Maximum Likelihood Estimation

S. Purcell.

Contents and Keywords

Further Reading

Edwards AWF (1972) Likelihood. Cambridge University Press.

Sham PC (1998) Statistics in Human Genetics. Arnold

Site created by S.Purcell, last updated 27.11.2000

Maximum Likelihood Estimation (MLE)

Introduction

This site provides a brief introduction to maximum likelihood estimation: the details are not essential to learn, but it is useful to have a grasp of some of the underlying principles.

Probability

The concept of likelihood, introduced by Sir R. A. Fisher, is closely related to the more common concept of *probability*. We speak about the probability of observing events. For example, for an unbiased coin, the probability of observing heads is 0.5 for every toss. This is taken to mean that if a coin were tossed a large number of times then we would expect, on average, to find half of the time the coin landed heads, half of the time tails.

There are certain *laws of probability* that allow us to make inferences and predictions based on probabilistic information. For example, the probabilities of different outcomes for a certain event must always add up to 1: if there is a 20% chance of rain today, there must be an 80% chance of no rain. Another very common law is that if two events are independent of one another (that is, they in no way influence each other), then the probability of certain pairs of outcomes will be the product of the two outcomes by themselves: if we toss a coin twice, the probability of getting 2 heads is 0.5 times 0.5 = 0.25.

Models: parameters and distributions

When we speak about the probability of observing events such as the outcome of a toss of a coin, we are implicitly assuming some kind of *model*, even in this simple case. In the case of a coin, the model would state that there is some certain, fixed probability for the particular outcomes. This model would have one *parameter*, *p* the probability of the coin landing on heads. If the coin is fair, then *p*=0.5. We can then speak about the probability of observing an event, given specific parameter values for the model. In this simple case, if *p* =0.5, then the probability of the coin landing heads on any one toss is also 0.5.

In the case of this simple example, it does not seem that we have gained very much - we seem to be merely calling what was previously a simple probability the *parameter* of a *model*. As we shall see, however, this way of thinking provides a very useful framework for expressing more complex problems.

Conditional probability

In the real world, very few things have absolute, fixed probabilities. Many of the aspects of the world that we are familiar with are not truly random. Take for instance, the probability of developing schizophrenia. Say that the prevalence of schizophrenia in a population is 1%. If we know nothing else about an individual, we would say that the probability of this individual developing schizophrenia is 0.01. In mathematical notation,

$$P(Sz) = 0.01$$

We know from empirical research, however, that certain people are more likely to develop schizophrenia than others. For example, having a schizophrenic first-degree relative greatly increases the risk of becoming schizophrenic. The probability above is essentially an average probability, taken across all individuals both with and without schizophrenic first-degree relatives.

The notion of *conditional probability* allows us to incorporate other potentially important variables, such as the presence of familial schizophrenia, into statements about the probability of an individual developing schizophrenia. Mathematically, we write

$$P(X \mid Y)$$

meaning the probability of X *conditional on Y* or *given Y*. In our example, we could write

$$P\ (Sz \mid \text{first degree relative has } Sz)$$

and

$$P\ (Sz \mid \text{first degree relative does not have } Sz)$$

Whether or not these two values differ is an indication of the influence of familial schizophrenia upon an individual's chances of developing schizophrenia.

Previously, we mentioned that all probability statements depend on some kind of model in some way. The probability of an outcome will be *conditional* upon the parameter values of this model. In the case of the coin toss,

```
        P (H | p=0.5)
```

where H is the event of obtaining a head and p is the model parameter, set at 0.5.

Let's think a little more carefully about what the full model would be for tossing a coin, if *p* is the parameter. What do we know about coin tossing?

- The outcome is a discrete, binary outcome for each toss - it is either heads or tails.
- We *assume* that the probability of either outcome does not change over time.
- We assume that the outcome of each toss of a coin can be regarded as independent from all other outcomes. That is, getting five heads in a row does not make it any more likely to get a tail on the next trial.
- In the case of a 'fair' coin, we assume a 50:50 chance getting either heads or tails - that is, *p*=0.5.

Say we toss a coin a number of times and record the number of times it lands on heads. The probability distribution that describes just this kind of scenario is called the *binomial* probability distribution. It is written as follows :

$$\frac{n!}{h!(n-h)!}p^h(1-p)^{n-h}$$

Let's take a moment to work through this. The notation is as follows:-

- *n* = total number of coin tosses
- *h* = number of heads obtained
- *p* = probability of obtaining a head on any one toss

(The ! symbol means *factorial* (5! = 1x2x3x4x5 = 120).)

We can think of this equation in two parts. The second part involves the joint probability of obtaining *h* heads (and therefore *n*-*h* tails) if a coin is tossed *n* times and has probability *p* of landing heads on any one toss (and therefore probability 1-*p* of landing tails). *Because we have assumed that each of the n trails is independent and with constant probability* the *joint probability* of obtaining *h* heads and *n*-*h* tails is simply the product of all the individual probabilities. Imagine we obtained 4 heads and 5 tails in 9 coin tosses. Then

$$p^4(1-p)^5$$

is simply convenient notation for

$$p\times p\times p\times p\times(1-p)\times(1-p)\times(1-p)\times(1-p)\times(1-p)$$

The first half of the binomial distribution function is concerned with the fact that there is more than 1 way to get, say, 4 heads and 5 tails if a coin is tossed 9 times. We might observe

```
   H, T, H, H, T, T, H, T, T.
```

or

```
   T, H, H, T, H, T, T, H, T.
```

or even

```
   H, H, H, H, T, T, T, T, T.
```

Every one of the permutations is assumed to have equal probability of occurring - the coefficient

$$\frac{n!}{h!(n-h)!}$$

represents the total number of permutations that would give 4 heads and 5 tails.

So, the probability of obtaining 4 heads and 5 tails for a fair coin is

$$\frac{9!}{4!5!}0.5^{4}0.5^{5} = 0.246$$

Return to front page

Site created by S.Purcell, last updated 1.10.2000

Maximum Likelihood Estimation (MLE)

Model-fitting

Now we are in a position to introduce the concept of likelihood.

If the probability of an event X dependent on model parameters *p* is written

```
   P ( X | p )
```

then we would talk about the likelihood

```
   L ( p | X )
```

that is, the likelihood of *the parameters given the data*.

For most sensible models, we will find that certain data are more probable than other data. The aim of maximum likelihood estimation is to find the parameter value(s) that makes the observed data most likely. This is because the likelihood of the parameters given the data is defined to be equal to the probability of the data given the parameters

(nb. technically, they are proportional to each other, but this does not affect the principle).

If we were in the business of making predictions based on a set of solid assumptions, then we would be interested in probabilities - the probability of certain outcomes occurring or not occurring.

However, in the case of *data analysis*, we have already observed all the data: once they have been observed they are fixed, there is no 'probabilistic' part to them anymore (the word data comes from the Latin word meaning 'given'). We are much more interested in the likelihood of the model parameters that underly the fixed data.

```
Probability
   Knowing parameters  -> Prediction of outcome

Likelihood
   Observation of data -> Estimation of parameters
```

A simple example of MLE

To re-iterate, the simple principle of maximum likelihood parameter estimation is this: find the parameter values that make the observed data most likely. How would we go about this in a simple coin toss experiment? That is, rather than assume that *p* is a certain value (0.5) we might wish to find the *maximum likelihood estimate* (MLE) of *p*, given a specific dataset.

Beyond parameter estimation, the likelihood framework allows us to make *tests* of parameter values. For example, we might want to ask whether or not the estimated *p* differs *significantly* from 0.5 or not. This test is essentially asking: is there evidence that the coin is biased? We will see how such tests can be performed when we introduce the concept of a *likelihood ratio test* below.

Say we toss a coin 100 times and observe 56 heads and 44 tails. Instead of *assuming* that *p* is 0.5, we want to find the MLE for *p*. Then we want to ask whether or not this value differs significantly from 0.50.

How do we do this? We find the value for *p* that makes the observed data most likely.

As mentioned, the observed data are now fixed. They will be constants that are plugged into our binomial probability model :-

- n = 100 (total number of tosses)
- h = 56 (total number of heads)

Imagine that *p* was 0.5. Plugging this value into our probability model as follows :-

$$L(p = 0.5 \mid data) = \frac{100!}{56!44!} 0.5^{56} 0.5^{44} = 0.0389$$

But what if *p* was 0.52 instead?

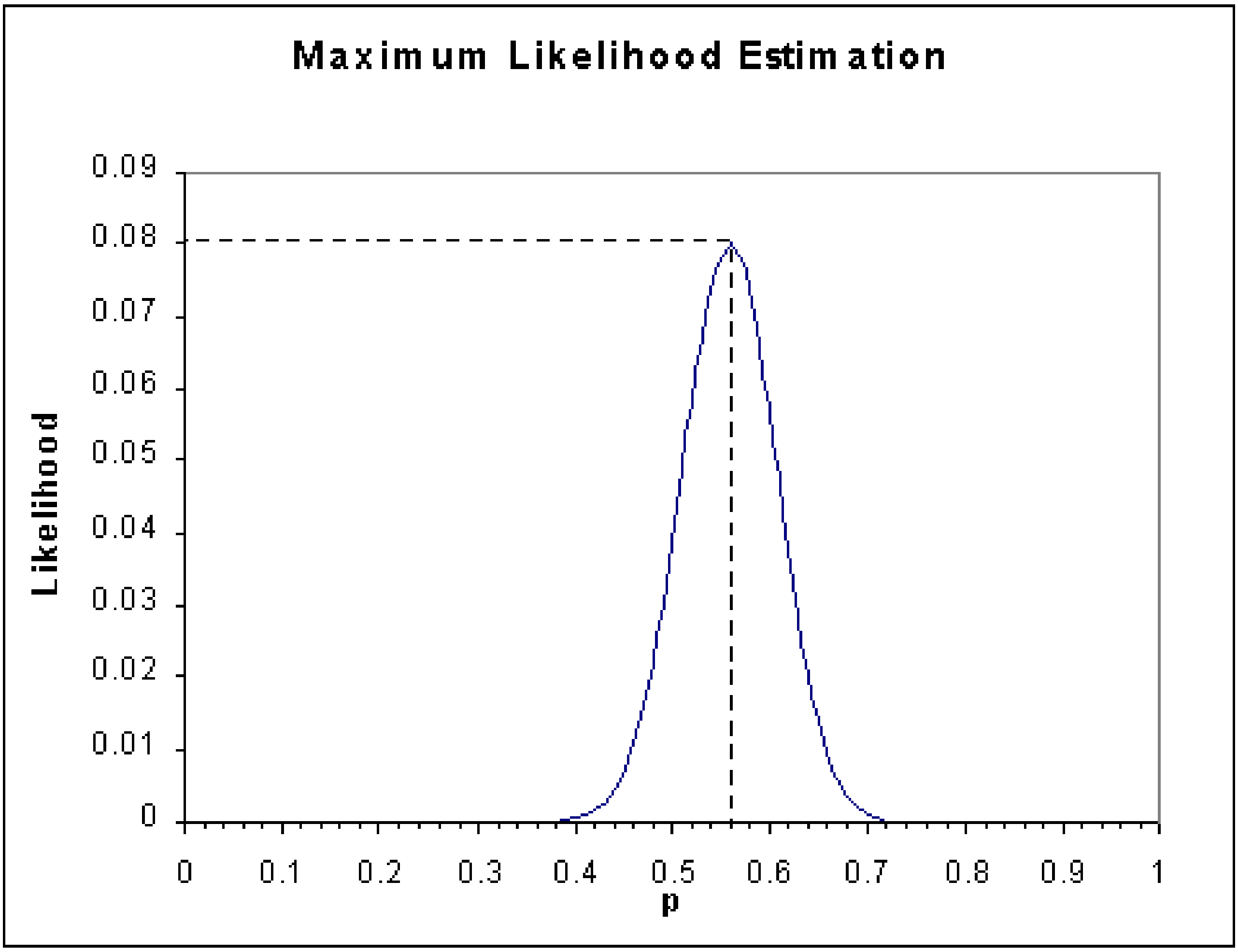
$$L(p = 0.52 \mid data) = \frac{100!}{56!44!} 0.52^{56} 0.48^{44} = 0.0581$$

So from this we can conclude that *p* is more likely to be 0.52 than 0.5. We can tabulate the likelihood for different parameter values to find the maximum likelihood estimate of *p*:

p	L

0.48	0.0222
0.50	0.0389
0.52	0.0581
0.54	0.0739
0.56	0.0801
0.58	0.0738
0.60	0.0576
0.62	0.0378

If we graph these data across the full range of possible values for *p* we see the following *likelihood surface*.



We see that the maximum likelihood estimate for *p* seems to be around 0.56. In fact, it is exactly 0.56, and it is easy to see why this makes sense in this trivial example. The best estimate for *p* from any one sample is clearly going to be the proportion of heads observed in that sample. (In a similar way, the best estimate for the population mean will always be the sample mean.)

So why did we waste our time with the maximum likelihood method? In such a simple case as this, nobody would use maximum likelihood estimation to evaluate *p*. But not all problems are this simple! As we shall see, the more complex the model and the greater the number of parameters, it often becomes very difficult to make even reasonable guesses at the MLEs. The likelihood framework conceptually takes all of this in its stride, however, and this is what makes it the work-horse of many modern statistical methods.

