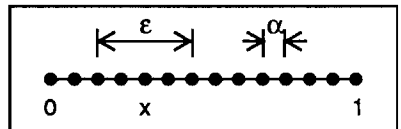

Figure 3. The discrete interval [0,1] with point spacing α, a model point x and the extent ε of uniform noise.

Single point, single dot.. The simplest point models consist of only a single point x. The model space M consists of the $1/\alpha$ possible locations for x. Let us take the prior distribution $\pi(x)$ on M to be uniform over these locations, so the prior probability of a particular x is $\pi(x) = \alpha$. If x generates a single dot d, it will be uniformly distributed over the ε/α points in the region $[x-\varepsilon/2, x+\varepsilon/2]$. We will ignore all end effects near 0 and 1 as being of vanishingly small probability. The conditional probability $p(d|x)$ of each dot in this size ε interval is therefore α/ε. Given a single dot d, we can use Bayes' theorem to determine the posterior distribution $p(x|d)$ of the point models x. With N for normalization, this will be

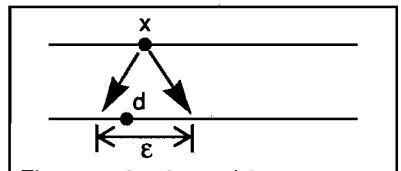

Figure 4. A point model at x generates a dot d with uniform probability in an interval of width ε.

$$p(x|d) = N\pi(x)p(d|x) = \frac{N\alpha^2}{\varepsilon} \qquad \text{(EQ 2)}$$

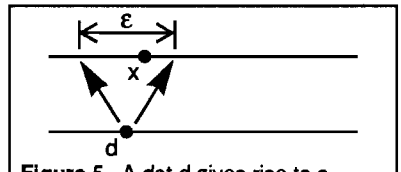

Figure 5. A dot d gives rise to a posterior model distribution over points x which is uniform in an interval of width ε.

for x in the interval $[d-\varepsilon/2, d+\varepsilon/2]$ and 0 elsewhere.

If our goal is to choose a single model and we use the simplest loss function, then each of the models in this interval is equally adequate. More realistically, we don't care about getting the right model exactly and would just like to be close. A commonly used loss function is the square of the distance between the predicted and actual model locations. The optimal choice of x under this loss function is the mean value of the distribution. In this simple case, the mean will be the center of the interval of possible models and is at the location of the dot, $x = d$.

Single point, n dots. When we generate n dots from a single point model x, the ambiguity in the choice of model drops. Again we choose

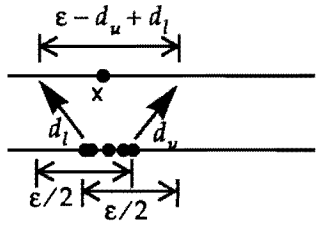

Figure 6. A set of n dots gives rise to a distribution uniform in an interval of models x.

a uniform prior $\pi(x)$ The conditional probability for the n dots given the model x will just be the product

$$p_n(d_1,\ldots,d_n|x) = \prod_{i=1}^{n} p_1(d_i|x) = \left(\frac{\alpha}{\varepsilon}\right)^n \qquad \text{(EQ 3)}$$

if each of the d_i is in the interval $[x-\varepsilon/2, x+\varepsilon/2]$ and 0 otherwise. Applying Bayes' theorem, we see that if the minimum of the n dot locations is d_l and the maximum is d_u, then the posterior probability $p(x|d)$ will be uniform in the interval $[d_u-\varepsilon/2, d_l+\varepsilon/2]$. of length $\varepsilon - d_u + d_l$. If the width of the interval determined by the dots is greater than ε, then the dots cannot have come from a single model point. The distribution of d_u and d_l is governed by order statistics and the mean of the width of the sample set $d_u - d_l$ is

$$\langle d_u - d_l \rangle = \frac{n-1}{n+1}\varepsilon. \qquad \text{(EQ 4)}$$

The expected uncertainty in the model x is therefore $(2\varepsilon)/(n+1)$ and so decreases linearly with the number of dots. As in the one dot case, it is natural to use a square error loss function and this gives rise to the optimal model at the midpoint of the region spanned by the dots:

$$x = \frac{d_l + d_u}{2}. \qquad \text{(EQ 5)}$$

Multiple point models. We begin to get interesting recognition phenomena only when we consider models which are composed of multiple points. We will choose the points in two stages, first choosing the number of points according to a distribution $\pi_{num}(m)$ and then given m we choose the coordinates $x_1, \ldots, x_n$ uniformly from the unit m cube. In this situation, given a set of dots we might be interested in determining the optimal number of model points m before choosing a particular set.

