Perception, 1982, volume 11, pages 557-576

Equation counting and the interpretation of sensory data

Whitman A Richards, John M Rubin, Donald D Hoffman
Department of Psychology, Massachusetts Institute of Technology, 74 Amherst Street, Cambridge, MA 02139, USA
Received 3 August 1981, in revised form 28 February 1982

Abstract. Many problems in biological information processing require the solution to a complex system of equations in many unknown variables. An equation-counting procedure is described for determining whether such a system of equations will indeed have a unique solution, and under what conditions the solution should be interpreted as 'correct'. Three examples of the procedure are given for illustration, one from auditory signal processing and two from vision.

1 Introduction

Biological systems routinely make good use of their sensory data. How organisms interpret patterns of sensory activity as events in the external world is the fundamental problem of perception. What makes the problem difficult is that the mapping of world events onto a sensor is typically many-to-one. A three-dimensional world is flattened onto the retina; sounds from all directions are funneled into a single signal of acoustic intensity varying with time. How might useful inverse mappings, from sensory activity to world events, be discovered? In principle, there are an infinity of interpretations of sensory events. Under what conditions can the correct interpretation be found?

Our approach is to consider types of sensory data that can be formally related to the external events that generated these data. For example, the image intensity of a patch of surface as seen on a retina will depend upon the strength and position of the illuminant, as well as the orientation and reflectance of the surface. The relation between these external physical variables and the observed image intensity can be put in the form of an equation (Horn 1977). On one side of the equation, there is only the image intensity, which is known. On the other, there is a complicated function of several external physical variables, the values of which are unknown. If the observer wishes to know the value of one of these physical variables, say the surface reflectance, then in some sense he must 'solve' this equation. Clearly with so many unknowns only one such equation is not sufficient to recover the reflectance. More information must be sought to restrict the number of solutions. For example, perhaps the observer can change his viewing angle to obtain another intensity measurement which is related to the first in a known way, thereby obtaining another equation. Or better yet, perhaps the orientation of the surface is known at some position, from which one can deduce the illuminant direction. To make the problem tractable, several such additional relations or equations must be found until the total number of equations equals the number of unknowns. Only then is there a chance that we can find a solution to the reflectance of the surface. This 'solvable' set of relations or equations then expresses a minimum set of conditions required to 'solve' the perceptual problem. (Whether or not the perceptual device actually makes use of this possible solution is a separate issue.) The technique of analysing a given sensory problem in terms of such equations that relate 'knowns' to 'unknowns' is called 'equation counting'.

The paper begins with a rather simple example of equation counting, namely, the detection of a narrow-band signal in noise. This problem involves only linear equations, but still illustrates the general features of the approach and raises three issues: (i) independence of the equations; (ii) constraints needed to yield a unique solution; and (iii) whether this unique solution is indeed 'correct'. We then introduce a theorem by Bezout which is needed to place bounds on the number of possible solutions to polynomial equations, as well as a Jacobian test for the independence of these equations. Finally, two other problem examples are given to illustrate further details. One example concerns recovering structure from visual motion; the other shows why three spectral samples are needed to distinguish shadows from reflectance changes.

2 A classic problem

A problem faced by many animals is the need to isolate a narrow-band species-specific signal from the background noise. Although examples may be found in every sense modality, the clearest probably occur in audition. Consider the bird listening to the call of its mate in the forest of other sounds, the dog perking his ears at his master's whistle, or the moth's task of isolating the cry of the bat as it homes in for its next meal. In each case the signal is confined to a relatively narrow band, as illustrated in figure 1, whereas the competing noise is much broader. Given that the frequency band of the signal is known (as it would be for the bird or the moth), how many intensity samples must be taken to isolate the signal from the noise?

Clearly, by referring to figure 1, we see that sampling in the signal band at frequency ν_1 will not allow us to isolate the signal. More formally, the ear will receive intensity I, at frequency ν_1 equal to the sum of the power produced by each source:

$$I(\nu_1) = S(\nu_1)+N(\nu_1), \qquad (1)$$

where $S(\nu_1)$ is the power of the narrow-band signal at ν_1 and $N(\nu_1)$ is the background noise at the same frequency. Since only I is available to the listener, S and N cannot be separated, for we have only one equation in two unknowns, S and N. More generally, if we allow additional samples at time intervals t_j, then equation (1) can be generalized to

$$I(\nu_1, t_j) = S(\nu_1, t_j)+N(\nu_1, t_j). \qquad (2)$$

Thus, for T time samples we will obtain T equations in $2T$ unknowns, which will not permit a unique solution for S.

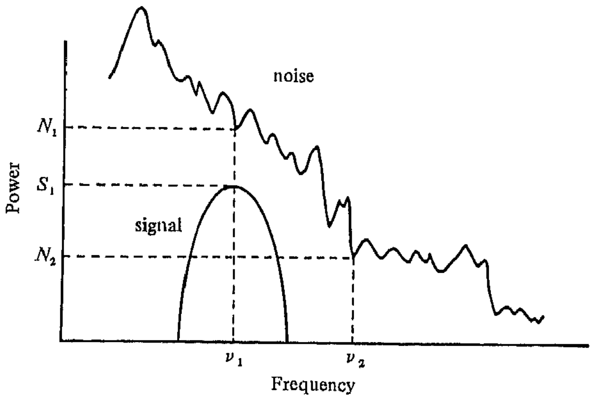

Figure 1. An illustration of a narrow-band signal against a background of noise. The noise is broad-band with a constant time-averaged spectrum.

Let us now make the obvious next step and consider frequency samples outside the signal band. The frequency ν_1 in equation (2) then becomes indexed to ν_i. However, since the signal is zero outside the band at ν_1, $S(\nu_i, t_j) = 0$ for $i \neq 1$. These conditions may be expressed as two families of equations:

$$I(\nu_i, t_j) = S(\nu_i, t_j) + N(\nu_i, t_j), \tag{3a}$$

$$S(\nu_i, t_j) = 0, \qquad (i \neq 1). \tag{3b}$$

Let F and T be the number of frequency and time samples, respectively; then there will be a total of $F \times T$ equations of form (3a) and $(F-1) \times T$ equations of form (3b). The total number of equations is thus $2FT-T$. Similarly, the total number of unknowns will be FT for S and FT for N, or $2FT$. In order to solve uniquely for S, the minimum condition is that the number of equations E equals (or exceeds) the number of unknowns U:

$$E \geqslant U. \tag{4}$$

For solution, equations (3a) and (3b) thus must pass the following inequality test:

$$2FT - T \geqslant 2FT, \tag{5}$$

or

$$0 \geqslant T,$$

which fails since $T \geqslant 1$. Thus, regardless of the number of time and frequency samples, a narrow-band signal cannot be extracted from the broad-band noise without specifying further constraints upon either the signal or the noise.

2.1 *Flat-noise condition*

Very often noise is relatively constant over frequency (or time), for example the hum of an air conditioner, a steady wind flow passing the body, or even body noise. This condition can be expressed by the following relation:

$$N(\nu_1, t_j) = N(\nu_i, t_j), \tag{6}$$

where $i \neq 1$ and ν_1 serves as the reference frequency. We now see that for a total of F frequency samples equation (6) adds $(F-1)T$ equations but no more unknowns. Applying the inequality test (4), we now find:

$$(2FT - T) + T(F-1) \geqslant 2FT, \tag{7}$$

or

$$FT \geqslant 2T,$$

or

$$F \geqslant 2. \tag{8}$$

Thus, the minimum condition for a unique solution occurs for two frequency samples at any temporal interval. If we ignore the time variable, equations (3a), (3b), and (6) then become

$$\begin{aligned} I(\nu_1) &= S(\nu_1) + N(\nu_1), \\ I(\nu_2) &= S(\nu_2) + N(\nu_2), \\ S(\nu_2) &= 0, \\ N(\nu_1) &= N(\nu_2). \end{aligned} \tag{9}$$

We now have four equations in four unknowns, which allows us to solve for $S(\nu_1)$, given that the noise spectrum is flat.

