Review: Statistics, Estimation and Decisions

The following is a (*very*) brief review of statistics, described from a Bayesian perspective, in the context of both experimental data analysis and modeling. I emphasize brevity and geometric intuition, at the expense of rigor and detail. I assume the reader is familiar with linear algebra (as reviewed in my handout *Geometric Review of Linear Algebra*), and least squares estimation (as reviewed in my handout *Least Squares Optimization*).

1 Probability Basics

Suppose we perform an experiment, measuring the intensity of a constant light source using a photometer. Each time we make this measurement, we get a slightly different answer. Although the answers will presumably cluster around the "correct" value, there is no way for us avoid the variability in the measurements. The field of **probability** provides an abstract language for describing the uncertainty of these measurements. The field of **statistics** tells us how to take a finite set of measurements and infer something about the world.

The primary entities of probability theory are **random variables** and their associated **probability distributions**.

The following ideas need to be explained:

- distributions: discrete/continuous, Examples: binary (coin), uniform, Gaussian
- cumulatives, transformations on densities, histogram-equalization
- multi-dimensional random variables: joint densities, marginals, conditionals, Bayes Rule
- statistical independence
- Expectations, moments
- Multi-dimensional Gaussians: parameterization, marginals/conditionals
- Sums of independent random variables implies convolution of distributions
- Central limit theorem

2 Statistical Estimation

The previous section was about abstract mathematical descriptions of probability. Now we imagine ourselves in a more practical context, in which we've made some observations (i.e.,

- Created: Spring 2004. Revised: June 2004.

- Author: Eero Simoncelli,
  Center for Neural Science, and Courant Institute of Mathematical Sciences,
  New York University
- Send corrections or comments to eero.simoncelli@nyu.edu

experimental measurements), from which we want to estimate some quantity. In general, each time we make a measurement, it comes out differently. This unpredictability might arise either from aspects of the environment that are beyond our control (eg., stray electromagnetic radiation), or from unobservable fluctuations within the system itself (eg., flow of individual ions through channels in a membrane), or from uncertainties introduced by the measurement process. We refer to the quantity we're trying to measure as the "signal", and all these sources of uncertainty as "noise". We describe the noise using random variables.

Consider an example in which we wish to measure the brightness of a constant light source. Our measurement is corrupted by the quantal nature of light, by disturbances in the medium (air) through which the light must propagate, and by inacuraccies in our measurement device. To make the problem simpler, it is often assumed that such measurement uncertainties are combined additively along with the "true" value to yield the measurement:

$$\mathbb{P}(m|b) = b + n \tag{1}$$

Here n is a random variable that represents the combination of three sources of uncertainty mentioned above. The right side of equation (1) is known as the "likelihood" function: it tells us the likelihood of our measurements given a particular value of b. Now, the problem of estimation is to *invert* this equation: We want estimate of b, given a finite set of measurements $\{m_k\}$. We'll notate our estimate as $\hat{b}(\{m_k\})$, with the "hat" indicating that this is not the true b, and the parentheses indicating that it is a function of the data. More compactly, we can also bundle the m_k's into a vectxor $\vec{m}$.

Before discussing specific estimators, it should be intuitively obvious that we'll want to minimize the error in our estimates – that is, the difference between the estimate and the true value. We decompose this error into two distinct pieces:

Bias : This is the average error in the estimator:

$$B(b) = \mathbb{E}_n \left[\hat{b}(\vec{m}) - b\right]$$

An estimator that is (on average) equal to the true value is called **unbiased**.

Variance : This is simply the variance of the error:

$$V(b) = \mathbb{E}_n \left[(\hat{b}(\vec{m}) - b - B(b))^2\right]$$

Notice that the mean squared error is just the sum of the squared bias and the variance.

Now how do we decide on an estimator? As with most such questions, the answer is "it depends on the problem". But it is worth knowing about three particular estimators that are most commonly used, and which are built upon each other. First, suppose that all we know about our problem is the likelihood function of equation (1). In this case, the simplest choice is to choose the value of b that makes the measurements most likely:

$$\hat{b}_{\mathrm{ML}}(\vec{m}) = \arg\max_b \mathbb{P}(\vec{m}|b)$$

This is known as the **Maximum Likelihood** estimator (MLE).