Two points, two dots. Let us begin by explicitly working out the case of using two dots to determine whether there are one or two model points, to see which features are important. We will denote the one point models by $(1, x_1)$ and the two point models by $(2, x_1, x_2)$. We will use d_1 and d_2 to denote the coordinates of the two dots. As above, the conditional probability $p(d_1, d_2|1, x_1)$ of the dots in the one point case is $(\alpha/\varepsilon)^2$ if the dots both lie in the interval $[x_1-\varepsilon/2, x_1+\varepsilon/2]$ and 0 otherwise. The figure shows the form of the probability distribution for two point models. The conditional probability $p(d_1, d_2|2, x_1, x_2)$ will be zero if either of the dots is not in one of the intervals, it will be $(\alpha/\varepsilon)^2/4$ if each dot is in a single interval, $(\alpha/\varepsilon)^2/2$ if one dot is in one interval and the other in both, and $(\alpha/\varepsilon)^2$ if both dots are in both intervals. The one point prior is

$$\pi(1,x_1) = \pi_{num}(1)\pi(x_1|1) = \pi_{num}(1)\alpha \qquad \text{(EQ 6)}$$

and the two point is

$$\pi(2,x_1,x_2) = \pi_{num}(2)\pi(x_1,x_2|2) = \pi_{num}(2)\alpha^2. \qquad \text{(EQ 7)}$$

We use Bayes' theorem to determine the posterior

$$\begin{aligned} p(1,x_1|d_1,d_2) &= N\pi(1,x_1)p(d_1,d_2|1,x_1) \\ &= N\pi_{num}(1)\alpha\left(\frac{\alpha}{\varepsilon}\right)^2 \end{aligned} \qquad \text{(EQ 8)}$$

if the ε sized interval centered on x_1 contains both d_1 and d_2 and 0 otherwise. Similarly,

$$\begin{aligned} p(2,x_1,x_2|d_1,d_2) &= N\pi(2,x_1,x_2)p(d_1,d_2|2,x_1,x_2) \\ &= N\pi_{num}(2)\alpha^2 C\left(\frac{\alpha}{\varepsilon}\right)^2 \end{aligned} \qquad \text{(EQ 9)}$$

where C is 0, 1/4, 1/2, or 1 depending on which dots are contained in which intervals. The figure shows the form of the distribution in (x_1, x_2) space, indicating the value of C in the different regions.

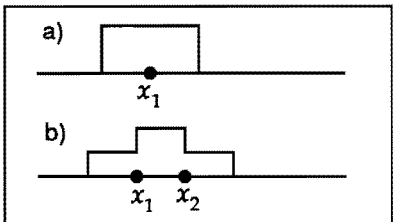

Figure 7. a) A single model point and the dot distribution it generates. b)Two model points and the dot distribution they generate.

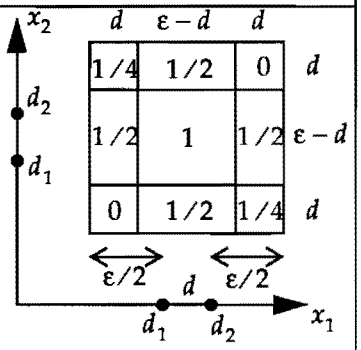

Figure 8. The form of the posterior in (x_1, x_2) space. The numbers give the value of the constanct C in the different regions of model space.

Given two dots a distance d apart, it is of interest to determine when the system should choose a one point model and when a two point model. We would therefore like to compute $p(1|d_1,d_2)$ and $p(2|d_1,d_2)$. To do this, we must sum the probabilities determined above over the possible values of x_1 and x_2. If the distance d between dots is less greater than ε, then it is impossible that they arose from a 1 point model and we must choose the 2 point explanation. Let us therefore analyze the cases where $d \le \varepsilon$. We obtain

$$\begin{aligned} p(1|d_1,d_2) &= \sum_{x_1} p(1,x_1|d_1,d_2) \\ &= \frac{\varepsilon - d}{\alpha} N\pi_{num}(1)\alpha\left(\frac{\alpha}{\varepsilon}\right)^2 \end{aligned} \qquad \text{(EQ 10)}$$