2.2 *Independence and uniqueness*

Although two frequency samples plus the constraint of 'flat noise' yield the same number of equations as unknowns, these equations must be shown to be independent. Certainly we can reduce equations (9) to obtain an explicit solution for $S(\nu_1)$, thereby demonstrating independence. However, in the more complex cases normally encountered such a reduction is often difficult or may be impossible (for example if fifth-degree polynomials are involved). We therefore seek a more general test for independence.

In the above example the obvious test is to recast equations (9) so all the unknowns are on the right-hand side (RHS) of the equality, and all the knowns are on the LHS. Then the determinant of the coefficients of the RHS can be calculated. By Cramer's Rule we know that, if this determinant is not zero, then the equations have a unique solution (Thomas 1951). To proceed, equations (9) are rearranged so the unknowns are ordered in the sequence $S(\nu_1)$, $N(\nu_1)$, $S(\nu_2)$, $N(\nu_2)$ and are each aligned in their separate columns on the RHS of the equality. Since there are four unknowns and four equations, the matrix of the coefficients of the unknowns will be as follows:

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} . \qquad (10)$$

The determinant of this matrix is easily found to be nonzero (its value is 1), and hence the set of equations (9) must have a unique solution.

We now can proceed with confidence to find the following solution for $S(\nu_1)$:

$$S(\nu_1) = I(\nu_1) - I(\nu_2). \qquad (11)$$

2.3 *Corroboration and constraint*

Unfortunately, *any pair* of sensory intensities $I(\nu_1)$ and $I(\nu_2)$ will provide a value for $S(\nu_1)$. How do we know that the obtained value for $S(\nu_1)$ is indeed correct? Clearly, if the noise stimulus is not flat over frequency, but varies as shown in figure 1, then the solution for $S(\nu_1)$ will be wrong because the assumed condition does not hold. Without some evidence supporting the flat-noise assumption, a meaningful interpretation of the intensity values $I(\nu_1)$, $I(\nu_2)$ cannot be made.

Ideally, any assumed condition, such as the flat-noise condition, that is introduced to match the number of equations to the unknowns should be a regularity in the world or a 'law' that is never (or rarely) broken by nature. Such conditions are difficult to discover, but when found and introduced into the system of equations provide powerful *constraints* on the solutions (Huffman 1971; Waltz 1975). Often the constraint may be a statistical regularity (Witkin 1980; Pentland 1980). Poor choices for constraints are those conditions that are very narrow and restrictive and which do not capture a very general property of the world.

In the case of detecting a narrow-band signal in flat noise the imposed condition is very restrictive. However, some attempt can be made to verify the validity of invoking this condition. For example, one possibility might be to examine other frequencies to see if the relation $N(\nu_1) = N(\nu_i)$ holds for a range of frequencies outside the signal band. [Note that the solution for $S(\nu_1)$ should also hold.] If so, then the chance that the flat-noise condition is invalid is reduced, although the uncertainty is never eliminated. Sampling at additional frequencies thus provides some (weak) corroboration for the interpretation, increasing its likelihood. (In fact, the condition assumed here has merely been replaced by another, less restrictive assumption about the smoothness of waveforms.) Stronger forms of corroboration will be discussed in later sections.

Finally, it should be noted that in cases where the imposed conditions are not verifiable, the appropriateness of the condition can often be rejected quite easily. For example, if $S(\nu_1)$ is found to be negative, then, since negative signals are not physically realizable, the assumption must not be valid. This strategy of rejecting certain conditions or possible states of the environment has been found useful elsewhere (Rubin and Richards 1981).

3 Nonlinear (polynomial) equations

3.1 *Bezout's Theorem*

In the above example all of the equations were linear, and simple techniques of linear algebra could be used. What if one or more of the equations were quadratic or a still higher-degree polynomial? In such cases, which are quite common, each nth order polynomial will at most have n distinct roots. How many possible solutions will there be if there are M polynomial equations of degree N? Can we even guarantee that there will in fact be a finite set of solutions? If this cannot be guaranteed, then the test that states that the number of equations E should at least equal the number of unknowns U is not useful, and the simple equation-counting procedure collapses at the onset. Fortunately, Bezout's Theorem tells us under what conditions a finite set of solutions can be found to N equations in N unknowns, and just what the maximum number of solutions will be (Van der Waerden 1940).

Theorem (Bezout): A set of N independent polynomial equations in N variables will have a maximum number of generic solutions equal to the product of the degrees of the equations[1].

The above theorem is critical for our procedure because it states that if the relations among the N variables can be cast as N independent polynomial equations (perhaps by a change in the form of the variables), then there will be a finite set of isolated solution points. Furthermore, we know the upper bound on the number of possible solutions. (See Appendix II for a brief discussion of a generalization of Bezout's Theorem by Sard to include any set of smooth functions on manifolds.) For linear equations it is clear that the product of the degrees of the equations will always be one, and only one solution set will be found. For third-order equations, which may include terms such as xyz, or y^2z, the number of possible N-tuples of variables that satisfy the N equations can be quite high. Among these is the physically meaningful solution that we seek, provided our hypotheses are correct.

3.2 *The Jacobian test*

Bezout's Theorem states that, in principle, N polynomial equations of any degree can provide a solution to N unknowns, if the equations are independent. In our simple first example the determinant of the matrix of coefficients of the unknowns was used to check for independence. More generally, the Jacobian of the set of equations should be evaluated (Kendig 1977; Guillemin and Pollack 1974). The Jacobian is formed by taking all N partial derivatives of each of the N equations $(\partial f_1/\partial x_1, \partial f_2/\partial x_2, ..., \partial f_n/\partial x_n)$, and placing these partial derivatives in an $N \times N$

[1] By a generic solution we mean that a slight perturbation in the values of the variables will not alter the solution appreciably (as would be the case if the solution were the special case of two circles just grazing each other rather than intersecting, for example).

matrix, where the columns represent each unknown and the rows correspond to the equations. Clearly, for linear equations, the Jacobian is simply the matrix of the coefficients of the unknowns of each equation.

Jacobian test (for independence): If the determinant of the Jacobian of the system of N equations in N unknowns is nonzero, then a countable set of isolated solution points can be found.

This test is simply an application of the Inverse Function Theorem, which gives a condition for a one-to-one and onto mapping between real variables. Note that if the determinant of the Jacobian collapses to zero (by a loss of rank), then this is *not* a proof that solution points cannot be found. The Jacobian test is therefore a test for sufficiency, not necessity.

3.3 *Summary of procedure*

To apply the equation-counting method to the recovery of event descriptions from limited sensory data we therefore proceed as follows:

(i) Set up polynomial equations describing the mapping of the external (unknown) variables into the (known) sensory data.

(ii) Embody as many constraints as necessary in the form of additional polynomial equations relating the variables in order that the total number of equations equals the number of unknowns that are to be recovered. Whenever possible, choose 'constraints' that can be supported by the data. Those that capture a regular or inconsistent property of the world are the best choice.

(iii) Apply the Jacobian test to demonstrate that the equations are independent. Bezout's Theorem then guarantees that there will be a finite number of solution points. If the Jacobian test fails, try to discover new constraints. (See also section 5.5.)

(iv) Proceed to solve for the variables of interest. (We know of no simple heuristics for this step.)

(v) Demonstrate that all constraints and conditions are valid. Usually this will involve taking an extra, independent measurement and verifying that the same solution is obtained. Some care must be taken with this step, however, as will be seen in the examples to follow.

