Appendix B

Statistical Methods

Machine vision could be called statistical geometry, since vision involves estimating geometrical information from image data. This book uses some basic material from statistics, such as the normal distribution and linear regression, which are reproduced in this appendix. The book also uses some new ideas in statistics for formulating measurement algorithms that are robust to unmodeled errors.

B.1 Measurement Errors

There are three types of performance parameters for a sensor or measurement procedure. Resolution or precision is the smallest change in the value that a sensor can report. Repeatability is the variation in repeated measurements of the same quantity. Accuracy is the variation in measurements of a known true value. It is easy to remember the relationship between accuracy and repeatability:

$$\text{Accuracy} = \text{Repeatability} + \text{Calibration}. \tag{B.1}$$

There is a similar relationship between the components of error:

$$\text{Error} = \text{Variance} + \text{Bias}. \tag{B.2}$$

The variance is the error in repeatability for a measurement; the bias is the systematic error due to lack of calibration. For example, suppose that you are measuring the length of a hallway with a yardstick, but instead of a

yardstick you accidentally pick up a meter stick. Your measurements would still have the same repeatability, assuming that you were just as careful in laying the measuring stick end to end as you measured the length of the hall. But there would be a systematic bias, a constant proportional to the true length of the hall, due to the incorrect length of the measuring stick. Bias can be removed through careful calibration, but variance (or repeatability) is a characteristic limitation of the measurement method.

The histogram is a useful tool for seeing the distribution, including both variance and bias, of a measurement. The range of measurements on the real line is partitioned into a finite number of intervals called buckets. A one-dimensional integer array of length equal to the number of buckets is used to count the number of occurrences of measurements that fall in the intervals. A plot of the histogram shows the distribution of measurements. The width of the plot is an indication of variance, and the difference between the location of the center of the distribution and the true measurement is an indication of bias.

A measurement y_i of some quantity x can be corrupted by additive error:

$$y_i = x + e_i, \tag{B.3}$$

where e_i is the error in the measurement. If the error were known or if the error could be removed through calibration, then each measurement would be an accurate estimate of the unknown quantity, and no further processing would be necessary. However, the repeatability of a measurement procedure is not perfect, and each measurement will include error drawn from some distribution leading to a distribution of measurements somehow related to the unknown parameter. Statistics is the science of measurement procedures and includes methods for estimating unknown parameters given various assumptions about the characteristics of the errors.

The average of a set of n measurements is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{B.4}$$

The median is computed by sorting the measurements and choosing the middle element (or averaging the two middle elements if the number of measurements is even). The mode is the location of the peak in the distribution of measurements. The average and median are methods for estimating an

unknown parameter from measurements corrupted by additive errors from a symmetric distribution. For example, pixels can be averaged over local neighborhoods, or the pixel values obtained at the same location in a sequence of images can be averaged to reduce the noise in the measurements of gray value.

Several different measurements of the amount of error can be computed. Mean absolute error (MAE) is $\pm\delta_x$, where

$$\delta_x = \frac{1}{n}\sum |x_i - \mu_x| \tag{B.5}$$

and μ_x is the average or median. Root mean square (RMS) error is $\pm\sigma_x$, where

$$\sigma_x^2 = \frac{1}{(n-1)}\sum (x_i - \mu_x)^2 \tag{B.6}$$

and μ_x is the average. Maximum error is $\pm\epsilon_x$, where

$$\epsilon_x = \max_i |x_i - \mu_x| \tag{B.7}$$

and μ_x is the average or median. Note that $\delta \leq \sigma \leq \epsilon$.

B.2 Error Distributions

The binomial distribution models measurement processes with a finite number of outcomes, called events. For example, a fair coin that is flipped once will show heads with probability 1/2 and tails with probability 1/2. This set of outcomes is modeled by the polynomial

$$P(x) = 0.5 + 0.5x, \tag{B.8}$$

where the coefficient for each power of x is the probability that heads will occur that many times. If the coin is tossed n times, then the probabilities of various numbers of heads can be determined by expanding the polynomial:

$$P(x; n) = (0.5 + 0.5x)^n. \tag{B.9}$$

The probability that heads will occur i times is the coefficient of x^i in the expanded polynomial. In general, a single measurement with k outcomes can be modeled by a polynomial of order k,

$$P(x) = p_0 + p_1 x + p_2 x^2 + \cdots + p_{k-1} x^{k-1}, \tag{B.10}$$

where p_i is the probability of outcome i and

$$\sum_{i=1}^{k} p_i = 1. \tag{B.11}$$

The cumulative results of various combinations of outcomes after a sequence of n measurements are modeled by raising the polynomial for one measurement to power n,

$$P(x;n) = \left(p_0 + p_1 x + p_2 x^2 + \cdots + p_{k-1} x^{k-1}\right)^n, \tag{B.12}$$

expanding the polynomial, and calculating each coefficient.