To do the corresponding sum for the 2 point case, we break it up into the parts in which both intervals contain one dot, one contains two dots, and both contain two dots. An examination of the geometry shows that the first case happens for $2(d/\alpha)^2$ choices of x_1 and x_2, the second for $4(d(\varepsilon - d)/\alpha^2)$ pairs and the third in $(\varepsilon - d)^2/\alpha^2$ cases. The posterior probability therefore becomes

$$p(2 \mid d_1, d_2) = \sum_{x_1, x_2} p(2, x_1, x_2 \mid d_1, d_2)$$

$$= N\pi_{num}(2)\,\alpha^2 \left(\frac{\alpha}{\varepsilon}\right)^2 \left(2\left(\frac{d}{\alpha}\right)^2 \frac{1}{4} + \frac{4d(\varepsilon - d)}{\alpha^2}\frac{1}{2} + \frac{(\varepsilon - d)^2}{\alpha^2} 1\right) \quad \text{(EQ 11)}$$

Expanding this out we obtain

$$p(2 \mid d_1, d_2) = N\pi_{num}(2)\,\alpha^2 \left(1 - \frac{d^2}{2\varepsilon^2}\right). \quad \text{(EQ 12)}$$

Combining the two results gives us the criterion that we should choose a 1 point model over a 2 point model whenever

$$\varepsilon - d > \frac{\pi_{num}(2)}{\pi_{num}(1)} \left(\varepsilon^2 - \frac{d^2}{2}\right). \quad \text{(EQ 13)}$$

If the priors are held constant as ε vanishes, then all of the terms on the right hand size are of second order in ε. To first order, then, we should always choose the one-point model when it is possible. Asymptotically, the ability of the more complex model to fit the data in more ways does not beat out the fact that the prior is lower because of the larger number of combinations that must be chosen from. Since to first order the criterion is $d < \varepsilon$, to second order it is

$$d < \varepsilon - \varepsilon^2 \frac{\pi_{num}(2)}{2\pi_{num}(1)}. \quad \text{(EQ 14)}$$

We reach the same first order conclusion if we look at the maximum posterior model as well. Because the probabilities are uniform, all one point models will have the same posterior distribution, as will all two point models which cover the dots in the same way. If $\varepsilon > d$, then again only the two point model is possible. If $\varepsilon < d$, then from equations 7 and 8 we see that the two point models whose two segments each cover both dots are preferred. The criterion for choosing a two point model over a one point model is then

$$\pi_{num}(1) / \pi_{num}(2) > \alpha \quad \text{(EQ 15)}$$

Asymptotically, this again vanishes. Notice that to higher order there are pairs of dots whose distance is near ε for which the whole set of two point models has higher probability than the set of one point models, but the best one point model has higher probability than the best two point model.

m points, n dots. Let us now consider the general case in which there are up to m point models which generate n dots. In light of the

discussion above, we will only seek the maximum apostiori probability model. Exactly as above the prior distribution will be:

$$\pi(m, x_1, \ldots, x_m) = \pi_{\text{num}}(m)\,\alpha^m. \qquad \text{(EQ 16)}$$

As above the results asymptotically do not depend on π_{num} if each of its values are non-zero and held constant as α goes to zero. We will therefore choose it to be uniform for each number of points up to n for convenience. As above, an isolated model point corresponds to a uniform interval whose points have probability $\alpha/(m\varepsilon)$. If a dot is contained in k intervals, then its probability will be k times this factor. If for each k there are n_k dots contained in k intervals, then we may compute the probability of those dots as

$$p(d_1, \ldots, d_n \mid m, x_1, \ldots, x_m) = \left(\frac{\alpha}{m\varepsilon}\right)^n \prod_k k^{n_k} \qquad \text{(EQ 17)}$$

The posterior is proportional to the product of these two terms. Because α is vanishing, the prior would like to make the number of model points m as small as possible. The other term would like there to be a large number dots in the overlap of a large number of segments. The first effect overwhelms the second. In fact we can easily see that overlaps of three segments will not occur for the MAP model. If there is a place where three segments overlap, then one of the segments must be completely contained in the region covered by the other two. We get the same dot coverage by eliminating this segment and so such triple intersections cannot occur for the MAP model. If we restrict ourselves to models with only pair overlaps, then only n_2 contributes. The posterior probability for these models is

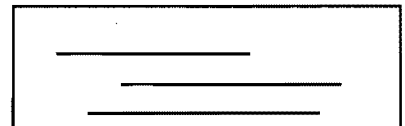

Figure 9. When three segments have a point of common overlap, one of the segments must be contained in the union of the other two.

$$p(m, x_1, \ldots, x_m \mid d_1, \ldots, d_n) = N\frac{\alpha^m}{n}\left(\frac{\alpha}{m\varepsilon}\right)^n 2^{n_2} \qquad \text{(EQ 18)}$$

We can see that the MAP model will firstly have the minimal number of model points possible, and then subject to that will maximize n_2.