(vi) The sensory data may now be given a preliminary interpretation. However, a final interpretation should await two further tests to be described subsequently. One is the exclusion of competing interpretations, the other is corroboration by means of an independent system of equations. (See section 6.)

4 Example 1: recovering structure from motion

The difference in visual impressions between a static scene and a dynamic movie is often quite striking. Somehow the motion created by viewing a rapid sequence of frames will transform an ambiguous 2-D shape into a vivid 3-D structure. Perhaps the most common example of this phenomenon occurs when we walk, run, or drive and immediately *know* the spatial configuration of the objects about us, regardless of whether we use two eyes or one. Although Ullman (1979) has shown how the spatial relations may be recovered from motion information in the general case, we wish to consider a simpler version of the same problem that has a more compact solution: namely, given a person in locomotion, how can he recover the orientation of the surface on which he walks?

Let the surface be covered with markings, or, for convenience, let a short 'stick' lie on the surface patch of particular interest. Then if the observer looks at the center of the stick as he moves ahead, the image of the stick as seen on his retina will rotate and change length as shown in frames F1, F2, and F3 of figure 2. Because the stick lies in a plane of fixed orientation relative to the moving observer, the orientation of the surface patch can be specified by the axis of rotation of the stick. The problem then is equivalent to recovering the axis of rotation of a rotating rod seen by a stationary observer.

Figure 2 illustrates the general form of this common problem. The stick or rod is rotating in 3-D and is projected onto a single 2-D retina. Let each of these retinal images be discrete time samples or frames as in a TV. Given only the three (or more) ambiguous 2-D image frames F1, F2, F3, how can the axis of rotation of the rod be recovered? This is a task that is solved easily by the human observer, although no information other than the 2-D motion of the end-points of the rod is available (Johansson 1975).

The inset to figure 2 shows the actual three-dimensional relation between the viewer, the rotating rod, and the axis about which the rod is spinning. Note that the axis of rotation (which defines the surface plane) can be any stationary vector and need not be vertical nor parallel to the xy image plane. The problem is to recover the correct axis of rotation (as well as the length of the rod).

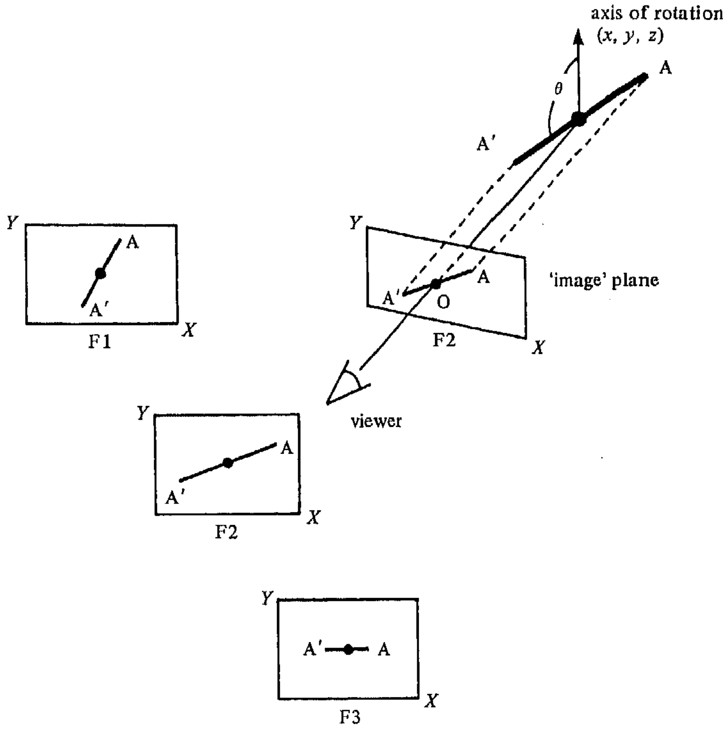

Figure 2. A simple rod rotating in 3-D about its midpoint.

4.1 *Rigid rod and rotation in a plane*

Let the coordinate system be centered at the projection of the midpoint of the rod. Then since the distance OA = OA′, we need consider the motion of OA only. Let the three-dimensional coordinates of end A be (x_1, y_1, z_1) for frame 1 and (x_i, y_i, z_i) for frame i. Then since the stick is a rigid rod, we have the constraint that the rod length remains constant for any frame:

$$x_1^2+y_1^2+z_1^2 = x_i^2+y_i^2+z_i^2 . \qquad (12)$$

For N frames the relation (12) will yield $(N-1)$ equations, each in two unknowns, z_1 and z_i (since x_i, y_i are observables in the image plane). So far we thus have $(N-1)$ equations in N unknowns.

To embody the condition of rotation about a fixed axis, we note that the angle θ between OA and its axis must remain constant. This can be expressed by forming the dot product between the rod segment OA with the presumed axis of rotation, **N**:

$$\mathrm{OA}_i \cdot \mathbf{N} = \cos\theta, \qquad (13a)$$

where the subscripted OA_i indicates the 2-D projection of the 3-D length OA onto the ith frame.

On letting the end position of the unit axial vector **N** have the coordinates x_0, y_0, z_0, equation (13a) reduces to

$$x_i \cdot x_0 + y_i \cdot y_0 + z_i \cdot z_0 = k \cos\theta\,, \qquad (13b)$$

where $k = (x_i^2+y_i^2+z_i^2)^{1/2}$.

But rotation in a plane requires that the angle θ between the axis **N** and OA be $\frac{1}{2}\pi$. Hence, $\cos\theta = 0$ and the value of k is irrelevant. For N frames relation (13b) thus gives us N equations in three more unknowns: x_0, y_0, z_0. However, because the length of the rotation axis is irrelevant also, **N** can be taken as the unit vector and we obtain the additional equation

$$x_0^2+y_0^2+z_0^2 = 1. \qquad (13c)$$

Altogether we thus have $(N-1)+N+1$ equations (E) in $N+3$ unknowns (U): z_i, x_0, y_0, z_0. (Note that all of these equations are polynomial.) The minimum number of equations can then be determined from the relation $E \geqslant U$:

$$2N \geqslant N+3, \qquad (14)$$

or

$$N \geqslant 3. \qquad (15)$$

4.2 *The Jacobian test*

The next step is to demonstrate that the equations (12) and (13) form a set of independent equations. We thus examine the Jacobian for $N = 3$ to see if its rank is maintained. It should be recalled that x_i, y_i for $i \neq 0$ are given in the image plane; the partial derivatives of z_i in equation (12) for $i = 2, 3$ yield the first two rows of the following matrix, while the remaining rows come from equations (13b) and (13c) respectively:

$$\begin{bmatrix} 2z_1 & -2z_2 & 0 & 0 & 0 & 0 \\ 2z_1 & 0 & -2z_3 & 0 & 0 & 0 \\ z_0 & 0 & 0 & x_1 & y_1 & z_1 \\ 0 & z_0 & 0 & x_2 & y_2 & z_2 \\ 0 & 0 & z_0 & x_3 & y_3 & z_3 \\ 0 & 0 & 0 & 2x_0 & 2y_0 & 2z_0 \end{bmatrix} . \qquad (16)$$

Evaluation of the determinant by MACSYMA shows that it is generally nonzero. However, certain relations between the variables may cause the Jacobian to drop rank. Some of these failure conditions can be noted by factoring the determinant. (Note that such failure conditions provide instances where any perceptual system that interprets data in accord with the system of equations should also fail. The factors thus provide example experiments for instant psychophysics.)