Return to front page

Site created by S.Purcell, last updated 21.09.2000

Maximum Likelihood Estimation (MLE)

MLE in Practice

Analytic MLE

Sometimes we can write a simple equation that describes the *likelihood surface* (e.g. the line we plotted in the coin tossing example) that can be differentiated. In this case, we can find the maximum of this curve by setting the first derivative to zero. That is, this represents the peak of a curve, where the gradient of the curve turns from being positive to negative (going left to right). In theory, this will represent the maximum likelihood estimate of the parameter.

Numerical MLE

But often we cannot, or choose not, to write an equation that can be differentiated to find the MLE parameter estimates. This is especially likely if the model is complex and involves many parameters and/or complex probability functions (e.g. the normal probability distribution).

In this scenario, it is also typically not feasible to evaluate the likelihood at all points, or even a reasonable number of points, in the *parameter space* of the problem as we did in the coin toss example. In that example, the parameter space was only one-dimensional (i.e. only one parameter) and ranged between 0 and 1. Nonetheless, because *p* can theoretically take any value between 0 and 1, the MLE will always be an approximation (albeit an incredibly accurate one) if we just evaluate the likelihood for a finite number of parameter values. For example, we chose to evaluate the likelihood at steps of 0.02. But we could have chosen steps of 0.01, of 0.001, of 0.000000001, etc. In theory and practice, one has to set a minimum *tolerance* by which you are happy for your estimates to be out. This is why computers are essential for these types of problems: they can tabulate lots and lots of values very quickly and therefore achieve a much finer resolution.

If the model has more than one parameter, the parameter space will grow very quickly indeed. Evaluating the likelihood exhaustively becomes virtually impossible - even for computers. This is why so-called *optimisation* (or *minimisation*) algorithms have become indispensable to statisticians and quantitative scientists in the last couple of decades. Simply put, the job of an optimisation algorithm is to *quickly* find the set of parameter values that make the observed data most likely. They can be thought of as intelligently playing some kind of hotter-colder game, looking for a hidden object, rather than just starting at one corner and exhaustively searching the room. The 'hotter-colder' information these algorithms utilise essentially comes from the way in which the likelihood *changes* as the they move across the parameter space. Note that it is precisely this type of 'rate of change' information that the analytic MLE methods use - differentiation is concerned with the rate of change of a quantity (i.e. the likelihood) with respect to some other factors (i.e. the parameters).

Other Practical Considerations

Briefly, we shall look at a couple of shortcuts and a couple of problems that crop up in maximum likelihood estimation using numerical methods:

Removing the constant

Recall the likelihood function for the binomial distribution:

$$\frac{n!}{h!(n-h)!}\,p^{h}(1-p)^{n-h}$$

In the context of MLE, we noted that the values representing the data will be fixed: these are *n* and *h*. In this case, the binomial 'co-efficient' depends only upon these constants. Because it does not depend on the value of the parameter *p* we can essentially ignore this first term. This is because any value for *p* which maximises the above quantity will also maximise

$$p^{h}(1-p)^{n-h}$$

This means that the likelihood will have no meaningful scale in and of itself. This is not usually important, however, for as we shall see, we are generally interested not in the absolute value of the likelihood but rather in the *ratio* between two likelihoods - in the context of a likelihood ratio test.

We may often want to ignore the parts of the likelihood that do not depend upon the parameters in order to reduce the computational intensity of some problems. Even in the simple case of a binomial distribution, if the number of trials becomes very large, the calculation of the factorials can become infeasible (most pocket calculators can not represent numbers larger than about 60!). (Note: in reality, we would quite probably use an approximation of the binomial distribution, using the normal distribution that does not involve the calculation of factorials).

Log-likelihood

Another technique to make life a little easier is to work with the natural log of likelihoods rather than the likelihoods themselves. The main reason for this is, again, computational rather than theoretical. If you multiply lots of very small numbers together (say all less than 0.0001) then you will very quickly end up with a number that is too small to be represented by any calculator or computer as different from zero. This situation will often occur in calculating likelihoods, when we are often multiplying the probabilities of lots of rare but independent events together to calculate the joint probability.

With log-likelihoods, we simply add them together rather than multiply them (log-likelihoods will always be negative, and will just get larger (more negative) rather than approaching 0). Note that if

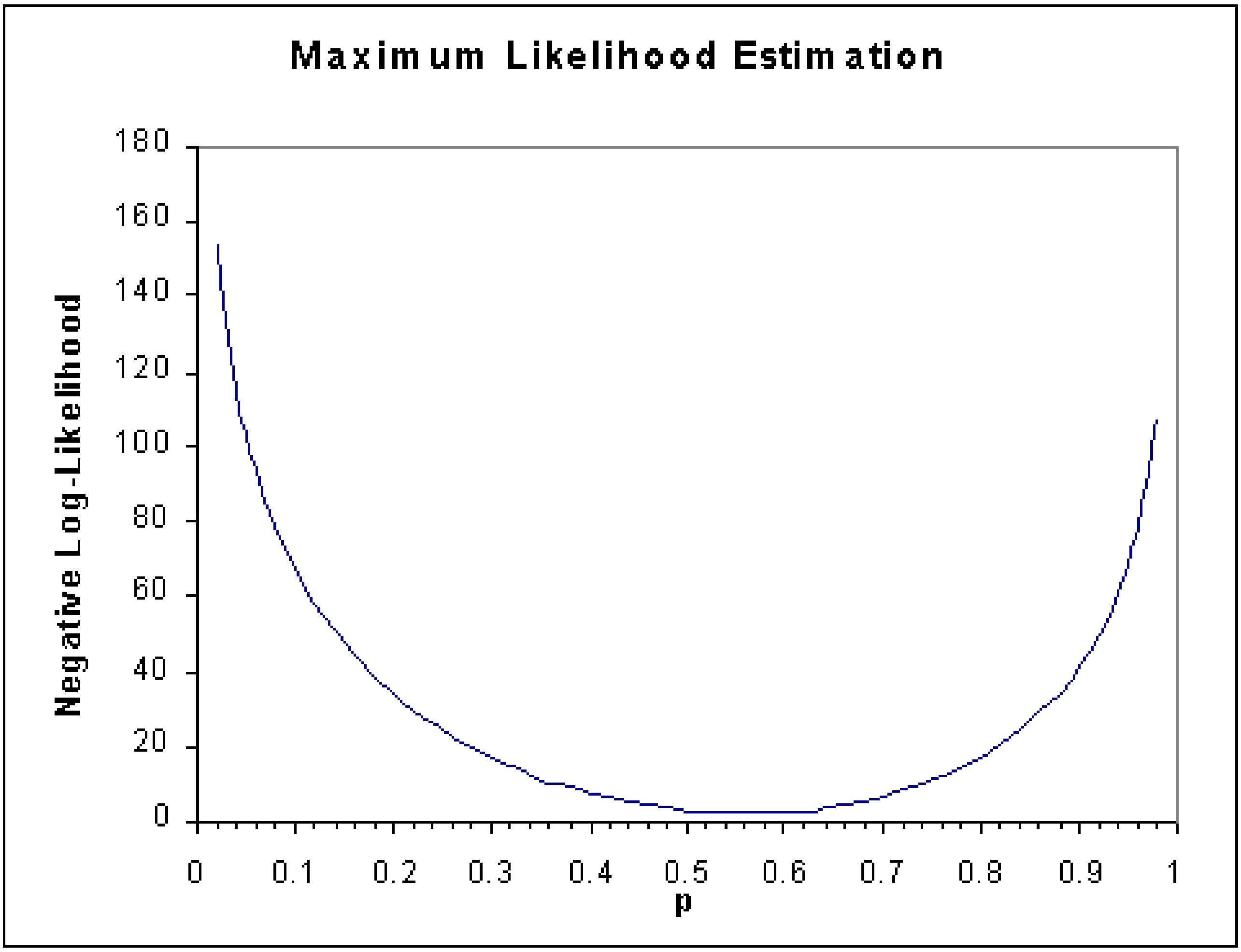
$$a = bc$$

then

$$log(a) = log(b) + log(c)$$

So, log-likelihoods are conceptually no different to normal likelihoods. When we optimise the log-likelihood (note: technically, we will be *minimising* the *negative* log-likelihood) with respect to the model parameters, we also optimise the likelihood with respect to the same parameters, for there is a one-to-one (monotonic) relationship between numbers and their logs.

For the coin toss example above, we can also plot the log-likelihood. We can see that it gives a similar MLE for *p* (note: here we plot the negative of the log-likelihood, merely because most optimisation procedures tend to be formulated in terms of minimisation rather than maximisation).



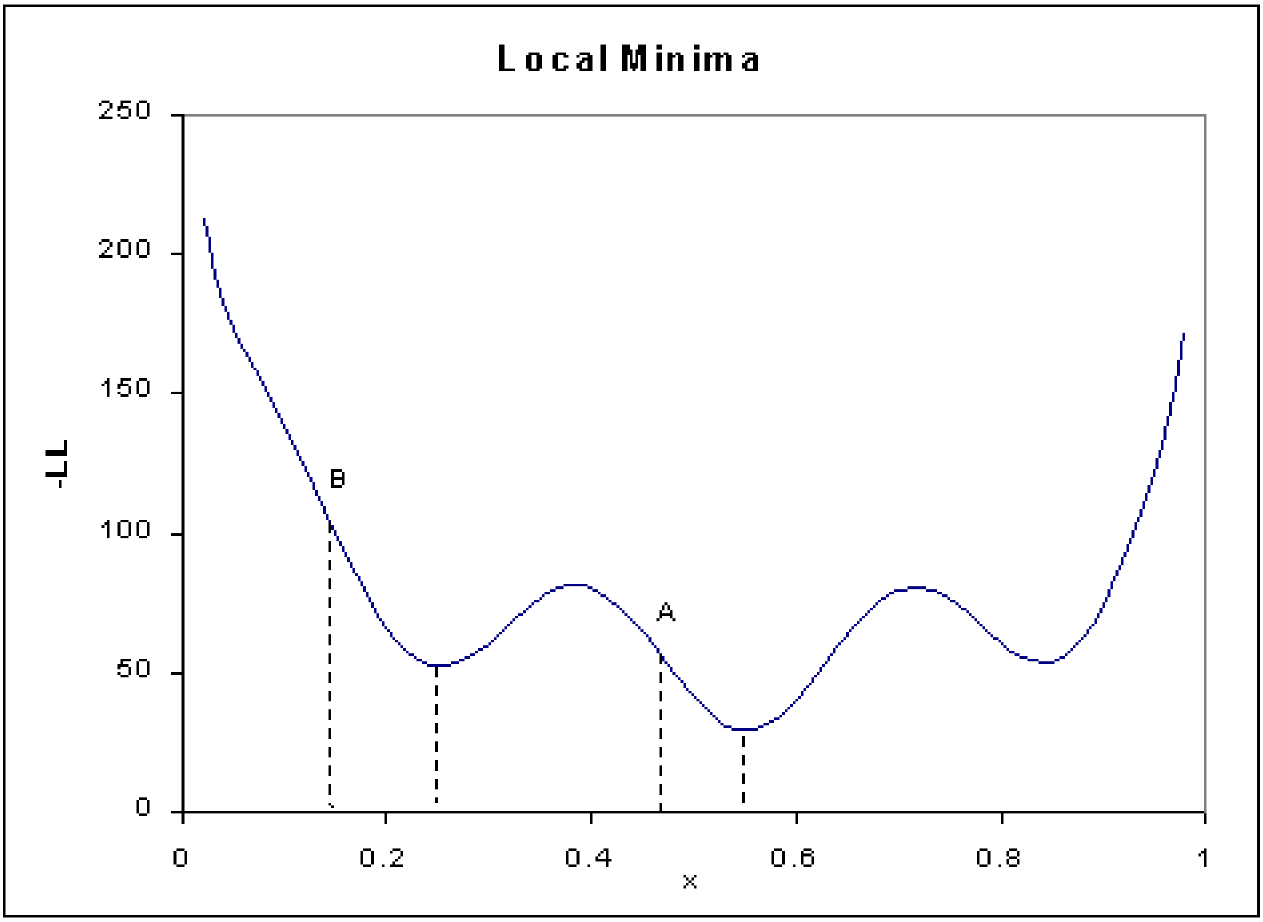
Model identification

It is worth noting that it is not always possible to find one set of parameter values that uniquely optimises the log-likelihood. This may occur if there are too many parameters being estimated for the type of data that has been collected. Such a model is said to be 'under-identified'.

A model that attempted to estimate additive genetic variation, dominance genetic variation *and* the shared environmental component of variance from just MZ and DZ twin data would be under-identified.

Local Minima

Another common practical problem when implementing model-fitting procedures is that of local minima. Take the following graph, which represents the negative log-likelihood plotted by a parameter value, *x*.



Model fitting is an iterative procedure: the user has to specify a set of *starting values* for the parameters (essentially an initial 'first guess') which the optimisation algorithm will take and try to improve on.

It is possible for the 'likelihood surface' to be any complex function of a parameter value, depending on the type of model and the data. In the case below, if the starting value for parameter *x* was at point *A* then optimisation might find the true, *global* minimum. However, if the starting value was at point *B* then it might not find instead only a local minimum. One can think of the algorithm crawling down the slope from *B* and thinking it has reached the lowest point when it starts to rise again. The implication of this would be that the optimisation algorithm would stop too early and return a sub-optimal estimate of the parameter *x*. Avoiding this kind of problem often involves specifying models well, choosing appropriate optimisation algorithms, choosing sensible starting values and more than a modicum of patience.