The uniform distribution is used to model measurements that are equally likely. For example, if a point is located anywhere with equal probability in the rectangular region defined by points (x_1, y_1) and (x_2, y_2) at the corners, then the probability that the point is at location (x, y) is

$$P(x,y) = \begin{cases} 1/A & \text{if } (x,y) \text{ is in the region} \\ 0 & \text{otherwise} \end{cases} \tag{B.13}$$

where A is the area of the region.

The normal distribution is a useful approximation to the errors in many measurement processes. The normal distribution $N(x;\mu,\sigma^2)$ is

$$N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{B.14}$$

where μ is the location of the center of the distribution and σ^2 is the variance. Since bias is usually eliminated by proper calibration, the normal distribution for errors is usually zero mean.

Some measurement processes occasionally produce gross errors, called outliers, in addition to normally distributed errors. Such an error process can be modeled as the mixture of a normal distribution and some unknown, broad-tailed distribution:

$$(1-\nu)N(x;\mu_1,\sigma_1^2) + \nu B(x;\mu_2,\sigma_2^2), \tag{B.15}$$

where ν represents the odds that an outlier will occur. There is a probability of ν that the measurement will be contaminated with error from the outlier

distribution and a probability of $\nu - 1$ that the measurement will be subject to normally distributed errors. Typically, the bias from both error processes is eliminated through proper calibration, and so both error distributions are zero mean:

$$(1 - \nu)N(x; 0, \sigma_1^2) + \nu B(x; 0, \sigma_2^2). \tag{B.16}$$

Since outliers are extreme errors, $\sigma_2 \gg \sigma_1$. As an example, the Cauchy distribution

$$C(x; a, b) = \frac{1}{\pi} \frac{a}{a^2 + (x - b)^2} \tag{B.17}$$

has such broad tails that the variance is infinite.

B.3 Linear Regression

Given n data points and a model with m parameters $a_1, a_2, \ldots, a_m$, the least-squares error in the fit of the model to the data is

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{y_i - y(x_i; a_1, a_2, \ldots, a_m)}{\sigma_i} \right)^2. \tag{B.18}$$

This is weighted least-squares regression, where the weights σ_i are the errors (standard deviations) in the measurements so that less noisy measurements are given more weight.

Often, the error is the same for each measurement, or the individual measurement errors are unknown and assumed to be identical, in which case the least-squares regression problem is to determine the parameters $a_1, a_2, \ldots, a_m$ that minimize

$$\chi^2 = \sum_{i=1}^{n} \left(y_i - y(x_i; a_1, a_2, \ldots, a_m) \right)^2. \tag{B.19}$$

If the model is linear in the parameters, then the problem is linear regression. A model is linear if it can be represented as a linear combination of basis functions:

$$y(x; a_1, a_2, \ldots, a_m) = a_1 \phi_1(x) + a_2 \phi_2(x) + \cdots + a_m \phi_m(x), \tag{B.20}$$

where the functions $\phi_i(x)$ do not depend on the model parameters. The coefficients $a_1, a_2, a_3, \ldots, a_n$ of the linear combination are the model parameters to be determined through regression. For example, the line

$$y = ax + b \tag{B.21}$$

is a linear model with parameters a and b.

A linear least-squares regression problem can be reliably solved using standard numerical routines for singular value decomposition [84, pp. 534–539], which also provides a measure of the error in the fitted parameters. Consider fitting a linear model

$$z = a_1 + a_2 x + a_3 y \tag{B.22}$$

to a set of data points $\{(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n)\}$. Each data point leads to a constraint

$$z_i \approx a_1 + a_2 x_i + a_3 y_i, \tag{B.23}$$

and the set of data points leads to a set of constraints that can be written in matrix form:

$$\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ & \vdots & \\ 1 & x_n & y_n \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}. \tag{B.24}$$

These equations are usually written in the more compact form

$$AX = B. \tag{B.25}$$

The A matrix is not square and cannot be inverted directly. One technique is to multiply both sides of Equation B.25 by A^T to form the *normal equations* and solve the normal equations using standard techniques for solving systems of linear equations such as LU decomposition. A better technique is to use singular value decomposition. There are several advantages to using singular value decomposition: (1) it is not necessary to premultiply Equation B.25 to form the normal equations, (2) singular value decomposition handles ill-conditioned systems of equations, and (3) the singular values provided as a by-product of SVD indicate redundancies (unnecessary terms) in the model.