4.3 *Bezout's Theorem and uniqueness*

Although the sets of equations (12) and (13) are shown to be 'independent' by the Jacobian test, Bezout's Theorem tells us that we may have up to $2^6 = 64$ possible solutions. (This is the product of the degrees of the six equations.) Which of these solutions do we pick?

Fortunately, it can be shown by algebraic reduction of the six equations that of these 64 possible solutions only two have real values—and one of these is simply a 'reflection' of the other about the image plane (Hoffman and Flinchbaugh 1982).(2) Thus, three snapshots or 'frames' showing the x, y positions of the end-points of a rotating rod are sufficient to solve for the rod length and its axis. (The reflection causes an ambiguity only in the direction of motion and orientation of the rod.) But since *any* triplet of x, y positions will yield a solution, how do we know that the measurements were taken from a rotating rod and not from a random set of points? Clearly, if the solution is imaginary, that set of triplets can be excluded. Are there then any real solutions that can arise from arbitrary triplets? If so, these triplets would yield false solutions. However, the probability of such false solutions can be shown to be zero (Hoffman and Flinchbaugh 1982). Nevertheless, as will be discussed below, additional tests still must be performed before accepting an interpretation of the data.

4.4 *Corroboration*

In addition to the problem of isolating a unique solution point, it is also necessary to show that the 'unique' solution is indeed plausible. (If the unique solution is not physically realizable, it can be rejected immediately.) In the case of the rod rotating in a plane about a fixed axis three frames (or snapshots) were sufficient to solve the six polynomial equations and to obtain a unique solution for the length of the rod and its axis of rotation. However, are we guaranteed that no other set of conditions could generate the data? Clearly not, for if the simple rod rotation is simulated in the laboratory on a TV monitor, then one obvious interpretation is that there are two points moving on the face of the TV. (In fact, if reflections appear on the screen so that strong 3-D cues are present, then the illusion of a rod rotating in 3-D is lost.)

Before a final interpretation is made it is therefore prudent to corroborate the solution to increase the probability of a correct interpretation. This can be accomplished by analyzing an independent set of data or hypotheses that are based on entirely distinct physical constraints. (In the case of structure from motion, stereopsis may be used.) When no corroboration is possible, it seems reasonable to accept the interpretation that is most favored by the real-world statistics(3).

5 Example 2: interpreting shadows and highlights—hidden dependencies

Quite often when the equation-counting method is used the constraint equations contain hidden dependencies that cause the Jacobian to drop rank and its determinant to equal zero. There are two general procedures for handling this situation so that an interpretation of the data can be made. The first is simply to introduce another independent constraint, the second is to identify the dependency and to reduce the number of physical variables accordingly. The disambiguation of shadows and highlights illustrates these two methods.

(2) When algebraic reduction is not possible, a useful strategy is to generate data from several known but arbitrary configurations, and by numerical evaluation determine if the correct solution is obtained (Ullman, personal communication). Numerical evaluation is recommended in any case as a further check for the isolation of solution points.

(3) In the rotating-rod case where the screen or reflections are not visible then, because there is no contrary 3-D information, the 3-D interpretation will be accepted as most likely.

Consider the very common situation in vision when two patches of surface A and B appear superficially different. Do A and B differ because they have different reflectances (albedos), or is one of the regions a highlight or a shadow on a surface of uniform reflectance? These two interpretations are different, since when B is a shadowed region the implication is that there is an object occluding the direct light of the source, whereas in the highlight case the difference between A and B is due to the specular properties of the surface and there is no cast shadow[4]. (If the darker region around the highlight were to be regarded as shadowed, then 99 per cent of the world would be interpreted as lying in shade!)

As shown in figure 3, let the observer view the surface from above, and let the surface be illuminated with at least two sources of illumination—one producing direct light, as from the sun, while the other source is diffuse, such as that characteristic of the sky and clouds.

We proceed by noting that the only information available to the viewer is the image intensities I_A, I_B from the two regions A and B. For simple Lambertian conditions, these image intensities will be the product of the strength of illumination and the reflectances of the surface material. Let the reflectance common to A and B be R_λ, where the subscript λ indicates R is a function of wavelength, and let S_λ be the incident flux from the direct light of the sun and D_λ the flux arising from the diffuse light from the sky, both of which are also functions of wavelength as indicated by the subscript[5]. If a region is neither highlighted nor shadowed, then

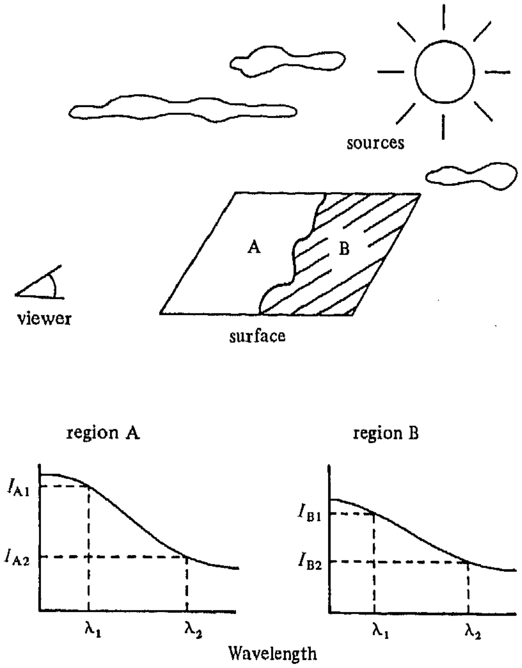

Figure 3. Direct and diffuse light illuminates the surface. Is region A a highlight or is region B in shadow? Possible image intensities over wavelength are illustrated in the lower pair of graphs.

[4] Note that for this analysis we are ignoring other distinctive features of a highlight: (i) the textural aspect of specularity, (ii) its directional component which produces a disparity between the two eyes, and (iii) that highlight edges are convex whereas shadow edges tend to be straight or concave.

[5] A planar surface is assumed; the effect of surface orientation on the source illumination can be considered incorporated into S_λ and D_λ.

the image intensity I will be given by

$$I = (S_\lambda + D_\lambda)R_\lambda . \tag{17a}$$

Equation (17a) thus describes the image intensity resulting from an unshadowed, matte surface.

5.1 *The highlight case*

If region A is the same flat surface as region B, except that it has a highlight, then B remains matte and I_B is defined by equation (17a). On the other hand, equation (17a) will not apply to the highlighted region A, which acts like a partial mirror. Some fraction f_H of the image intensity I_A arising from region A will be due to the mirror-like properties of the surface, whereas the remaining fraction $(1-f_\text{H})$ will be due to the matte component (Evans 1948; Horn 1977).

The matte component of the image intensity of region A, I_A, is found simply by multiplying the fraction $(1-f_\text{H})$ by equation (17a) for a matte surface to obtain $(1-f_\text{H})(S_\lambda+D_\lambda)R_\lambda$. The mirror-like or specular component of I_A is a bit harder to characterize, for the direct source illumination S_λ will be seen only if the surface is oriented such that the source light is reflected into the observer's eye. Otherwise, only a small component of the diffuse illumination will be seen, namely P_λ, which corresponds to the intensity of a particular surface patch that lies in the line of sight as reflected off surface A. The total source illumination of A will thus be the combination of S_λ and P_λ (if different from S_λ), causing the mirror-like components of I_A to be $f_\text{H}(S_\lambda+P_\lambda)$. Therefore, the image intensity we expect from a highlighted region A will be:

$$I_{\text{A}\lambda} = f_\text{H}(S_\lambda+P_\lambda)+(1-f_\text{H})(S_\lambda+D_\lambda)R_\lambda , \tag{17b}$$

where the first term on the RHS is the mirror or specular component and the second term is the matte component of the highlight[6].