Return to front page

Site created by S.Purcell, last updated 21.09.2000

Maximum Likelihood Estimation (MLE)

The likelihood ratio test

Model-fitting provides a framework within which we can not just estimate the maximum likelihood estimates for parameters: we can also test whether or not they are significantly different from other fixed values.

The likelihood ratio test provides the means for comparing the likelihood of the data under one hypothesis (usually called the *alternate* hypothesis) against the likelihood of the data under another, more restricted hypothesis (usually called the *null* hypothesis, for the experimenter tries to *nullify* this hypothesis in order to provide support for the former).

For example, we may wish to ask: was the coin we tossed 100 times fair? This is rephrased as :

Alternate hypothesis (H_A) : *p* does not equal 0.50
Null hypothesis (H_0)   : *p* equals 0.50

The likelihood ratio test answers this question: are the data significantly less likely to have arisen if the null hypothesis is true than if the alternate hypothesis is true?

We proceed by calculating the likelihood under the alternate hypothesis, then under the null, then we calculate test the difference between these two likelihoods

$$2 (LL_A - LL_0)$$

Note that if *a=b/c* then *log(a)=log(b)-log(c)*. This is why it is called a likelihood ratio test, but we look at the difference between log-likelihoods.

The difference between the likelihoods is multiplied by a factor of 2 for technical reasons, so that this quantity will be distributed as the familiar χ^2 statistic. This can then be assessed for statistical significance using standard χ^2 significance levels. In most simple cases, the degrees of freedom for the test will equal the difference in the number of parameters being estimated under the alternate and null models. In the case of the coin, we estimate one parameter under the alternate (*p*) and none under the null (as *p* is fixed) so the χ^2 has 1 degree of freedom.

In the case of the coin tossing experiment, comparing the log-likelihood under the alternate (i.e. when *p* is estimated at its MLE) and the null (i.e. when *p* is fixed at 0.50):

	Alternate	Null
p	0.56	0.50
Likelihood	0.0801	0.0389
Log Likelihood	-2.524	-3.247

$$2(L_A - L_0) = 2 * (-2.524 + 3.247) = 1.446$$

Therefore, as the critical significance level for a 1 degree of freedom χ^2 is 3.84

(see the Probability Function Calculator also on this site) we can conclude that the fit is *not* significantly worse under the null. That is, we have no reason to reject the null hypothesis that the coin is fair. So, the answer to the question is that the data are indeed consistent with the coin being a fair coin.

Return to front page

Site created by S.Purcell, last updated 21.09.2000

Maximum Likelihood Estimation (MLE)

MLE for twin data

How does all of this apply to twins and the kind of complex, quantitative traits that we wish to study?

The fundamental principles of maximum likelihood still apply in exactly the same manner as for the coin-tossing experiment. What change are the data we measure and the form of the probability model that describes these data.

In the case of coin tossing, we observed two items of data: *n* the total number of tosses and *h* the number of heads. For twins, in the most basic case, we would collect three pieces of information for each twin pair:

- a trait measure for twin 1
- a trait measure for twin 2
- whether they are identical or not (MZ vs DZ)

For the coin tossing, we used a binomial distribution to model the data. Typically, for quantitative traits, we would assume that our observations come from a *normally-disbtributed* trait population (bell-shaped curve). As the unit of analysis is a twin pair (i.e. involving two variables rather than one) we need to use the *bivariate* form of the normal distribution. This specifically describes distributions of *pairs* of scores.

Finally, in the coin-tossing experiment we had one parameter in our model, representing the probability of obtaining a head. In the case of twins, we would generally have three parameters (four if we include a means model, see below):

- *a* : proportion of variance attributable to additive genetic variation
- *c* : proportion of variance attributable to shared environmental variation
- *e* : proportion of variance attributable to nonshared environmental variation
- *m* : trait mean

Traditionally, we would say the that binomial distribution takes two parameters, *n* the total number of trials and *p* the probability of success. A *random variable*, say *X* that has a binomial distribution is written :

```
X~B(n, p)
```

and we are interested in *P(X=x)*: that is, the probability that the the random variable *X* has the specific value *x*. In our coin tossing example, *h*, the observed number of heads, is equivalent to *x*. Recall,

$$\frac{n!}{h!(n-h)!}p^h(1-p)^{n-h}$$

Similarly, the normal distribution has two parameters. These parameters are in terms of the mean and variance of the distribution rather than probabilities of success and numbers of trials.

[reword this para].These we shall call μ which is the trait mean and σ which is the trait standard deviation. A *random variable*, say X that has a normal distribution is written :

$$X \sim N(\mu, \sigma)$$

The standard formula which defines P(X=x) for the *bivariate* normal distribution is

$$P(\mathbf{X}=\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}}\exp\left\{-\frac{(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}{2}\right\}$$

Exactly what the component terms of these formula represent is not important - in any case, it is beyond the scope of this tutorial. The important point to note is that the normal probability function is determined by only two parameters (although these parameters are actually matrices):

- μ : a vector of means (two means in the bivariate case)
- Σ : the covariance matrix (a two-by-two matrix in the bivariate case)

(Each pair's trait scores are in the vector **x** and *p* represents the number of variables, i. e. 2 in the bivariate case.)

But we said that the coin tossing model only had one parameter, and that the model we fit to twin data would have 3 or 4 parameters? This is the distinction between parameters of a probability distribution and model parameters. Most model fitting involves some kind of *re-parameterisation*, but there is a direct correspondence between the two types of parameters.

The following table gives the relationships:

Binomial Probability Model		**Coin Tossing Model**
N (number of trials)	---->	N (observed data)
P (probability of success)	---->	P (estimated parameter)
Normal Probability Model		**Twin Design Model**
μ (mean vector)	---->	m (estimated or fixed parameter)
Σ (covariance matrix)	---->	a, c, e (estimated or fixed parameters)

In the case of the coin tossing experiment, there was a one-to-one correspondence between the *parameters* of the binomial probability function and the underlying model. That is, *p* the probability of 'success' in the binomial model is very directly equivalent to *p* the probability of getting heads in our model.

In the case of fitting a normal distribution to twin data, *parameters* can either refer to the direct parameters of the normal distribution (the mean vector and covariance matrix) or the parameters of the underlying genetic model (proportion of trait variation attributable to additive genetic variation, etc.)

Model-fitting for twin data proceeds by specifying the mean vector and covariance matrix of the normal distribution *in terms of the genetic parameters of interest*. As we shall see in the next section, this is done according to basic biometrical assumptions and allows to us estimate quantities of interest providing we have collected suitably informative data.

Now we are ready to model fit to twin data

As mentioned elsewhere in this course, twin analysis essentially models the covariation between identical and non-identical twins. The comparison of an MZ twin correlation with a DZ twin correlation allows us to estimate the effects of additive genetic influences, shared environmental influences and nonshared environmental influences.

Specifically, we are re-parameterising the twin covariance structure in terms of the parameters a, c and e (as mentioned above). The covariance matrix for a sample of twin pairs contains three unique values:

- the variance of twin 1
- the variance of twin 2
- the covariance between twin 1 and twin 2

According biometrical theory the trait variance can be *decomposed* into independent *components of variance*, and the trait covariance, conditional on twin zygosity, can be expressed in terms of these components of variance also.

- Trait variance = a + c + e
- MZ covariance = a + c
- DZ covariance = 0.5a + c

We can therefore write the trait covariance matrices for MZ and DZ twins in terms of these three components of variance. For MZ twins

$$\Sigma_{MZ} = \begin{bmatrix} V_A + V_C + V_E & V_A + V_C \\ V_A + V_C & V_A + V_C + V_E \end{bmatrix}$$

whilst for DZ twins

$$\Sigma_{DZ} = \begin{bmatrix} V_A + V_C + V_E & V_A/2 + V_C \\ V_A/2 + V_C & V_A + V_C + V_E \end{bmatrix}$$

The Means Model

Because the twin design is primarily an analysis of *individual differences* we are typically only interested in the components of variance - that is, modelling the twin covariance structure. The normal distribution requires a *means model* however. We could either let all four means (i.e. twin 1 and twin 2 for MZ and DZ twins) be estimated independently, or we could constrain all four measures to be estimated at the same value. The latter option would be the typical choice: conditional on the means not being significantly different from each other, this will provide a more powerful test for fewer parameters are being estimated. (Note: *if the means are different* in a standard twin design, this may well be indicative of some problem in ascertainment or data management.)

Raw data versus Summary Statistics

We can either formulate models in terms of the raw unit of observation or it may be possible to model certain *summary statistics* instead. In the coin tossing example, the summary statistics were the total number of tosses and the number of heads. These two summary statistics contained all of the information relevant to the problem - that is, given these summary statistics it was not important that we knew the *actual* sequence of heads and tails.

In a similar way, the mean vector and covariance matrix are said to be *sufficient* summary statistics in the sense that, under the assumption of normality, we gain nothing by analysing the raw data (i.e. all actual scores for each twin pair) if we know what the mean vectors and covariance matrices are for all MZ and all DZ pairs.

Indeed, it is common practice to ignore the means model and only analyse the covariance matrices for twins. Model-fitting to summary statistics instead of raw data has a slightly more complicated form, which essentially allow computational shortcuts. These shortcuts were more or less essential in the 1960s and 1970s when MLE techniques were first being implemented. Nowadays, analysis of raw data is computationally not a problem.

From the point of view of *using* model-fitting software such as *Mx* it makes little or no difference whether or not the model is fitted to raw data or summary statistics. The main difference is, obviously, just in how the information are entered into the program:

Raw Data

```
     Input                      Output
                         (estimated parameters)


Twin1  Twin2  Zyg                 a
-0.23  -0.41  1                   c
0.43   1.32   1                   e
-0.47  0.76   2                   m
1.23   0.65   2
-1.62  -0.44  1
...    ...    .
```

```
Covariance Matrices
     Input                      Output
                          (estimated parameters)

MZ  1.32                          a
    0.87  1.28                    c
                                  e
DZ  1.29
    0.54  1.35
```

However, analysing raw data does have certain advantages :

- outliers can be easily detected
- covariates can be easily incorporated
- missing values can be dealt with efficiently
- more complex gene-by-environment interaction models can be implemented easily

For basic twin ACE models, fitting to covariance matrices will be sufficient.

MLE for twin data

For the purpose of understanding MLE in the context of analysing twin data, it is more transparent to think in terms of the the analysis of raw data. Model-fitting proceeds in the standard way :

1. select starting values for the parameters (*a, c, e, m*)
2. evaluate the log-likelihood for the first twin pair using the normal probability distribution and zygosity-specific models of the twin covariance
3. sum the log-likelihoods over all twin pairs in the sample
4. optimise the sample log-likelihood with respect to the model parameters
5. the *output* is then the values for the sample parameters and the log-likelihood
6. the likelihood ratio test can be used to compare the full model which estimates all the parameters with submodels that constrain one or more of the parameters to be zero
7. select the most parsimonious model that explains the data

Return to front page

Site created by S.Purcell, last updated 21.09.2000

Maximum Likelihood Estimation (MLE)

MLE analysis of linkage data

If we have a sample in which the number of recombinants and non- recombinants for two specific loci can be counted, then we can estimate the recombination fraction between between those two loci.

The test for linkage is simply the test of whether the recombination fraction (θ) is 0.5 (the null hypothesis of no linkage) or less than 0.5 (the alternative hypothesis of linkage).

You might have noticed a striking similarity to the coin-flipping example here. The good news is that the analysis is virtually identical. Note that, in real life, we would not expect to observe fully informative gametes for all pedigrees, and more complex methods have to fill in the gaps, but the principles are much the same.

Suppose that we observe N fully informative gametes, of which R are recombinants. How do we test for linkage and estimate the recombination fraction, θ ?

Since each gamete has probability θ of being recombinant and probability (1- θ) of being non-recombinant, the likelihood function is

$$L(\theta \mid N, R) = \theta^{R} (1-\theta)^{N-R}$$

Note : strictly speaking, the likelihood is proportional to this quantity rather than equal to it - notice that the constant part of the binomial formula has been dropped.

The log-likelihood function is therefore

$$\ln L(\theta \mid N, R) = R \ln \theta + (N-R) \ln(1-\theta)$$

The null hypothesis of no linkage implies θ =0.5, so the value of the log-likelihood function is

$$\ln L_0 = N \ln(\tfrac{1}{2})$$

As we know that the maximum likelihood estimate for θ is simply the proportion of recombinant gametes

$$\hat{\theta} = \frac{R}{N}$$

when R<(n/2), otherwise

$$\hat{\theta} = \frac{1}{2}$$

for biological reasons

Under the alternative of linkage, the maximum log-likelihood is

$$\ln L_A = R \ln(R/N) + (N - R) \ln(1 - R/N)$$

where R<(n/2) and

$$\ln L_A = \ln L_0$$

when R>(N/2).

The likelihood ratio statistic $2(\ln L_A - \ln L_0)$ provides a direct test for linkage. Note: this likelihood ratio statistics is distributed as a 50:50 mixture of chi-squared with one degree of freedom and point probability mass of 0. In this way, a one-tailed test of linkage is provided.