Let us first consider the problem of finding the minimum number of segments of size ε which cover a set of dots in the interval. It turns out that this is algorithmically quite simple. We need only lay down the intervals from left to right, putting the left edge of the first interval at the first dot and putting down successive intervals with no overlap so that their left edge is on a dot. To see that this gives the minimal number of intervals, consider the left endpoints of these segments. Each endpoint is further than ε from its two neighbors and therefore cannot be in the same segment with them. There must therefore be at least as many segments as there are of these endpoints, so the covering given is the smallest possible. (The complementary problem of finding the set of a given number k of segments with the shortest total length

Figure 10. A set of dots is minimally covered by intervals by proceeding from left to right.

which covers a set of dots is also easy to solve efficiently. Simply choose the n-k-1 largest gaps to remain empty and fill in the rest with segments.)

How might we then maximize n_2 subject to fixing the number of segments to be the minimal number? In the proof above we picked out a set of key dots such that the ith dot must be contained in the ith segment. We are free to slide the segment about on this dot, however, to maximize the overlaps. It doesn't hurt to slide the final configuration of segments so that each segment has its left end on one of the dots. With this convention, each segment can potentially be in only a finite number of states (one for each dot it can reach to the left of its key point). If we fix the state of one of the segmenst, it restricts the states of the segment to its right to those which cover all the points. We may efficiently find the optimal segment settings by dynamic programming. We proceed from right to left, determining for each segment the maximal number of dots to its right which are included in two segments for each of its possible states. If we have determined this for the rightmost i segments, we get it for the (i+1)th segment by maximizing the sum of newly created dots in segment pairs with the previously computed number to the right over the legal positions of the segment to the right. In this way we obtain the optimal choice of segments in a time $O(n^2)$.

In this section we will discuss some typical model spaces for generating dot pictures. All of our models will lead to probability distributions on the unit square $\rho_m(x, y)$. A dot picture is formed from such a model by independently drawing n dots from this distribution. Our models will themselves be constructed from simpler models derived from points, line segments, circles, arcs, etc.

Point models. The simplest model is a single point. If there is no noise then the corresponding dot picture will always just consist of n dots located exactly at the model point's location. Because in reality our computational processes will use finite precision arithmetic, it is useful to consider computational procedures which are robust to small amounts of noise. We really are interested in the small noise limit, however, and this is one of the important simplifications over the general case which makes heuristic techniques possible.

When there is noise, the model point will not generate dots exactly at its location. There are many compact probability distributions which may be used to describe small deviations from the model center. We are really only interested in results which are fairly independent of the exact details of this small noise. Such independence may be obtained under reasonable conditions if the magnitude of the noise is parameterized and we consider the small noise limit. We will use ε for this parameter. Two obvious candidates are the distribution which is uniform in a disk centered on the point, parametrized by the radius of the disk and the two-dimensional Gaussian distribution parameterized by the standard deviation. We will use the Gaussian because of several nice properties. If the real noise is additively due to independent sources, then the central limit theorem shows that the actual distribution will approach Gaussianity. The convolution of two Gaussians is a Gaussian. Most importantly, because the square of the Euclidean distance appears in the exponent, Gaussian distributions are naturally related to geometric concepts and thus tend to give rise to computationally tractable subproblems.

Gaussian point model. The point model space M_{pt} is two-dimensional and may be parameterized by the coordinates of the point, which we will denote by (x_m, y_m). The image distribution is just a Gaussian centered at the model point:

Figure 11. A Gaussian point model gives rise to a cloud of dots centered on the model point.

$$\rho_{(x_m, y_m)}(x, y) = \frac{1}{2\pi\varepsilon^2} e^{-\frac{(x-x_m)^2 + (y-y_m)^2}{2\varepsilon^2}} \qquad \text{(EQ 19)}$$

A dot picture arising from such a model will consist of a cloud of dots centered at the model point. Notice that ρ achieves its maximum at (x_c, y_c) and its value there is

$$\rho_{\max} = \frac{1}{2\pi\varepsilon^2}. \qquad \text{(EQ 20)}$$

Point prior. What should the prior on the model space M_{pt} be? It is natural to choose priors which are as uniform as possible on the corresponding model spaces. We will denote the prior distribution by $\pi(x_m, y_m)$. Here we will just take this to be everywhere equal to 1. Notice that if we want the dot picture to be contained in the unit square, there may be difficulties near the edge in that dots will lie outside with non-zero probability. We will ignore all such difficulties because their probability becomes vanishingly small as the noise magnitude vanishes.