5.2 *The shadow case*

If region B is the same surface as A, but B is in shadow, then region B will be illuminated only by the diffuse light D_λ. The effect of shadowing is thus to reduce the illumination from $(S_\lambda+D_\lambda)$ to D_λ. Recognizing that shadows often have penumbrae, we may let f_S be the fraction of the direct illumination that contributes to the shaded region. For shadow, therefore, equation (17a) may be modified as follows:

$$I_{\text{B}\lambda} = (f_\text{S}S_\lambda+D_\lambda)R_\lambda , \tag{18a}$$

which may be rewritten (for future reference) as

$$I_{\text{B}\lambda} = f_\text{S}(S_\lambda+D_\lambda)R_\lambda+(1-f_\text{S})D_\lambda R_\lambda . \tag{18b}$$

For complete shade $f_\text{S} = 0$ and the image intensity $I_{\text{B}\lambda}$ arising from region B is described only by the product of the diffuse light times the reflectance. For no shade $f_\text{S} = 1$; and for penumbrae f_S lies between 0 and 1.

5.3 *Preliminary equation counting*

When we first encounter an intensity difference between two regions in an image, we have no prior knowledge of what caused this difference. In our simple example it could be either a shadow or a highlight condition. To decide, one might be tempted to create a more complex image-intensity equation that combines the effects of all possible factors—including not only shadows and highlights, but also surface

[6] Note that the equation describing the highlight condition is similar to that used for transparency.

orientations, transparency, material changes, or whatever. We would then proceed to find enough image samples that would allow us to solve simultaneously for all the variables in the expression.

Alternatively one might proceed in a more modular fashion, and seek more simple expressions that characterize each physical process separately. Then fewer image samples will be needed to solve for the variables. However, this latter, more modular approach may run into difficulty when several physical processes occur together.

To explore briefly the trade-offs between these two alternatives consider first a simple version of a more complex image-intensity equation, namely a surface that is illuminated to create simultaneously both shadows and highlights. By combining equations (17b) and (18a) we find that the image-intensity equation will have the following form:

$$I_\lambda = f_\mathrm{H}(f_\mathrm{S}S_\lambda + P_\lambda) + (1 - f_\mathrm{H})(f_\mathrm{S}S_\lambda + D_\lambda)R_\lambda\,, \tag{19}$$

where the subscript λ indicates a wavelength dependency and f_H and f_S are respectively the highlight and shadow fractions, which are unknowns, as are S_λ, P_λ, D_λ, and R_λ. Equation (19) thus characterizes a region that is shadowed and/or highlighted to varying degrees, f_S and f_H respectively.

If I_λ, f_H, f_S are now indexed to indicate the spatial region, we can apply the standard equation-counting procedure to determine the minimum number of wavelength and spatial samples needed to solve for the physical variables S_λ, R_λ, D_λ, $f_{i\mathrm{H}}$ and $f_{i\mathrm{S}}$ in terms of the known $I_{i\lambda}$, and then attempt to determine whether the solution for these physical variables implies a shadow or highlight.

Unfortunately, equation counting fails in this case. The Jacobian collapses (is singular). The singularity is due to hidden dependencies in the set of equations of the form (19), such as $D_\lambda R_\lambda$ which always occur together. The possibility of encountering such dependencies increases as more and more physical effects are combined together into one equation.

5.4 *Eliminating dependencies*

If a complex equation containing many variables is to be retained, then the most obvious strategy for eliminating dependencies among equations is to search for other independent relations or constraints. Often this may be difficult, and a more desirable course is to try to reduce the number of unknowns by combining some of the physical variables which are not critical to the interpretation. For example, if the pairs $f_\mathrm{S}S_\lambda$ and $D_\lambda R_\lambda$ occur together everywhere, then we must replace each pair by another single variable, otherwise the Jacobian will always remain singular. (Incidently, such paired variables should be phenomenally indistinguishable anyway, as in the more familiar size-distance trade-off.) Such a reduction of variables need not affect the ability to distinguish a shadow from a highlight.

In effect, by replacing paired variables by a single new variable we are reducing an intractably complex expression to a simpler form that tests the relations of interest more directly. This reduction may lead one toward the more modular approach. For example, note that the shadow equation (18b), the highlight equation (17b) or the combination (19) all have the same basic form [equation (20), below] if the wavelength-dependent portions of the two terms of the RHS of the equations are replaced by the new variables L_λ and M_λ (corresponding to the 'lit' and 'matte' contributions to the image intensity). We shall now show how such simple, more 'abstract' characterizations of the variables can still disambiguate shadows from highlights, using the modular approach together with some simple extra constraints.

5.5 *Solving for the highlights*

To allow us to introduce sufficient constraining relations to solve for a shadow or highlight condition consider the view of two different abutting surfaces as shown in figure 4. The choice of axes is such that the reflectances of each surface, R_1 and R_2, differ in the Y direction with the intensity gradient due to the shadow or highlight decreasing in the X direction, as shown by the stippling. Samples of the image intensities are taken at various positions on each side of the edge, as indicated by the circles labelled A1, A2, B1, B2, etc. Is the brighter region a highlight on a darker (unhighlighted, matte) region, or, alternatively, is the brighter region fully lit and the darker region in shadow? Quite simply, is the change in the image intensity in the X direction due to a shadow or to a highlight? For this two-dimensional case, the highlight equation (17b) will assume the following form:

$$I_{XY\lambda} = f_X L_{Y\lambda} + (1 - f_X) M_{Y\lambda} , \qquad (20)$$

where $I_{XY\lambda}$ is the image intensity corresponding to one of the regions A1, B1, C1 or A2, B2, C2, and $L_{Y\lambda}$ and $M_{Y\lambda}$ are new variables corresponding to $(S_\lambda + P_\lambda)$ or $(S_\lambda + D_\lambda)R_\lambda$, respectively. Note that, since only two wavelength variables L_λ and M_λ are involved along the X axis, these variables need to be indexed by Y only, where Y corresponds to a surface (1 or 2). The remaining variables have an X index which will be replaced by a sample position (A, B or C).

By simple equation counting it can be verified that the minimum number of samples along X or Y and for λ will be respectively either 3, 1, 3 or 3, 3, 1. (Note that Y and λ appear together and hence can be symmetrically indexed.) A further reduction can be obtained by noting that region C1 or C2, etc is always matte, and hence f_C is zero. Thus $I_{CY\lambda} = M_{Y\lambda}$. The minimum for X, Y, λ is then 3, 1, 2 or 3, 2, 1, which corresponds to a set of six equations in six unknowns. The determinant of the Jacobian of either system of equations is still zero, however.

To solve the equations we need to introduce one more constraint or reduce the number of variables. For highlights an additional constraint can be added by noting that the spectral composition of the purely specular component is independent of the underlying reflectance R_1, R_2. Thus along Y, $L_{i\lambda} = L_{j\lambda}$. The minimum X, Y, λ samples are now $X = 2$, $\lambda = 1$, $Y = 2$ (the symmetry between Y and λ has been

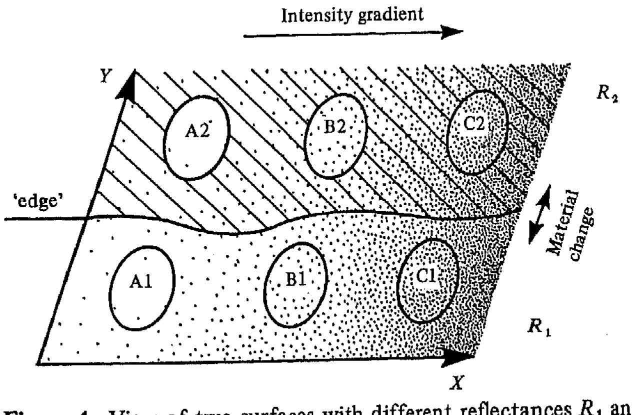

Figure 4. View of two surfaces with different reflectances R_1 and R_2. The image intensity decreases gradually in the X direction, as indicated by the stippling. Is the intensity decrease due to a highlight on the left, or to a shadow on the right? Samples of the image intensities are taken at various positions on each side of the edge, as indicated by the circles labelled, A1, A2, B1, B2, and C1, C2.