In linkage analysis, it is customary to take the common (base 10) logarithm of the likelihood function, and then define the difference between the log-likelihood at a certain value of θ and the log-likelihood at θ =0.5 to be the "lod-score" at that value of θ . The maximum lod-score occurs at the MLE of θ : its value is equal to the likelihood ratio statistic divided by a factor of 2ln10 (approximately 4.6).

An Example

Suppose that between two loci we observe

- 27 recombinants
- from 139 fully informative gametes

What is the evidence for linkage?

The MLE estimate of the recombination fraction is therefore

$$27 \;/\; 139 = 0.1942$$

The log-likelihood at the MLE of the recombination fraction is

$$\begin{aligned} \ln L_A &= 27 * \ln(0.1942) + (139 - 27) * \ln(1-0.1942) \\ &= -68.43 \end{aligned}$$

whereas under the null of no linkage it is

$$\begin{aligned} \ln L_0 &= 139 * \ln(0.5) \\ &= -96.35 \end{aligned}$$

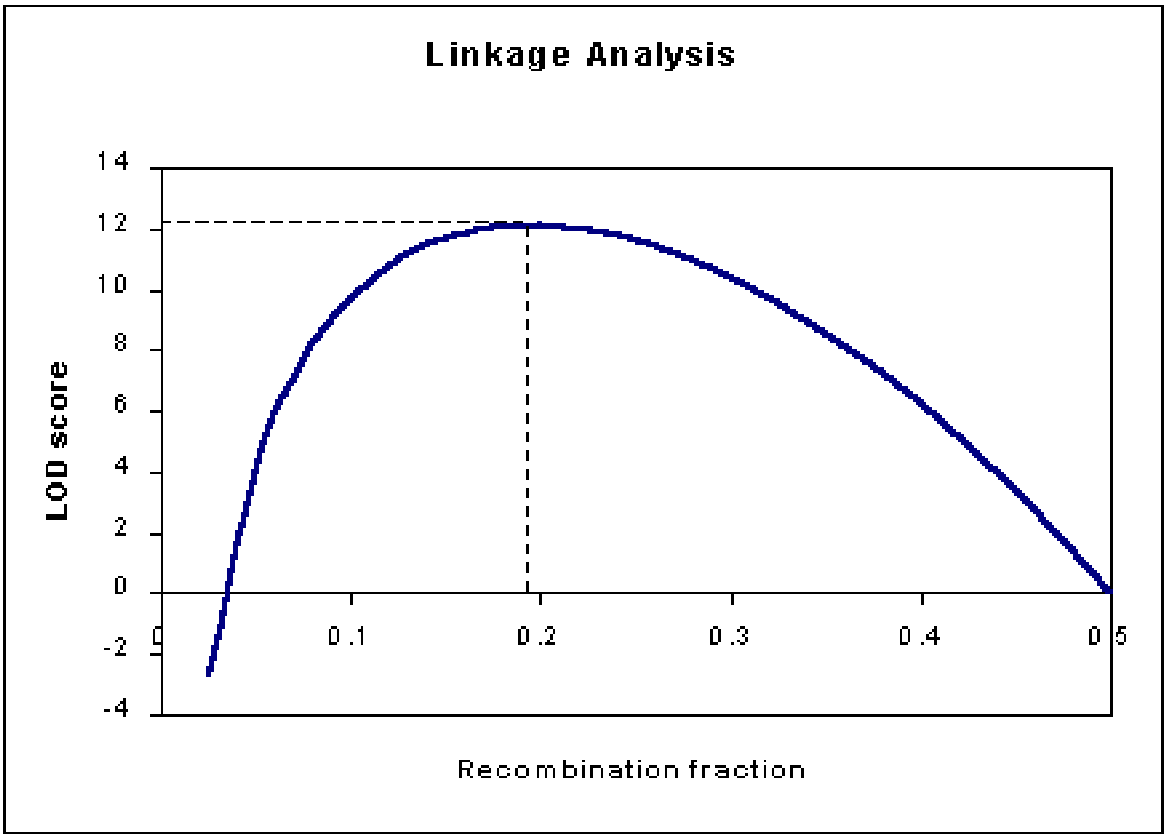
This gives a value of

$$\begin{aligned} 2(L_A - L_0) &= 2 * -68.43 - (-96.35) \\ &= 55.84 \end{aligned}$$

This is clearly highly significant, corresponding to a lod-score of approximately

$$\begin{aligned} \text{LOD} &= 55.84 \;/\; 4.6 \\ &= 12.1 \end{aligned}$$

We can plot the lod-score curve for different values of θ :



From this we can draw up so-called *support-intervals* that give an equivalent of a confidence interval around the point maximum likelihood estimate of the recombination fraction. Typically, one would drop down one lod score unit either side of the MLE - in this case, this localises the linkage as approximately 0.13 - 0.27.

Return to front page

Site created by S.Purcell, last updated 21.09.2000

Maximum likelihood

From Wikipedia, the free encyclopedia

Maximum likelihood estimation (MLE) is a popular statistical method used to make inferences about parameters of the underlying probability distribution of a given data set.

The method was pioneered by geneticist and statistician Sir R. A. Fisher between 1912 and 1922 (see external resources below for more information on the history of MLE).

Contents

Prerequisites

Following discussions assume that readers are familiar with basic notions in probability theory such as probability distributions, probability density functions, random variables and expectation. It also assumes they are familiar with standard basic techniques of maximising continuous real-valued functions, such as using differentiation to find a function's maxima.

The principles of MLE

Given a parameterized family D_θ of probability distributions associated with either a known probability density function (continuous distribution) or a known probability mass function (discrete distribution), denoted as f_θ, we may draw a sample $x_1, x_2, \ldots, x_n$ of n values from this distribution and then using f_θ we may compute the probability density associated with our observed data:

$$f_\theta(x_1, \ldots, x_n \mid \theta).$$

As a function of θ with $x_1, \ldots, x_n$ fixed, this is the likelihood function

$$L(\theta) = f_\theta(x_1, \ldots, x_n \mid \theta).$$

When θ is not observable, the method of maximum likelihood uses the value of θ that maximizes *L*(θ) as an estimate of θ. This is the **maximum likelihood estimator (MLE)** $\widehat{\theta}$ of θ

This contrasts with seeking an unbiased estimator of θ, which may not necessarily yield the most likely value of θ but which will yield a value that (on average) will neither tend to over-estimate nor under-estimate the true value of θ.

The maximum likelihood estimator may not be unique, or indeed may not even exist.

Examples

Discrete distribution, discrete and finite parameter space

Consider tossing an unfair coin 80 times (i.e., we sample something like $x_1 = \mathrm{H}, x_2 = \mathrm{T}, \ldots, x_{80} = \mathrm{T}$ and count the number of HEADS H observed). Call the probability of tossing a HEAD *p*, and the probability of tossing TAILS 1 − *p* (so here *p* is the parameter which we referred to as θ above). Suppose we toss 49 HEADS and 31 TAILS, and suppose the coin was taken from a box containing three coins: one which gives HEADS with probability *p* = 1 / 3, one which gives HEADS with probability *p* = 1 / 2 and another which gives heads with probability *p* = 2 / 3. The coins have lost their labels, so we don't know which one it was. Using **maximum likelihood estimation** we can calculate which coin it was most likely to have been, given the data that we observed. The likelihood function (defined above) takes one of three values:

$$\mathbb{P}(\text{we toss 49 HEADS out of 80} \mid p = 1/3) \quad = \quad \binom{80}{49}(1/3)^{49}(1-1/3)^{31} \approx 0.000$$

$$\mathbb{P}(\text{we toss 49 HEADS out of 80} \mid p = 1/2) \quad = \quad \binom{80}{49}(1/2)^{49}(1-1/2)^{31} \approx 0.012$$

$$\mathbb{P}(\text{we toss 49 HEADS out of 80} \mid p = 2/3) \quad = \quad \binom{80}{49}(2/3)^{49}(1-2/3)^{31} \approx 0.054$$

We see that the likelihood is maximised by parameter $\widehat{p} = 2/3$, and so this is our *maximum likelihood estimate* for *p*.

Discrete distribution, continuous parameter space

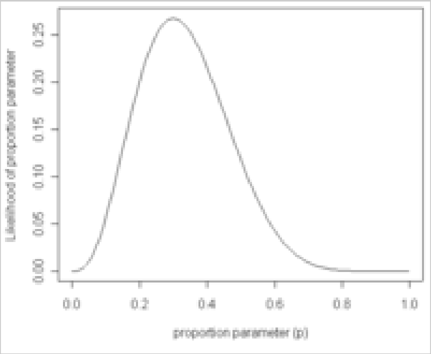
Now suppose our special box of coins from example 1 contains an infinite number of coins: one for every possible value $0 \leq p \leq 1$. We must maximise the likelihood function:

$$L(\theta) \quad = \quad f_D(\text{observe 49 HEADS out of 80} \mid p) = \binom{80}{49} p^{49}(1-p)^{31}$$

over all possible values $0 \leq p \leq 1$.

One may maximize this function by differentiating with respect to *p* and setting to zero:

$$\begin{aligned} 0 &= \frac{d}{dp}\left(\binom{80}{49}p^{49}(1-p)^{31}\right) \\ &\propto 49p^{48}(1-p)^{31} - 31p^{49}(1-p)^{30} \\ &= p^{48}(1-p)^{30}\left[49(1-p) - 31p\right] \end{aligned}$$



Likelihood of different proportion parameter values for a binomial process with *t* = 3 and *n* = 10; the ML estimator occurs at the mode with the peak (maximum) of the curve.

which has solutions *p* = 0, *p* = 1, and *p* = 49 / 80. The solution which maximises the likelihood is clearly *p* = 49 / 80 (since *p* = 0 and *p* = 1 result in a likelihood of zero). Thus we say the *maximum likelihood estimator* for *p* is $\widehat{p} = 49/80$.

This result is easily generalised by substituting a letter such as *t* in the place of 49 to represent the observed number of 'successes' of our Bernoulli trials, and a letter such as *n* in the place of 80 to represent the number of Bernoulli trials. Exactly the same calculation yields the *maximum likelihood estimator*:

$$\widehat{p} = \frac{t}{n}$$

for any sequence of *n* Bernoulli trials resulting in *t* 'successes'.

Continuous distribution, continuous parameter space

One of the most common continuous probability distributions is the normal distribution which has probability density function

$$f(x\mid\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\ \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The corresponding probability density function for a sample of *n* independent identically distributed normal random variables is

$$f(x_1,\ldots,x_n\mid\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right),$$

or more conveniently:

$$f(x_1,\ldots,x_n \mid \mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}\right).$$

This family of distributions has two parameters: μ,σ^2. This may be alarming to some, given that in the discussion above we only talked about maximising over a single parameter. However there is no need for alarm: we simply maximise the likelihood $L(\mu,\sigma) = f(x_1,\ldots,x_n \mid \mu,\sigma^2)$ over each parameter separately, which of course is more work but no more complicated. In the above notation we would write $\theta = (\mu,\sigma^2)$.

When maximising the likelihood, we may equivalently maximise the log of the likelihood, since log is a continuous strictly increasing function over the range of the likelihood. [Note: the log-likelihood is closely related to information entropy and Fisher information]. This often simplifies the algebra somewhat, and indeed does so in this case:

$$0 = \frac{\partial}{\partial\mu}\log\left(\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left(-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}\right)\right)$$

$$= \frac{\partial}{\partial\mu}\left(\log\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} - \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}\right)$$

$$= 0 - \frac{-2n(\bar{x}-\mu)}{2\sigma^2}$$

which is solved by $\widehat{\mu} = \bar{x} = \sum_{i=1}^{n} x_i/n$. This is indeed the maximum of the function since it is the only turning point in μ and the second derivative is strictly less than zero.

Similarly we differentiate with respect to σ and equate to zero:

$$0 = \frac{\partial}{\partial\sigma}\log\left(\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left(-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}\right)\right)$$

$$= \frac{\partial}{\partial\sigma}\left(\frac{n}{2}\log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}\right)$$

$$= -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{\sigma^3}$$

which is solved by $\widehat{\sigma}^2 = \sum_{i=1}^{n}(x_i - \widehat{\mu})^2/n$.

Contrast this with the *unbiased* sample variance estimator

$$s^2 = \sum_{i=1}^{n}(x_i - \widehat{\mu})^2/(n-1).$$

Formally we say that the *maximum likelihood estimator* for θ = (μ ,σ^2) is:

$$\widehat{\theta} = \left(\widehat{\mu}, \widehat{\sigma}^2\right) = \left(\bar{x}, \sum_{i=1}^{n}(x_i - \bar{x})^2/n\right).$$

Properties

Functional invariance

If $\widehat{\theta}$ is the maximum likelihood estimator (MLE) for θ, then the MLE for α = *g*(θ) is $\widehat{\alpha} = g(\widehat{\theta})$. The function *g* need not be one-to-one. For detail, please refer to the proof of Theorem 7.2.10 of *Statistical Inference* by George Casella and Roger L. Berger.

Asymptotic behaviour

Maximum likelihood estimators achieve minimum variance (as given by the Cramer-Rao lower bound) in the limit as the sample size tends to infinity. When the MLE is unbiased, we may equivalently say that it has minimum mean squared error in the limit.