Take the example of light measurement, and assume that n is zero-mean, Gaussian distributed, with variance σ^2. Assume we make N measurements, and that these are statistically independent. Then the likelihood function is a product of the individual likelihoods:

$$\begin{aligned}
\mathbb{P}(\vec{m}|b) &= \prod_k \mathbb{P}(m_k|b) \\
&= \prod_k \exp[-(m_k - b)^2/2\sigma^2]
\end{aligned}$$

To compute the estimate, we could maximize this expression, but it's simpler to maximize the log:

$$\begin{aligned}
\hat{b}(\vec{m}) &= \arg\max_b \log \mathbb{P}(\vec{m}|b) \\
&= -\arg\max_b \sum_k (m_k - b)^2/2\sigma^2
\end{aligned}$$

Taking the derivative of the righthand expression with respect to b and setting equal to zero gives:

$$\sum_k 2(m_k - \hat{b}(\vec{m})) = 0$$

or

$$\hat{b}(\vec{m}) = \frac{1}{N}\sum_k m_k$$

After all that, the answer is quite simple: take the average of the measurements! [Now, verify that this is unbiased, and compute the variance.]

Now we consider a more sophisticated estimator. Suppose we had some knowledge of the values that b could assume. For example, we might know that it must lie within a particular range. Or perhaps some values, while possible, are extremely unliky to occur in the real world. This kind of knowledge may be represented with a probability distribution on b, known as the **prior** distribution: $\mathbb{P}(b)$.

Given this knowledge, we can use Bayes' rule to turn the likelihood into the inverse conditional probability, and we can then maximize that:

$$\begin{aligned}
\hat{b}_{\mathrm{MAP}}(\vec{m}) &= \arg\max_b \mathbb{P}(b|\vec{m}) \\
&= \arg\max_b \mathbb{P}(\vec{m}|b)\mathbb{P}(b)/\mathbb{P}(\vec{m}) \\
&= \arg\max_b \mathbb{P}(\vec{m}|b)\mathbb{P}(b)
\end{aligned}$$

In the last step, we dropped the denominator from the expression because it does not depend on b, and thus has no influence on the maximum. This estimator is known as the **maximum aposteriori** (MAP) estimator.

Finally, we might want to augment the problem by including some sort of cost function (also called a "loss" function) that describes how much penalty is incurred by making each particular error. In general, this is a function of both the true value and the estimated value: $L(b, \hat{b})$. A **Bayesian** estimator attempts to minimize the average (expected) loss:

$$\begin{aligned}\hat{b}_{\text{Bayes}}(\vec{m}) &= \arg\min_{\hat{b}} \mathbb{E}_b[L(b,\hat{b})|\vec{m}] \\ &= \arg\min_{\hat{b}} \int_b L(b,\hat{b})\mathbb{P}(b|\vec{m})\end{aligned}$$

[Ex: Gaussian linear case. Note that as the number of measurements increases, the estimate gets closer and closer to the ML estimate.].

A special case of this estimator is the **Bayes Least Squares** (BLS) estimator, in which the loss function is just squared error:

$$\begin{aligned}\hat{b}_{\text{BLS}}(\vec{m}) &= \arg\min_{\hat{b}} \int_b (b-\hat{b})^2\mathbb{P}(b|\vec{m}) \\ &= \int_b b\mathbb{P}(b|\vec{m})q = \mathbb{E}_b[b|\vec{m}]\end{aligned}$$

where the second line is achieved by differentiating the first line, setting the result equal to zero, and solving for $\hat{b}$. That is, the BLS estimator is simply the conditional mean of the parameter given the data!

Note that in the Gaussian linear case studied above, the BLS is identical to the MAP estimator, since the peak of a Gaussian distribution is the same as the mean. But for non-Gaussian posterior densities, the BLS and MAP are often different.

Using empirical estimates for variance

Often, we don't actually have a good model for the noise source, and thus can't get an expression for the variability in our estimates. In such cases, we can look at the variability empirically, by repeatedly computing the estimate on a "fake" data set that is obtained by resampling from the true data...

Statistical Decision Theory

The previous section described the problem of estimating the value of some unknown quantity. Another commonly encountered problem is that of making a decision based on a set of uncertain measurements. The problem is closely related to the estimation problem (and can often be posed as a subcase).

ML solution

MAP / BAYES

In perceptual psychology, a restricted form of decision theory, known as **Signal Detection Theory** has been used to describe the process by which observers detect stimuli in experiments....

d'

ROC curves