removed by the specularity constraint), leading to the following equations:

$$
\begin{aligned}
&I_{\mathrm{B}1} = f_{\mathrm{B}}L_1 + (1-f_{\mathrm{B}})M_1,\\
&I_{\mathrm{B}2} = f_{\mathrm{B}}L_2 + (1-f_{\mathrm{B}})M_2,\\
&I_{\mathrm{C}1} = f_{\mathrm{C}}L_1 + (1-f_{\mathrm{C}})M_1,\\
&I_{\mathrm{C}2} = f_{\mathrm{C}}L_2 + (1-f_{\mathrm{C}})M_2,\\
&f_{\mathrm{C}} = 0, \qquad L_1 = L_2,
\end{aligned}
\tag{21}
$$

where the indexing is for Y only, since there is only a single-wavelength sample.

The Jacobian of the reduced set of the above equations obtained by substituting $L_1 = L_2$ and $f_{\mathrm{C}} = 0$ is:

$$
\begin{bmatrix}
L_1 - M_1 & f_{\mathrm{B}} & (1-f_{\mathrm{B}}) & 0\\
L_1 - M_2 & f_{\mathrm{B}} & 0 & (1-f_{\mathrm{B}})\\
0 & 0 & 1 & 0\\
0 & 0 & 0 & 1
\end{bmatrix}
= f_{\mathrm{B}}(M_2 - M_1),
$$

which is nonsingular provided $M_1 \neq M_2$ and $f_{\mathrm{B}} \neq 0$.

Thus solutions can be obtained for f_{B}, M_1, M_2, and particularly L_1, the specular component of the light reflected off the surface:

$$
L_1 = \frac{I_{\mathrm{B}2}I_{\mathrm{C}1} - I_{\mathrm{B}1}I_{\mathrm{C}2}}{(I_{\mathrm{C}1} - I_{\mathrm{C}2}) - (I_{\mathrm{B}1} - I_{\mathrm{B}2})}, \tag{22a}
$$

$$
1 - f_{\mathrm{B}} = \frac{I_{\mathrm{B}1} - I_{\mathrm{B}2}}{I_{\mathrm{C}1} - I_{\mathrm{C}2}}. \tag{22b}
$$

5.6 *Solving for shadows by combining variables*

Returning to figure 4, we may now reinterpret the regions A1, A2, B1, B2 in terms of a shadow gradient in the X direction. (A penumbra will be needed for this constraint, implying that the minimum number of spatial samples along X is three, although only two will be used as in the highlight case.)

For shadows the equation (19) then has the same form as the first four equations (21), with $L_i = (S_i + D_i)R_i$ and $M_i = D_iR_i$, where S and D are, respectively, the source and diffuse light and R is the reflectance. Since for shadows $L_1 \neq L_2$ (ie there is no spectral component superimposed on A1, A2, or B1, B2), an additional constraining equation must replace this specular constraint. For illustration, we will introduce a 'gray-world' assumption, namely that the average of all surface reflectances (albedos) is spectrally flat or 'gray'. Hence the diffuse light D_i is simply some fraction γ of the source light:

$$
D_i = \gamma S_i, \tag{23a}
$$

and

$$
M_i = \gamma S_i R_i, \tag{23b}
$$

$$
L_i = (1+\gamma)S_i R_i. \tag{23c}
$$

Because S and R appear together in two of the above equations, they cannot be solved for separately, and the Jacobian test will fail when applied to equations (23). To eliminate this dependency define a new variable $S^* = SR$. The shadow equations then become:

$$
\begin{aligned}
I_{B1} &= f_B(1+\gamma)S_1^* + (1-f_B)\gamma S_1^* = (f_B+\gamma)S_1^*,\\
I_{B2} &= f_B(1+\gamma)S_2^* + (1-f_B)\gamma S_2^* = (f_B+\gamma)S_2^*,\\
I_{C1} &= \gamma S_1^*(=M_1),\\
I_{C2} &= \gamma S_2^*(=M_2),
\end{aligned}
\qquad (24)
$$

with the four unknowns f_B, γ, S_1^*, S_2^*.

Unfortunately, the determinant of the Jacobian of this set of equations is still zero, suggesting that dependencies are still present:

$$
\begin{bmatrix}
(f_B+\gamma) & 0 & S_1^* & S_1^* \\
0 & (f_B+\gamma) & S_2^* & S_2^* \\
\gamma & 0 & 0 & S_1^* \\
0 & \gamma & 0 & S_2^*
\end{bmatrix} = 0\,. \qquad (25)
$$

Rather than introduce a new constraint, we shall proceed to determine whether any of the physical variables can be combined to reduce further the number of unknowns. The most obvious choices are ratios or products of the entries in the Jacobian array. These terms are the coefficients of the variables in the original set of equations, and consequently are the factors that would be used to multiply two of the equations to eliminate one variable. (In essence, we are exploring various triangular forms of the matrix of rank one less than the original.) The appropriate ratios are thus those between the rows in the same columns, because it is these factors that will be cross-multiplied to eliminate the variable that is indentified with that column of the Jacobian matrix. Thus the appropriate ratios of the above Jacobian that should be explored first are $(f_B+\gamma)/\gamma$, which appear in columns 1 and 2, and S_1^*/S_2^*, which appear in columns 3 and 4. Inspection of equations (24) shows that the solution for these reduced variables is quite simple:

$$\frac{S_1^*}{S_2^*} = \frac{I_{B1}}{I_{B2}} = \frac{I_{C1}}{I_{C2}} = \frac{S_1R_1}{S_2R_2}\,, \qquad (26a)$$

$$\frac{f_B+\gamma}{\gamma} = \frac{I_{B1}}{I_{C1}} = \frac{I_{B2}}{I_{C2}}\,. \qquad (26b)$$

The extra solution for each paired variable now reveals the dependency between the image intensities that caused the rank reduction of the Jacobian of (24), namely the relation

$$I_{B1}I_{C2} = I_{C1}I_{B2}\,, \qquad (26c)$$

which is common to both (26a) and (26b). If the gray-world condition applies and if C(Y) is a shadow on B(X), then the shadow relation (26c) will be true.

Unfortunately, there are an unlimited number of image-intensity values that will satisfy the 'shadow' relation (26c). How are we to be sure that they all correspond to the shadow condition and not to a reflectance change or even a highlight? To answer this question, we proceed in two stages. First we shall show that the shadow solution (26) never will correspond to a highlight, and hence shadows and highlights are at least disambiguated because their solutions are distinct. Then, we shall illustrate how the probability of other confounding spectral relations such as different materials can be set arbitrarily low by independent corroboration of the original solution.

6 Distinctness of shadow and highlight solutions (exclusion of competing interpretations)

Our basic procedure to prove distinctness of the shadow S and highlight H solutions will be to show that there is at least one relation between the four available image intensities (I_{B1}, I_{B2}, I_{C1}, I_{C2}) that has different values for the shadow and highlight conditions. These values will always be different (if the constraints are valid) because the relation corresponds to two different physical variables (one for shadow, the other for highlights) that have nonoverlapping values.