For independent observations, the maximum likelihood estimator often follows an asymptotic normal distribution.

Bias

The bias of maximum-likelihood estimators can be substantial. Consider a case where *n* tickets numbered from 1 to *n* are placed in a box and one is selected at random (*see uniform distribution*). If *n* is unknown, then the maximum-likelihood estimator of *n* is the value on the drawn ticket, even though the expectation is only (*n* + 1) / 2. In estimating the highest number *n*, we can only be certain that it is greater than or equal to the drawn ticket number.

See also

- The mean squared error is a measure of how 'good' an estimator of a distributional parameter is (be it the maximum likelihood estimator or some other estimator).

- The article on the Rao-Blackwell theorem for a discussion on finding the best possible unbiased estimator (in the sense of having minimal mean squared error) by a process called Rao-Blackwellisation. The MLE is often a good starting place for the process.

- The reader may be intrigued to learn that the MLE (if it exists and is unique) will always be a function of a sufficient statistic for the parameter in question.

- Maximum likelihood estimation is related to generalized method of moments.

- See the article on inferential statistics for an alternative to the maximum likelihood estimate.

External resources

- A paper detailing the history of maximum likelihood, written by John Aldrich—

Retrieved from "http://en.wikipedia.org/wiki/Maximum_likelihood"

Category: Estimation theory

Learning with Maximum Likelihood

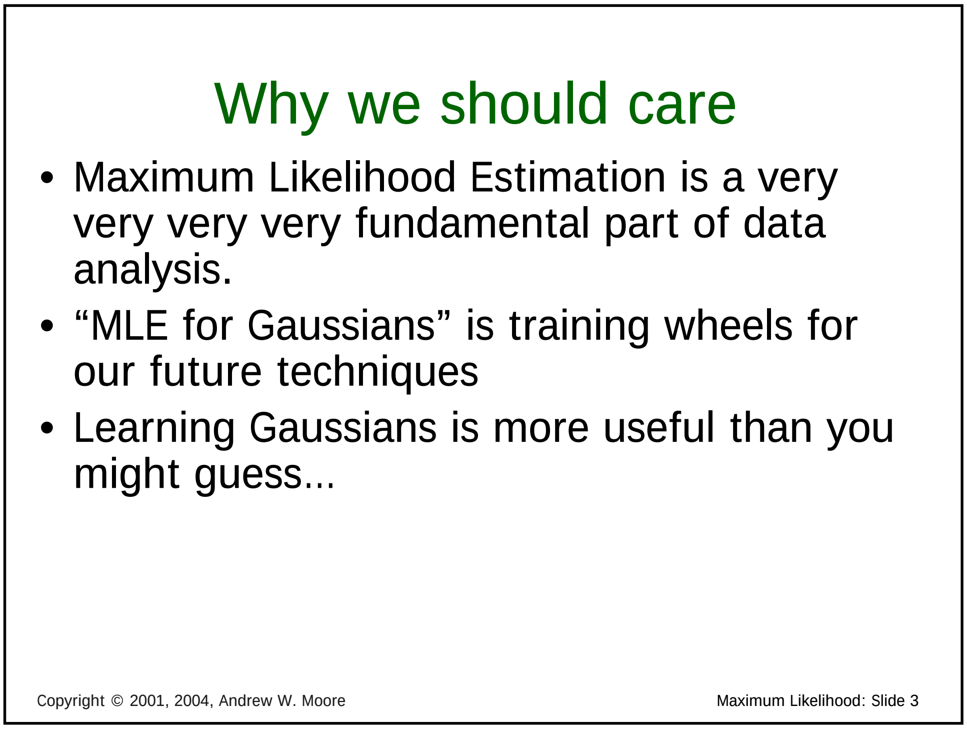
Andrew W. Moore
Professor
School of Computer Science
Carnegie Mellon University
www.cs.cmu.edu/~awm
awm@cs.cmu.edu
412-268-7599

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew's tutorials: http://www.cs.cmu.edu/~awm/tutorials . Comments and corrections gratefully received.

Sep 6th, 2001

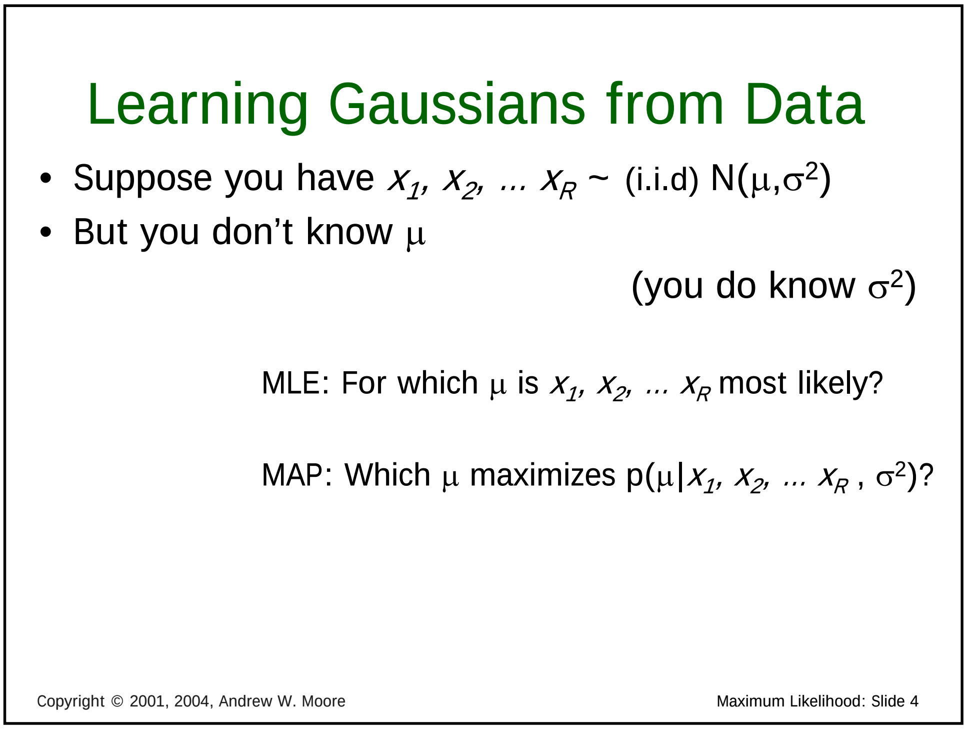
Maximum Likelihood learning of Gaussians for Data Mining

- Why we should care
- Learning Univariate Gaussians
- Learning Multivariate Gaussians
- What's a biased estimator?
- Bayesian Learning of Gaussians



Why we should care

- Maximum Likelihood Estimation is a very very very very fundamental part of data analysis.
- “MLE for Gaussians” is training wheels for our future techniques
- Learning Gaussians is more useful than you might guess…



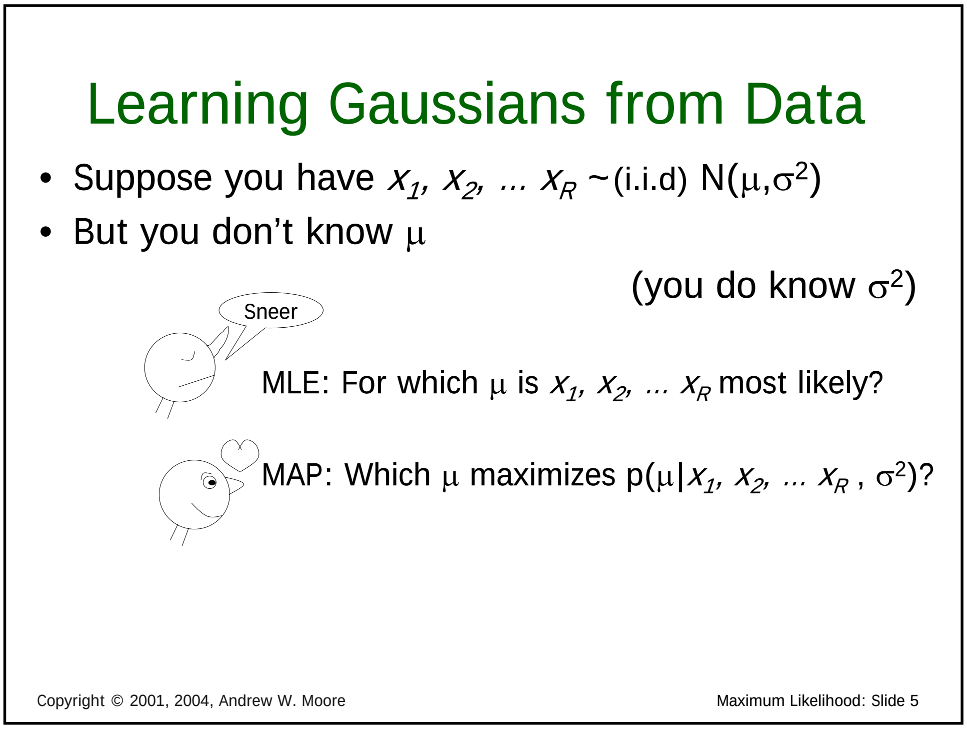
Learning Gaussians from Data

- Suppose you have $x_1, x_2, \ldots x_R$ ~ (i.i.d) N(μ,σ^2)
- But you don’t know μ

(you do know σ^2)

MLE: For which μ is $x_1, x_2, \ldots x_R$ most likely?

MAP: Which μ maximizes p(μ|$x_1, x_2, \ldots x_R$, σ^2)?



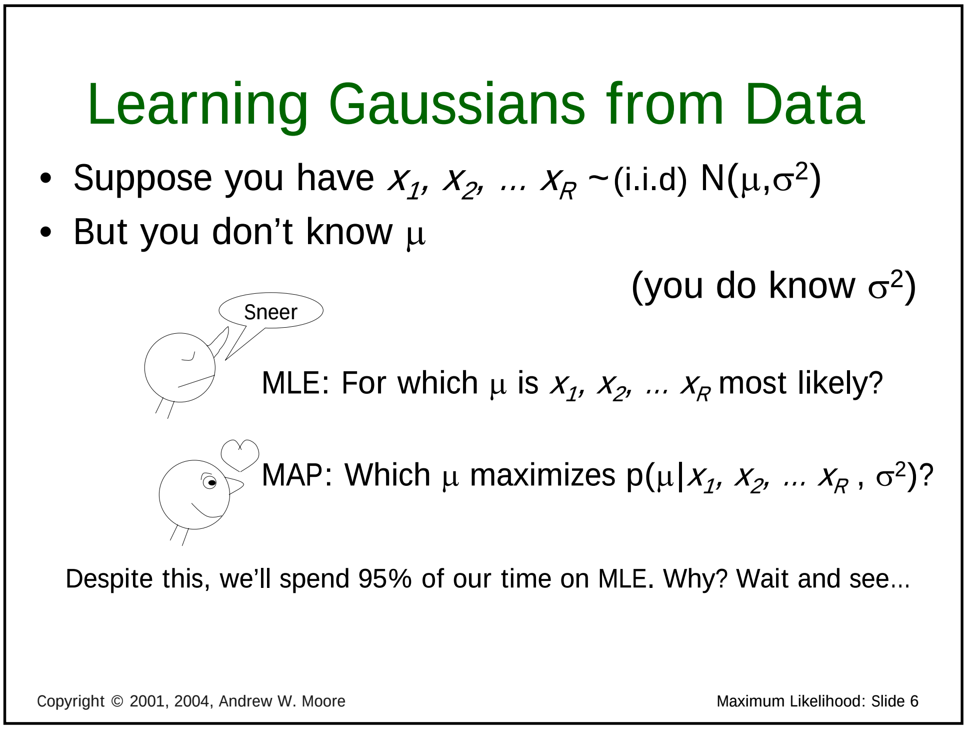
Learning Gaussians from Data

- Suppose you have $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2)
- But you don't know μ

(you do know σ^2)

MLE: For which μ is $x_1, x_2, \ldots x_R$ most likely?

MAP: Which μ maximizes p($\mu|x_1, x_2, \ldots x_R$, σ^2)?



Learning Gaussians from Data

- Suppose you have $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2)
- But you don't know μ

(you do know σ^2)

MLE: For which μ is $x_1, x_2, \ldots x_R$ most likely?

MAP: Which μ maximizes p($\mu|x_1, x_2, \ldots x_R$, σ^2)?

Despite this, we'll spend 95% of our time on MLE. Why? Wait and see…

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2)
- But you don't know μ (you do know σ^2)
- MLE: For which μ is $x_1, x_2, \ldots x_R$ most likely?

$$\mu^{mle} = \arg\max_{\mu} \; p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

Algebra Euphoria

$$\mu^{mle} = \arg\max_{\mu} \; p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

= (by i.i.d)

= (monotonicity of log)

= (plug in formula for Gaussian)

= (after simplification)

Algebra Euphoria

$$\mu^{mle} = \underset{\mu}{\arg\max}\ p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

$$= \underset{\mu}{\arg\max} \prod_{i=1}^{R} p(x_i \mid \mu, \sigma^2)$$

(by i.i.d)

$$= \underset{\mu}{\arg\max} \sum_{i=1}^{R} \log p(x_i \mid \mu, \sigma^2)$$

(monotonicity of log)

$$= \underset{\mu}{\arg\max} \frac{1}{\sqrt{2\pi}\ \sigma} \sum_{i=1}^{R} - \frac{(x_i - \mu)^2}{2\sigma^2}$$

(plug in formula for Gaussian)

$$= \underset{\mu}{\arg\min} \sum_{i=1}^{R} (x_i - \mu)^2$$

(after simplification)

Intermission: A General Scalar MLE strategy

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out ∂LL/∂θ using high-school calculus
3. Set ∂LL/∂θ=0 for a maximum, creating an equation in terms of θ
4. Solve it*
5. Check that you've found a maximum rather than a minimum or saddle-point, and be careful if θ is constrained

*This is a perfect example of something that works perfectly in all textbook examples and usually involves surprising pain if you need it for something new.