Since singular value decomposition is used frequently for the algorithms in this book and *Numerical Recipes* [197] is an excellent source for good

numerical algorithms, the method for solving a linear regression problem using the singular value decomposition routine in *Numerical Recipes* will be presented in detail. (Note that some elements of C programming are described in Appendix C.) The interface for the numerical routine for singular value decomposition is

```
void svdcmp (a, n, m, w, v)
float **a, *w, **v;
int n, m;
```

The routine takes an array `a` with `n` rows and `m` columns and replaces it with the singular value decomposition

$$A = UWV^T. \tag{B.26}$$

The array `a` is replaced by U in the singular value decomposition, the array `w` receives the singular values in the diagonal matrix W, and the array `v` receives the V matrix. Note that `n` is the number of observations (measurements), while `m` is the number of parameters in the linear model. It is easy to fill the entries in the array `a` according to Equation B.24 and call **svdcmp** to obtain the singular value decomposition. After obtaining the singular value decomposition, use the routine **svbksb** to solve for the model parameters:

```
void svbksb (u, w, v, n, m, b, x)
float **u, *w, **v, *b, *x;
int n, m;
```

The `u`, `w`, and `v` arrays are the `a`, `w`, and `v` arrays computed by **svdcmp**. The dimensions `n` and `m` are the same as for **svdcmp**. The `b` array is the right-hand side of Equation B.24, and the array `x` is the parameter vector (solution). In other words, the combination of **svdcmp** and **svbksb** solve the nonsquare system of linear equations presented as Equation B.24. The code fragment for invoking the routines is

```
float wmin, wmax;
svdcmp (a, n, m, w, v);
wmax = 0.0;
for (j = 1; j <= m; j++)
  if (w[j] > wmax) wmax = w[j];
```

```
wmin = wmax * 1.0e-6;
for (j = 1; j <= m; j++)
  if (w[j] < wmin) w[j] = 0.0;
svbksb (a, w, v, n, m, b, x);
```

Small singular values indicate problems in the regression model. The code provided above sets small singular values to zero, which is one safe way to handle the problem. The constant `1.0e-6` is a typical value but may be different for some applications. If there are small singular values, it is important to carefully analyze the model and determine why the small singular values occur. There may be terms in the model that are unnecessary.

Singular value decomposition is a very reliable algorithm but can fail to produce a good regression estimate for several reasons:

- The wrong model may be used in formulating the regression problem.
- The measurement errors σ_i may be too large.
- The measurement errors may not be from a normal distribution. For example, the errors could be from a broad-tailed distribution.

The probability distribution for χ^2 when the minimum value of the regression norm is obtained is the chi-square distribution with $\nu = n - m$ degrees of freedom. The probability that chi-square should exceed χ^2 by chance is

$$Q = \frac{1}{\Gamma(\frac{\nu}{2})} \int_{\chi^2}^{\infty} e^{-t} t^{\frac{\nu}{2}-1} \, dt. \tag{B.27}$$

The regression error χ^2 should be calculated as part of the regression procedure. The integral can be calculated using numerical methods, but values of Q are provided in statistical tables. If $Q > 0.001$, then the regression fit should probably be rejected. This provides an objective way to evaluate the model fitting procedure. In practice, the value for Q is selected based on knowledge of the application. The value for ν is determined by the number of data points and the order of the model. The corresponding value for χ^2 can be found in statistical tables. If the measured value for χ^2 obtained during regression exceeds the tabulated value, then the algorithm has not succeeded, and the results should be discarded. If excessive values for χ^2 are encountered frequently, then the model is probably wrong or linear regression is not the right approach, and perhaps robust regression should be used instead.

B.4 Nonlinear Regression

If the model is nonlinear in its parameters, the least-squares regression problem is to minimize χ^2 for the nonlinear model. Since the formula for the model is known, both the gradient and Hessian can be calculated. This allows the Levenberg-Marquardt method to be used for solving nonlinear regression problems [197, pp. 540–547].

Seber and Wild [217] is an excellent text on nonlinear regression. In some cases, the arguments to the Levenberg-Marquardt routine in *Numerical Recipes* may not match the intended application, but *Numerical Recipes* provides routines for the Newton-Raphson method, which is discussed in practical texts on nonlinear regression [23].

Further Reading

There are many excellent textbooks on probability and statistics. Any book that describes experimental methods in science and engineering or statistical methods in the social sciences would be sufficient. For example, Beck and Arnold [24] cover linear and nonlinear methods for regression with emphasis on applications in engineering and science. Box, Hunter, and Hunter [43] have written a classic text on experimental methods. Drake [69] has written a basic introduction to probability and statistics, while Papoulis [192] provides more comprehensive coverage. Gaussian error models are used in communications theory [257], which has influenced machine vision and pattern recognition. Vanmarcke [244] and Cressie [63] present probability and statistical methods for spatial data which may be useful as further readings in machine vision. Robust regression methods are summarized in the article by Efron and Tibshirani [72]. Finally, *Numerical Recipes in C* is an excellent source for statistics algorithms [197] and explains regression particularly well.