To proceed we ask first what highlight conditions satisfy the shadow solution (26). (Subsequently we shall examine the opposite case—asking which shadow conditions will 'look like' highlights.)[7] We thus assume relation (26) holds and solve for one of the highlight conditions. Consider equation (22a) that specifies the magnitude of the specular components of the highlight. Note that the numerator is identical to the shadow equation (26) if the LHS of equation (26) is subtracted from the RHS. In this case, however, the numerator of (22a) will be zero. Hence the shadow condition requires that $L_{\text{specular}} = 0$ and consequently there can be no highlight interpretation. Thus, given that the shadow condition (26) holds, there will be no highlight interpretation.

To check for the reverse case, namely under which conditions the image-intensity relations for the highlight condition will also yield a shadow interpretation, we may examine the second highlight equation (22b). In particular, we wish to solve for the physical interpretation of the intensity relations of (22b) given a shadow condition. This can be accomplished simply by substituting equations (24) into the RHS of expression (22b). We find that, given the shadow conditions, then

$$\frac{I_{B1}-I_{B2}}{I_{C1}-I_{C2}} = \frac{f_B+\gamma}{\gamma} = \frac{f_B}{\gamma}+1. \tag{27}$$

Figure 5 now shows the possible values of the image-intensity ratio given by the LHS of (27) for shadows and the RHS of (22b) for highlights.

We note that both f (the fraction of specularity or shadow) and γ (the fraction of direct light), range between 0 and 1. Hence for highlights $1-f$ must lie between 0 and 1, whereas for shadows $1+f/\gamma$ will be greater than or equal to 1. The only common condition is when $f = 0$, which corresponds to a homogeneous matte area. Thus highlights and shadows will never be confused from the image intensities (provided the gray-world assumption is correct), if the calculation given by the LHS of (27) is made. It is of some interest that this operation on image intensities is equivalent to examining the output of the double-opponent color cell found in most biological color vision systems (see Rubin and Richards 1981).

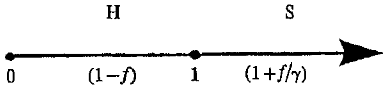

Figure 5. Solution space for shadow S and highlight H conditions.

6.1 *Corroboration*

Although the highlight H and shadow S solutions are unique and distinct, it is still possible that other properties of surfaces, such as pigment density changes or changes in reflectances, could satisfy equations (22) or (25) and be misinterpreted as either a highlight H or shadow S. Thus a shadow or highlight interpretation should not yet be given to the solutions H and S. To exclude *all* other possibilities is difficult (see Rubin and Richards 1981, however). Nevertheless, the odds for an incorrect H or S

[7] For another example treatment, see Ullman's (1979) analysis of false targets for his structure-from-motion theorems.

interpretation can be reduced by applying an independent test for the validity of the shadow or highlight equations. We call such a procedure 'corroboration'.

One simple independent corroborative test is to note whether the equation-counting procedure suggested more than one minimal condition for solution. In particular, we noted in section 5.5 that equation (20) had a symmetry in wavelength (λ) and space (Y). We chose as a starting point one spectral sample and two samples in the Y dimension. An independent test would therefore be to use two spectral samples rather than one, and only one sample in the Y dimension. This case corresponds to examining the gradients of a highlight, or the penumbra of a shadow.

A second and more common type of corroborating procedure is simply to take another set of measurements independent of the first, and determine whether the solutions for the physical constants remain the same or not. If they do not, then the interpretation must be rejected. If they are confirmed, then the odds on a misinterpretation are reduced. Ideally, the corroboration should be based upon measurements taken from a physical dimension different from that used in the original solution. In any case, since we are corroborating the value of a physical parameter, the corroborating measurements must not be confounded with the dimensions of that physical parameter. In this respect the relation (27) that tests for the highlight or shadow condition is most satisfactory, for the values f_{B} and γ are dimensionless and are not functions of wavelength, for example. For the shadow condition we thus can take a third spectral sample I_{B3}, I_{C3} and substitute these image intensities for I_{B2}, I_{C2}. Since the physical constant $(f_{\mathrm{B}}+\gamma)/\gamma$ of equation (26b) is not a function of wavelength, its value should remain unchanged if the image-intensity changes are indeed due to a shadow. In effect, we are confirming that the S solution point remains fixed along the solution ray illustrated in figure 5. If it does, then the shadow (or highlight) interpretation is reaffirmed and the chance of misinterpretation is unlikely provided that the competing interpretations are not processes that behave like shadows. Consequently at least three wavelength samples are required before a reliable shadow interpretation can be made.

In the case of recovering structure from motion—our earlier example—the corroboration of the axis of rotation could entail adding additional *frames* or snapshots to see if the same axis and rod length are recovered. Clearly this procedure is not entirely independent because the strategy for solution remains the same and some possible confounding interpretations may not be excluded (eg the correct interpretation that the points are on a TV monitor in 2-D).

A more independent corroborative test would be to use stereopsis, for this computation of the depth relations between the feature points is quite different from the structure-from-motion analysis. The ideal corroborative procedure should thus use an entirely different computational analysis, which is based upon relations that have quite different failure conditions[8].

7 Summary

Although the equation-counting procedure has been used in the past to give some insight into the complexity required to solve problems in many nonlinear variables (eg Leith et al 1981), researchers in perception have often neglected to recognize that certain other conditions must be fulfilled before a meaningful solution can be guaranteed (Meiri 1980). These conditions are summarized in the flow diagram of figure 6. They include the Jacobian test for the independence of the system of

[8] For biological systems we probably should view 'corroboration' as an early step in the perceptual process—perhaps at the level of Marr's 2-1/2D sketch (Marr 1976, 1982)—that acts on the output of modules analyzing information derived from motion, disparity, color, texture, etc, as well as nonvisual information, such as tactile roughness, shape, or even in some cases acoustic information.

equations, uniqueness of solution, exclusions of competing interpretations, and two kinds of corroboration. If these conditions can be met, then the equation-counting procedure provides a powerful theoretical tool for understanding how, in principle, biological systems can make reliable interpretations and assertions from the greatly impoverished sensory data available to them.

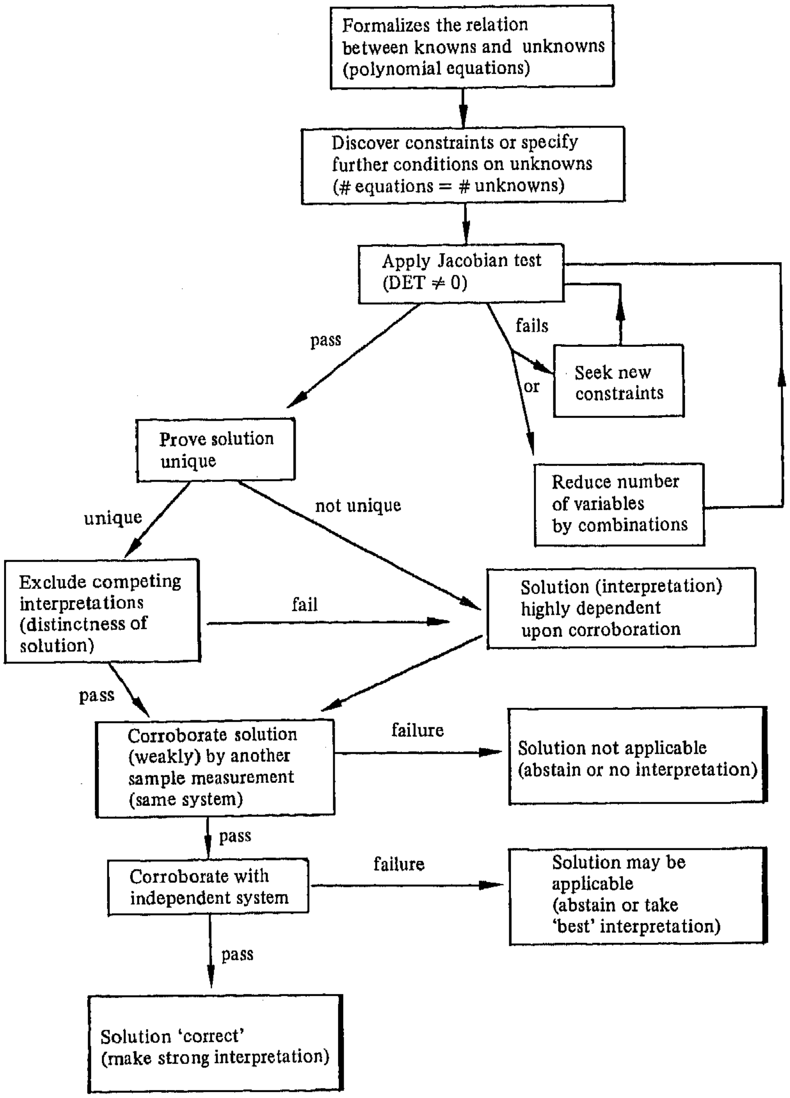
Formalizes the relation between knowns and unknowns (polynomial equations)
Discover constraints or specify further conditions on unknowns (# equations = # unknowns)
Apply Jacobian test (DET ≠ 0)
pass
fails
or
Seek new constraints
Reduce number of variables by combinations
Prove solution unique
unique
not unique
Exclude competing interpretations (distinctness of solution)
fail
Solution (interpretation) highly dependent upon corroboration
pass
Corroborate solution (weakly) by another sample measurement (same system)
failure
Solution not applicable (abstain or no interpretation)
pass
Corroborate with independent system
failure
Solution may be applicable (abstain or take 'best' interpretation)
pass
Solution 'correct' (make strong interpretation)