The MLE μ

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

$$= \arg\min_{\mu} \sum_{i=1}^{R} (x_i - \mu)^2$$

$$= \mu \ \text{ s.t. } \quad 0 = \frac{\partial \text{LL}}{\partial \mu} =$$

= (what?)

The MLE μ

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

$$= \arg\min_{\mu} \sum_{i=1}^{R} (x_i - \mu)^2$$

$$= \mu \ \text{ s.t. } \quad 0 = \frac{\partial \text{LL}}{\partial \mu} = \frac{\partial}{\partial \mu} \sum_{i=1}^{R} (x_i - \mu)^2$$

$$- \sum_{i=1}^{R} 2(x_i - \mu)$$

$$\text{Thus} \quad \mu = \frac{1}{R} \sum_{i=1}^{R} x_i$$

Lawks-a-lawdy!

$$\mu^{mle} = \frac{1}{R}\sum_{i=1}^{R} x_i$$

- The best estimate of the mean of a distribution is the mean of the sample!

At first sight:
This kind of pedantic, algebra-filled and ultimately unsurprising fact is exactly the reason people throw down their "Statistics" book and pick up their "Agent Based Evolutionary Data Mining Using The Neuro-Fuzz Transform" book.

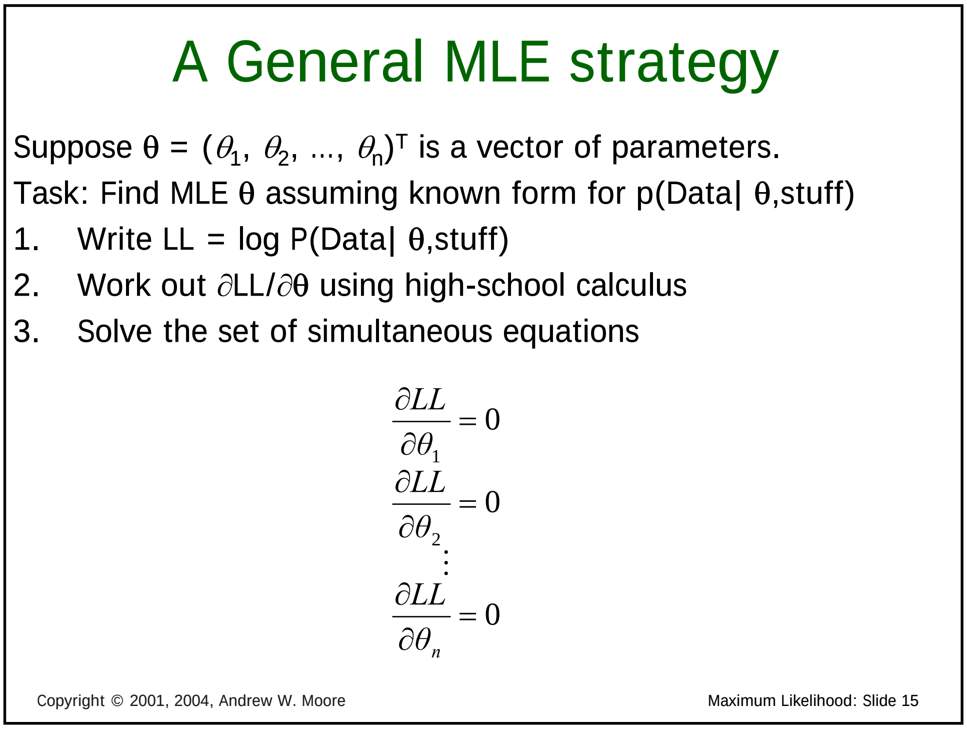
A General MLE strategy

Suppose $\theta = (\theta_1, \theta_2, ..., \theta_n)^T$ is a vector of parameters.

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out ∂LL/∂θ using high-school calculus

$$\frac{\partial LL}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial LL}{\partial \theta_1} \\ \frac{\partial LL}{\partial \theta_2} \\ \vdots \\ \frac{\partial LL}{\partial \theta_n} \end{pmatrix}$$



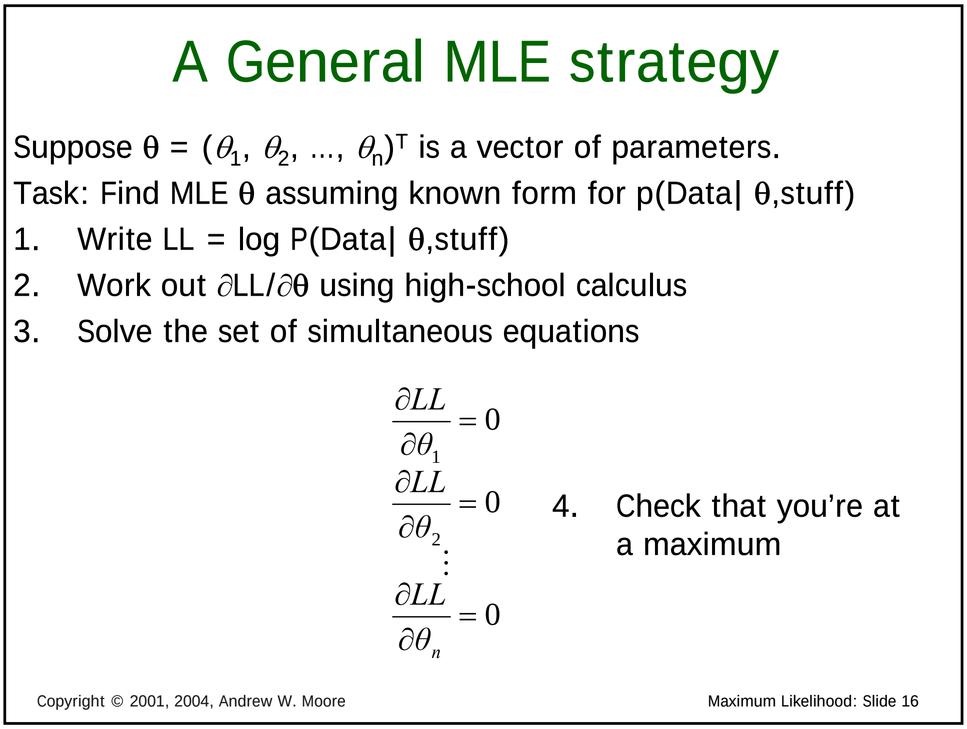
A General MLE strategy

Suppose θ = (θ_1, θ_2, ..., θ_n)^T is a vector of parameters.

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out ∂LL/∂θ using high-school calculus
3. Solve the set of simultaneous equations

$$\frac{\partial LL}{\partial \theta_1} = 0$$
$$\frac{\partial LL}{\partial \theta_2} = 0$$
$$\vdots$$
$$\frac{\partial LL}{\partial \theta_n} = 0$$



A General MLE strategy

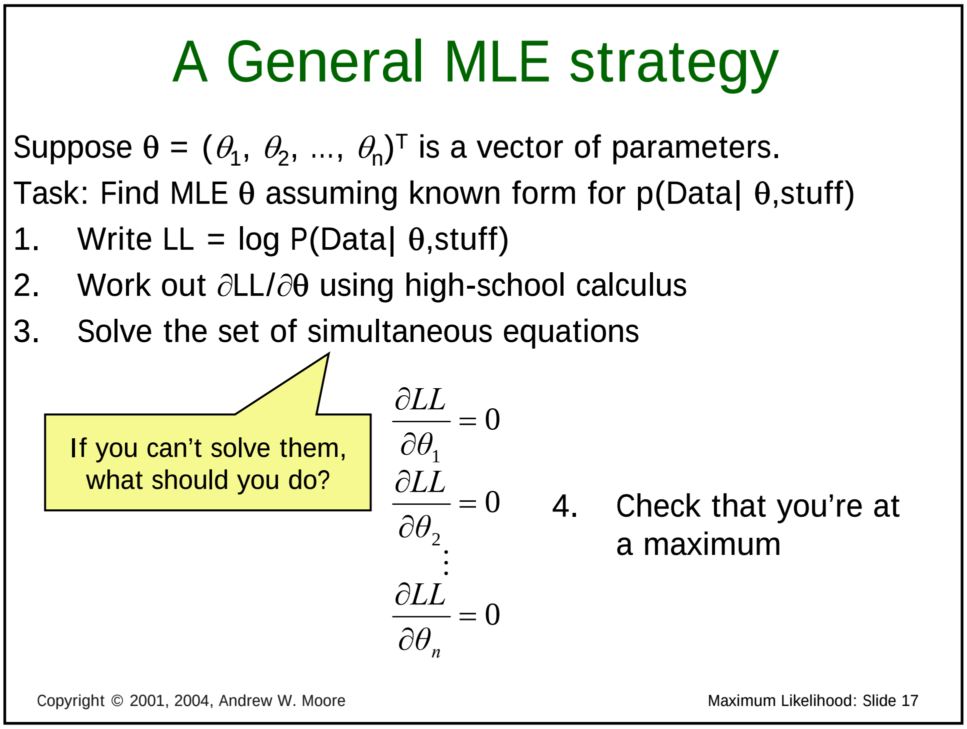
Suppose θ = (θ_1, θ_2, ..., θ_n)^T is a vector of parameters.

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out ∂LL/∂θ using high-school calculus
3. Solve the set of simultaneous equations

$$\frac{\partial LL}{\partial \theta_1} = 0$$
$$\frac{\partial LL}{\partial \theta_2} = 0$$
$$\vdots$$
$$\frac{\partial LL}{\partial \theta_n} = 0$$

4. Check that you’re at a maximum



A General MLE strategy

Suppose θ = (θ_1, θ_2, …, θ_n)^T is a vector of parameters.

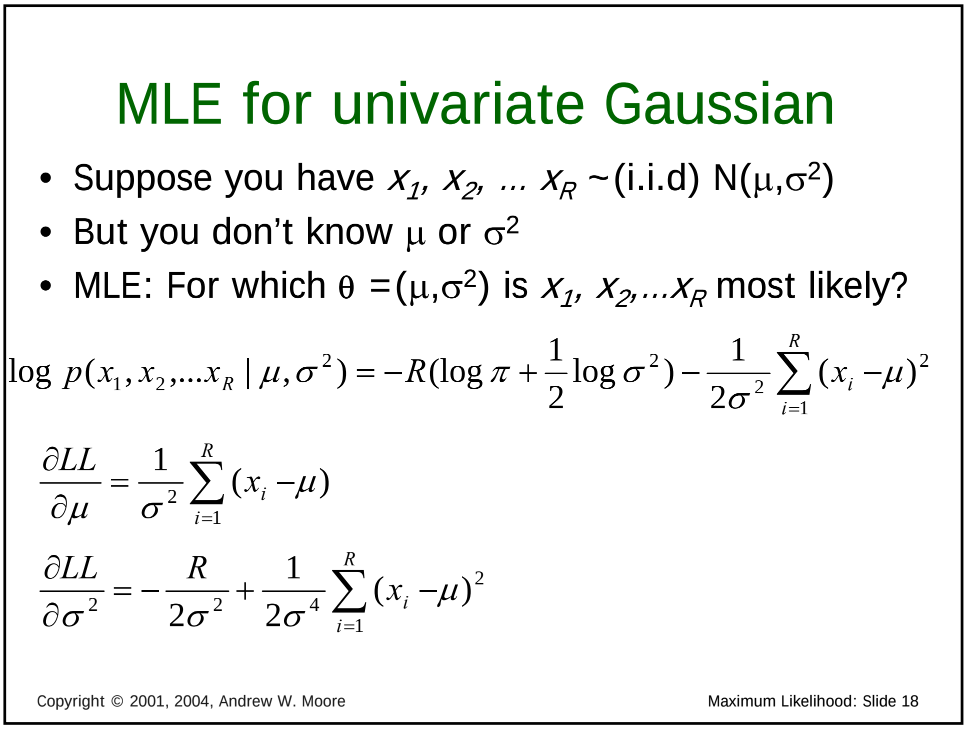
Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out ∂LL/∂θ using high-school calculus
3. Solve the set of simultaneous equations

If you can't solve them, what should you do?

$$\frac{\partial LL}{\partial \theta_1} = 0$$
$$\frac{\partial LL}{\partial \theta_2} = 0$$
$$\vdots$$
$$\frac{\partial LL}{\partial \theta_n} = 0$$