Segment model. A natural way to generate dots from a segment is to first choose a point on the segment and then perturb it by additive Gaussian noise. We will take the sampling from the segment to be uniform. The image distribution is the convolution of the Gaussian with

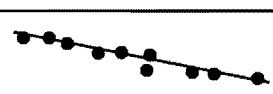

Figure 12. A Gaussian segment model gives rise to a cloud of dots along the segment.

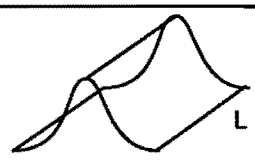

Figure 13. The distribution due to a Gaussian segment looks like a length L slab with Gaussian cross section.

the singular distribution supported on the segment. It will be peaked along the segment, uniform along it (except near the endpoints) and the cross section will be a Gaussian with width ε. We can see this by considering a horizontal line of length L whose left end is at the origin. Other lines of the same length generate distribtuions which are just rigid rotations and ranslations of this one. If $G_{(x_c, y_c)}(x, y)$ is the Gaussian distribution given above, then the line distribution is

$$\rho(x, y) = \frac{1}{L}\int_0^L G_{\alpha, 0}(x, y)\, d\alpha \qquad \text{(EQ 21)}$$

If the line is much longer than width of G, then aside from end effects, the x part of the Gaussian will integrate out and we will be left with just a length L slab with a Gaussian cross section. For points in the strip above and below the segment, the probability depends on only the distance to the segment and has the value:

$$\rho(x, y) = \frac{1}{\sqrt{2\pi}L\varepsilon} e^{-\frac{y^2}{2\varepsilon^2}} \qquad \text{(EQ 22)}$$

The value at the maximum (on the line segment itself) is

$$\rho_{max} = \frac{1}{\sqrt{2\pi}L\varepsilon}. \qquad \text{(EQ 23)}$$

We will see that it is the 1/L factor which will cause shorter lines to be prefered to longer ones.

Segment prior. The space of segments is a four dimensional space (parameterized by the coordinates of the two endpoints, for example). The choice of segment prior is not as clear as it was for points. For lines of a given length, we would like the probability to be uniform under rotations and translations of the segment. A choice which is simple to implement is to choose the enpoints uniformly in the unit square.

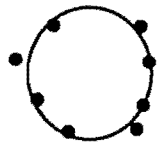

Figure 14. A Gaussian circle model gives rise to a cloud of dots along the circle.

Circle models. It is again natural to pick samples uniformly from a circle and then additively perturb them by a Gaussian. Taking the limit of small noise is like zooming in on the circle and it looks more and more like a straight line. In fact the conclusions reached above for segments also hold for circles in which the length L is replaced by the circcumference $2\pi R$ where R is the radius of the circle. The distance to the line in the exponent of the Gaussian is replaced by the distance to the circle. The maximum value of the distribution is achieved on the circle itself and has value:

$$\rho_{max} = \frac{1}{\sqrt{2\pi}2\pi R\varepsilon}. \qquad \text{(EQ 24)}$$

Circle prior. The space of circles is three-dimensional (parameterized by the coordinates of the center and the radius, for example).

Multiple point models. More interesting model classes arise as combinations of the previous ones. The simplest such model would consist of 2 model points weighted equally. The resulting distribution will be a mixture of two Gaussians centered at the two point locations:

$$(x, y) = \frac{1}{2}\left(G_{(x_{c1}, y_{c1})}(x, y) + G_{(x_{c2}, y_{c2})}(x, y)\right) \qquad \text{(EQ 25)}$$

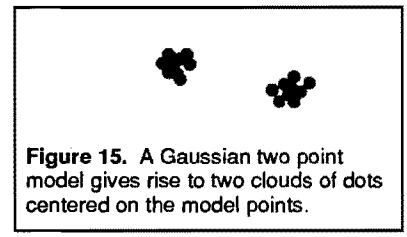

Figure 15. A Gaussian two point model gives rise to two clouds of dots centered on the model points.

This models space is four dimensional. In the limit of small noise, the probability density at a point is dominated by the contribution of the closer model point. The distribution maxima are at the two model point locations and are half the single point maximum:

$$\rho_{max} = \frac{1}{4\pi\varepsilon^2}. \qquad \text{(EQ 26)}$$