Figure 6. Outline of steps in equation-counting procedure.

Acknowledgements. This report describes research in Natural Computation at the Artificial Intelligence Laboratory and Psychology Department of the Massachusetts Institute of Technology. Support for this research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-75-C-0643 and by NSF and AFOSR under grant 7923110-MCS. D D Hoffman was also supported by a Hughes Aircraft Company graduate fellowship and J M Rubin was supported by an NSF Fellowship. We greatly appreciate the help of Professor M Artin and T Shifrin, who pointed us to the Bezout Theorem and the Jacobian test, and also the insightful comments of Professor S Ullman and A Pentland. Technical assistance in the preparation of the manuscript was provided by C J Papineau.

References

Evans R M, 1948 *An Introduction to Color* (New York: John Wiley)

Guillemin V, Pollack A, 1974 *Differential Topology* (Englewood Cliffs, NJ: Prentice Hall)

Hoffman D D, Flinchbaugh B E, 1982 "The interpretation of biological motion" *Biological Cybernetics* 42 195-204 (also available as MIT AI Memo 608)

Horn B, 1977 "Understanding image intensities" *Artificial Intelligence* 8 201-231

Huffman D, 1971 "Impossible objects as nonsense sentences" in *Machine Intelligence* 6 Eds B Meltzer, D Michie (Edinburgh: Edinburgh University Press)

Johansson G, 1975 "Visual motion perception" *Scientific American* 232(6) 76-89

Kendig K, 1977 *Elementary Algebraic Geometry* (Charlottesville, VA: University of Virginia Press)

Leith E N, Chen H, Cheng Y S, 1981 "Diffraction-limited imaging through a phase-distorting medium" *Optics Letters* 6 4-6

Marr D C, 1976 "Early processing of visual information" *Philosophical Transactions of the Royal Society of London Series B* 275 483-524

Marr D C, 1982 *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (San Francisco, CA: W H Freeman)

Meiri A Z, 1980 "On monocular perception of 3-D moving objects" *IEEE Transactions on Pattern Analyses & Machine Intelligence* PAMI-2 582-583

Milnor J W, 1978 *Topology from the Differential Viewpoint* (Charlottesville, VA: University of Virginia Press)

Pentland A, 1980 *Finding the Illuminant Direction* Massachusetts Institute of Technology, Cambridge, MA, MIT AI Memo 584

Rubin J M, Richards W A, 1981 *Color Vision: When are Changes Material?* Massachusetts Institute of Technology, Cambridge, MA, MIT AI Memo 631

Thomas G B, 1951 *Calculus and Analytic Geometry* (Reading, MA: Addison-Wesley)

Ullman S, 1979 *The Interpretation of Visual Motion* (Cambridge, MA: MIT Press)

Van der Waerden B L, 1940 *Modern Algebra* volume II (New York: F Ungar)

Waltz D, 1975 "Understanding line drawings of scenes with shadows" in *The Psychology of Computer Vision* Ed. P Winston (New York: McGraw-Hill)

Witkin A, 1980 *Shape from Contour* Massachusetts Institute of Technology, Cambridge, MA, MIT AI Technical Report 589

APPENDIX I: Redundancy

Unfortunately, owing to measurement and sampling errors, real-world data are not precise. The hardware performing the calculations may also be quite noisy, as is the case for many neural networks. Without exact data and calculations solution vectors will not be completely isolated, but rather are more properly represented as a probability distribution about the exact solution point. To reduce the likelihood of misinterpretation, several overconstraining equations are often helpful. (By 'overconstraining' we mean the inclusion of equations in addition to those needed to obtain a unique solution.) Their value will depend in part upon how many variables (unknowns) are included in the solution point. Intuitively, the more unknowns there are, the greater the potential noise and the less the contribution of any one overconstraining equation will be. To capture this property, we suggest the following measure of the redundancy of a system containing overconstraining equations:

$$\text{Redundancy} = 1 - \left(1 - \frac{1}{U}\right)^{C}, \tag{A1}$$

where C is the number of independent combinations of the equations and U is the number of unknowns. As U increases, this measure decreases to zero. The effect of the additional overconstraining equations, on the other hand, is to reduce the deleterious effect of increasing U in a manner analogous to probability summation, yet the redundancy measure will never exceed 1 (the ideal). The redundancy measure has the practical value of providing an estimate of how many extra equations (or data samples) are needed to isolate a solution point to a certain probability, given known measurement signal-to-noise ratios.

APPENDIX II: Sard's Theorem for nonpolynomial functions

In many cases, the equations relating the unknown variables will not be polynomial and Bezout's Theorem will not apply. These exceptions include such common functions as exponential, logarithmic, or trigonometric functions. Sometimes a change of variables can be made to recast the nonpolynomial relations in polynomial form. If this is done, then care must be taken to restrict the range over which the polynomial form applies.

More generally, if a function is smooth on a manifold, then Sard's Theorem can be used (Guillemin and Pollack 1974; Milnor 1978). Suppose that the following system of independent equations holds:

$$f_1(x_1, ..., x_k) = p_1,$$
$$\vdots$$
$$f_n(x_1, ..., x_k) = p_n.$$

This system can then be represented more generally as a mapping from R^k to R^n:

$$F: R^k \rightarrow R^n,$$

or

$$F(x_1, ..., x_k) = \{f_1(x_1, ..., x_k), ..., f_n(x_1, ..., x_k)\}.$$

By Sard's Theorem, we know that if F is a smooth mapping and if F is invertible for the values p, then the dimension of $F^{-1}(p)$ is $(k-n)$. Since when $k = n$ the dimension of $F^{-1}(p)$ is zero, there can be at most a countable number of (isolated) solutions.

Some care must be taken in assuming that Sard's Theorem guarantees a finite number of solutions for any system of equations involving differentiable functions. It does not. For example, consider the simple periodic function $\sin x$. Such a function is uniquely invertible only over a specified range. Polynomial functions are thus a 'safer' class of functions to use for equation counting, for their appropriate range is usually more obvious.