4. Check that you're at a maximum



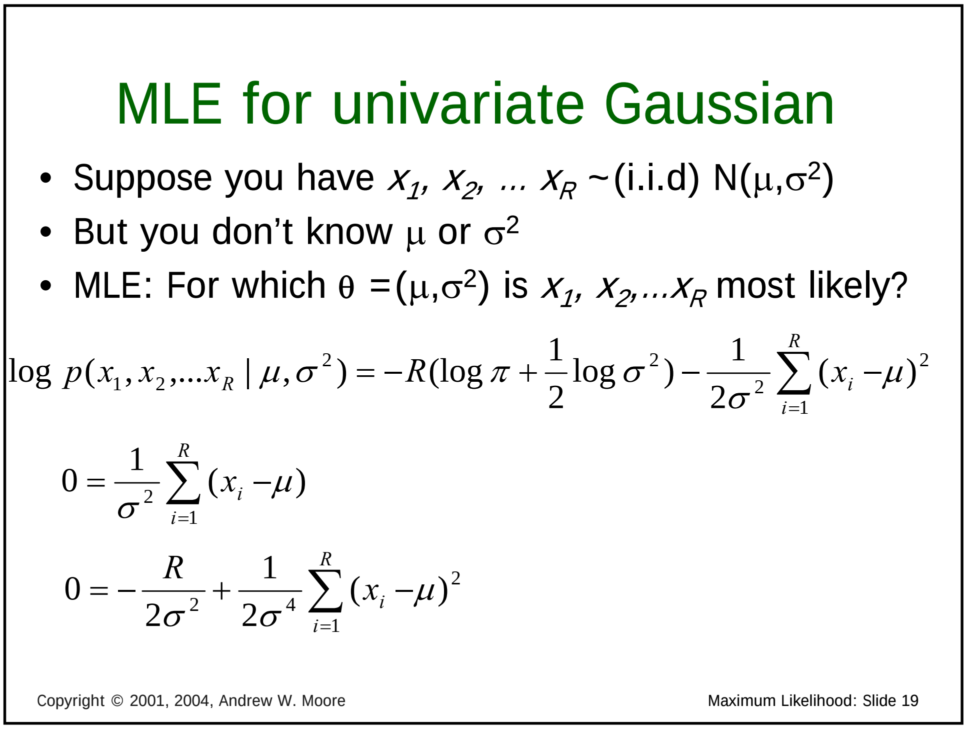
MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2)
- But you don't know μ or σ^2
- MLE: For which θ =(μ,σ^2) is $x_1, x_2, \ldots x_R$ most likely?

$$\log\, p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2) = -R(\log \pi + \frac{1}{2}\log \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{R}(x_i - \mu)^2$$

$$\frac{\partial LL}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{R}(x_i - \mu)$$

$$\frac{\partial LL}{\partial \sigma^2} = -\frac{R}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{R}(x_i - \mu)^2$$



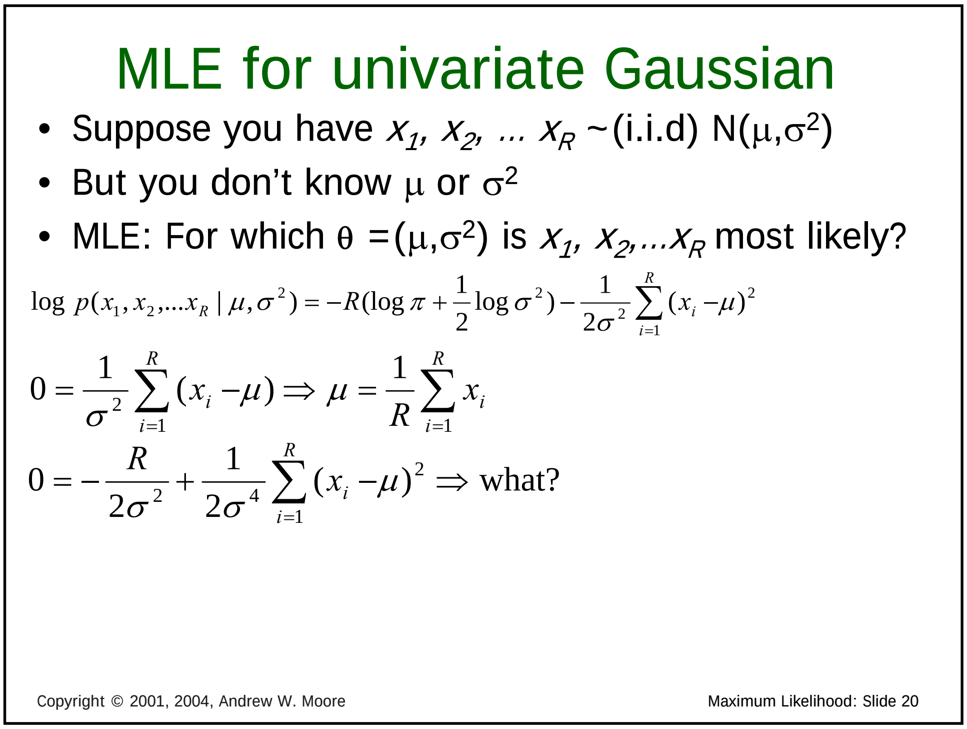
MLE for univariate Gaussian

- Suppose you have *x₁, x₂, ... x_R* ~(i.i.d) N(μ,σ²)
- But you don't know μ or σ²
- MLE: For which θ =(μ,σ²) is *x₁, x₂,...x_R* most likely?

$$\log\ p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2) = -R(\log \pi + \frac{1}{2}\log \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{R}(x_i - \mu)^2$$

$$0 = \frac{1}{\sigma^2}\sum_{i=1}^{R}(x_i - \mu)$$

$$0 = -\frac{R}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{R}(x_i - \mu)^2$$



MLE for univariate Gaussian

- Suppose you have *x₁, x₂, ... x_R* ~(i.i.d) N(μ,σ²)
- But you don't know μ or σ²
- MLE: For which θ =(μ,σ²) is *x₁, x₂,...x_R* most likely?

$$\log\ p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2) = -R(\log \pi + \frac{1}{2}\log \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{R}(x_i - \mu)^2$$

$$0 = \frac{1}{\sigma^2}\sum_{i=1}^{R}(x_i - \mu) \Rightarrow \mu = \frac{1}{R}\sum_{i=1}^{R} x_i$$

$$0 = -\frac{R}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{R}(x_i - \mu)^2 \Rightarrow \text{what?}$$

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2)
- But you don't know μ or σ^2
- MLE: For which θ =(μ,σ^2) is $x_1, x_2, \ldots x_R$ most likely?

$$\mu^{mle} = \frac{1}{R}\sum_{i=1}^{R} x_i$$

$$\sigma^2_{mle} = \frac{1}{R}\sum_{i=1}^{R} (x_i - \mu^{mle})^2$$

Unbiased Estimators

- An estimator of a parameter is unbiased if the expected value of the estimate is the same as the true value of the parameters.
- *If* $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2) then

$$E[\mu^{mle}] = E\left[\frac{1}{R}\sum_{i=1}^{R} x_i\right] = \mu$$

μ^{mle} is unbiased

Biased Estimators

- An estimator of a parameter is biased if the expected value of the estimate is different from the true value of the parameters.
- *If* $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2) then

$$E\left[\sigma^2_{mle}\right] = E\left[\frac{1}{R}\sum_{i=1}^{R}(x_i - \mu^{mle})^2\right] = E\left[\frac{1}{R}\left(\sum_{i=1}^{R} x_i - \frac{1}{R}\sum_{j=1}^{R} x_j\right)^2\right] \neq \sigma^2$$

σ^2_{mle} is biased

MLE Variance Bias

- *If* $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2) then

$$E\left[\sigma^2_{mle}\right] = E\left[\frac{1}{R}\left(\sum_{i=1}^{R} x_i - \frac{1}{R}\sum_{j=1}^{R} x_j\right)^2\right] = \left(1 - \frac{1}{R}\right)\sigma^2 \neq \sigma^2$$

Intuition check: consider the case of R=1

Why should our guts expect that σ^2_{mle} would be an underestimate of true σ^2?

How could you prove that?

Unbiased estimate of Variance

- *If x_1, x_2, … x_R* ~(i.i.d) N(μ,σ^2) then

$$E\left[\sigma^2_{mle}\right] = E\left[\frac{1}{R}\left(\sum_{i=1}^{R} x_i - \frac{1}{R}\sum_{j=1}^{R} x_j\right)^2\right] = \left(1 - \frac{1}{R}\right)\sigma^2 \neq \sigma^2$$

So define $\sigma^2_{\text{unbiased}} = \dfrac{\sigma^2_{mle}}{\left(1 - \dfrac{1}{R}\right)}$ So $E\left[\sigma^2_{\text{unbiased}}\right] = \sigma^2$

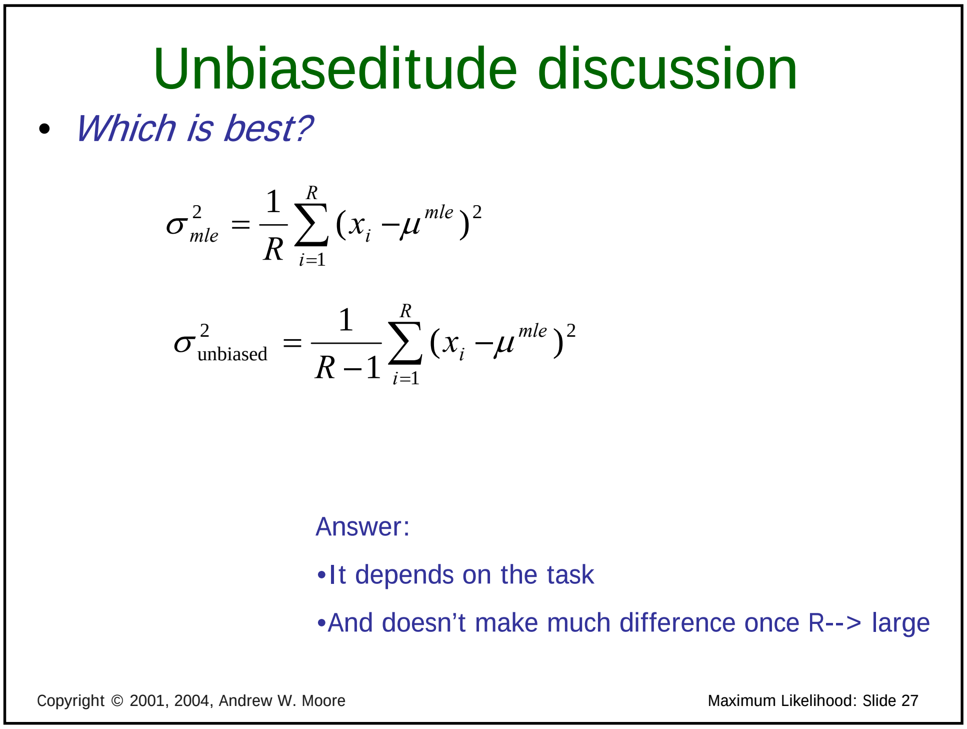
Unbiased estimate of Variance

- *If x_1, x_2, … x_R* ~(i.i.d) N(μ,σ^2) then

$$E\left[\sigma^2_{mle}\right] = E\left[\frac{1}{R}\left(\sum_{i=1}^{R} x_i - \frac{1}{R}\sum_{j=1}^{R} x_j\right)^2\right] = \left(1 - \frac{1}{R}\right)\sigma^2 \neq \sigma^2$$

So define $\sigma^2_{\text{unbiased}} = \dfrac{\sigma^2_{mle}}{\left(1 - \dfrac{1}{R}\right)}$ So $E\left[\sigma^2_{\text{unbiased}}\right] = \sigma^2$

$$\sigma^2_{\text{unbiased}} = \frac{1}{R-1}\sum_{i=1}^{R}(x_i - \mu^{mle})^2$$



Unbiaseditude discussion

- *Which is best?*

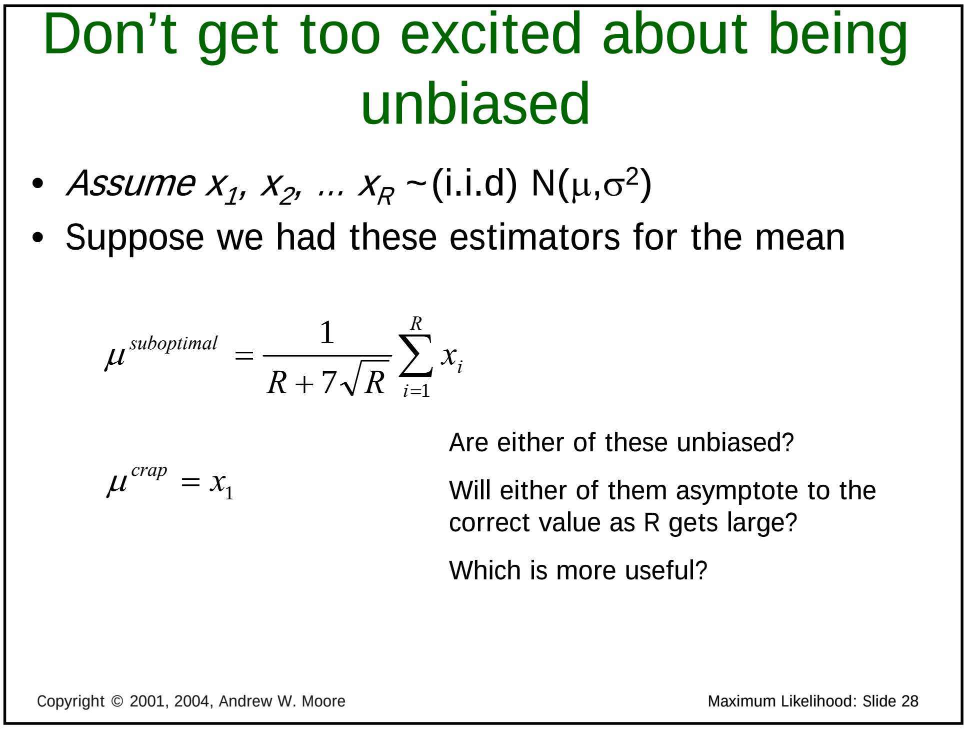
$$\sigma^2_{mle} = \frac{1}{R}\sum_{i=1}^{R}(x_i - \mu^{mle})^2$$

$$\sigma^2_{\text{unbiased}} = \frac{1}{R-1}\sum_{i=1}^{R}(x_i - \mu^{mle})^2$$

Answer:

•It depends on the task

•And doesn't make much difference once R--> large



Don't get too excited about being unbiased

- *Assume x_1, x_2, … x_R* ~(i.i.d) N(μ,σ²)
- Suppose we had these estimators for the mean

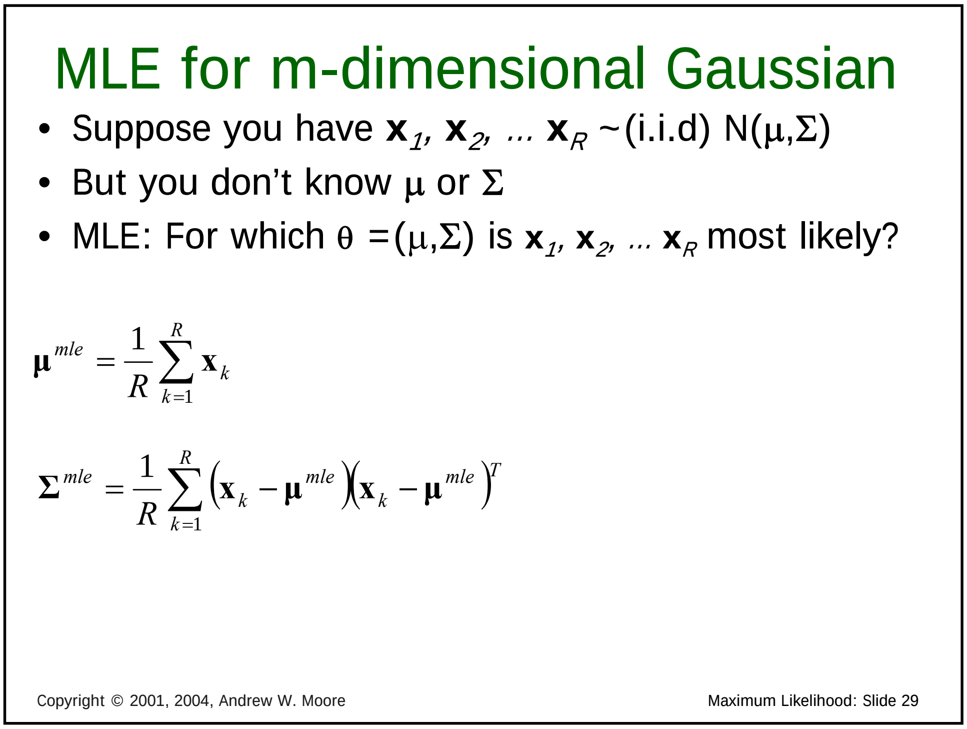
$$\mu^{suboptimal} = \frac{1}{R+7\sqrt{R}}\sum_{i=1}^{R}x_i$$

$$\mu^{crap} = x_1$$

Are either of these unbiased?

Will either of them asymptote to the correct value as R gets large?

Which is more useful?

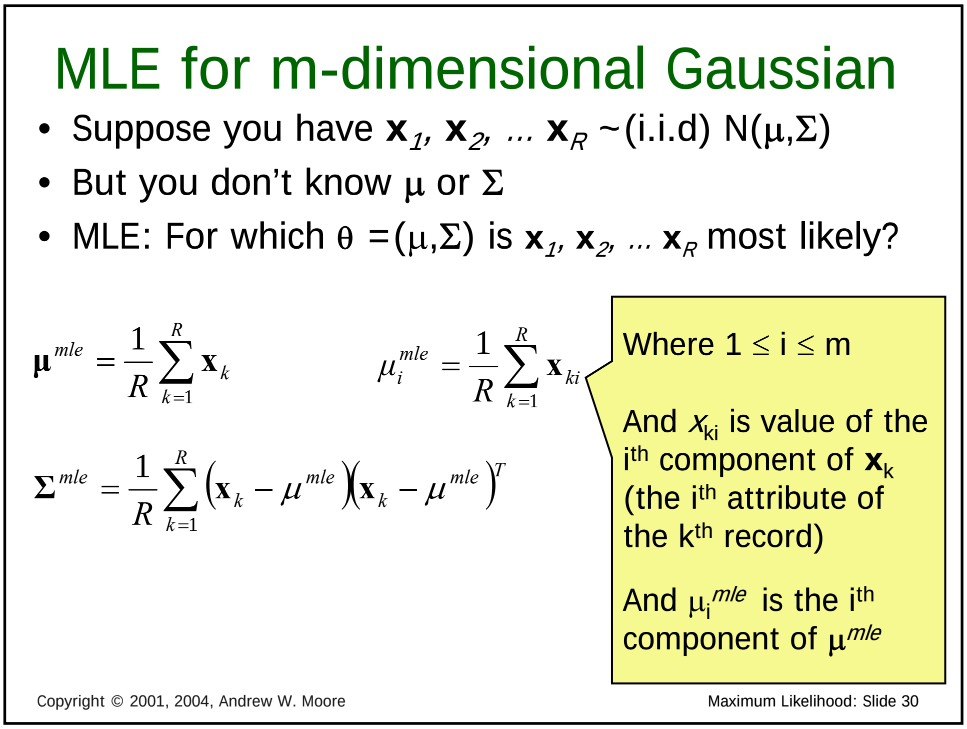


MLE for m-dimensional Gaussian

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MLE: For which θ =(μ,Σ) is $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ most likely?

$$\boldsymbol{\mu}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_k$$

$$\boldsymbol{\Sigma}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\left(\mathbf{x}_k - \boldsymbol{\mu}^{mle}\right)\left(\mathbf{x}_k - \boldsymbol{\mu}^{mle}\right)^T$$



MLE for m-dimensional Gaussian

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MLE: For which θ =(μ,Σ) is $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ most likely?

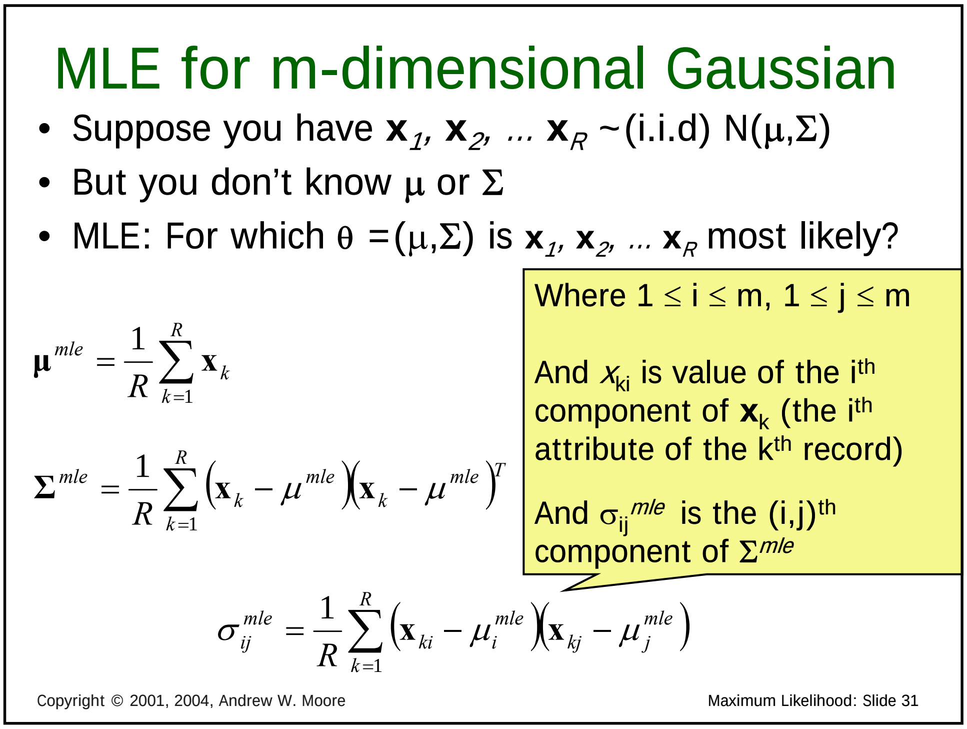
$$\boldsymbol{\mu}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_k \qquad \mu_i^{mle} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_{ki}$$

$$\boldsymbol{\Sigma}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\left(\mathbf{x}_k - \mu^{mle}\right)\left(\mathbf{x}_k - \mu^{mle}\right)^T$$

Where $1 \le i \le m$

And x_{ki} is value of the i^{th} component of $\mathbf{x}_k$ (the i^{th} attribute of the k^{th} record)

And μ_i^{mle} is the i^{th} component of μ^{mle}



MLE for m-dimensional Gaussian

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MLE: For which θ =(μ,Σ) is $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ most likely?

$$\boldsymbol{\mu}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_k$$

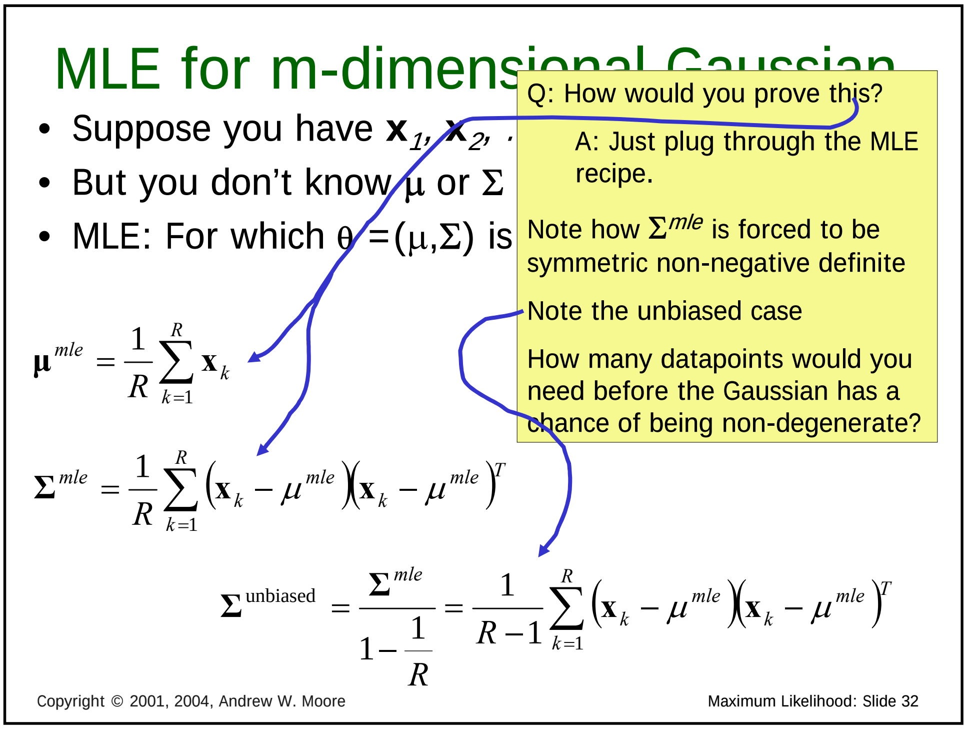
$$\boldsymbol{\Sigma}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\left(\mathbf{x}_k - \mu^{mle}\right)\left(\mathbf{x}_k - \mu^{mle}\right)^T$$

Where 1 ≤ i ≤ m, 1 ≤ j ≤ m

And x_{ki} is value of the ith component of $\mathbf{x}_k$ (the ith attribute of the kth record)

And σ_{ij}^{mle} is the (i,j)th component of Σ^{mle}

$$\sigma_{ij}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\left(\mathbf{x}_{ki} - \mu_i^{mle}\right)\left(\mathbf{x}_{kj} - \mu_j^{mle}\right)$$



MLE for m-dimensional Gaussian

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2$, .
- But you don't know μ or Σ
- MLE: For which θ =(μ,Σ) is

Q: How would you prove this?

A: Just plug through the MLE recipe.

Note how Σ^{mle} is forced to be symmetric non-negative definite

Note the unbiased case

How many datapoints would you need before the Gaussian has a chance of being non-degenerate?

$$\boldsymbol{\mu}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_k$$

$$\boldsymbol{\Sigma}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\left(\mathbf{x}_k - \mu^{mle}\right)\left(\mathbf{x}_k - \mu^{mle}\right)^T$$

$$\boldsymbol{\Sigma}^{\text{unbiased}} = \frac{\boldsymbol{\Sigma}^{mle}}{1 - \frac{1}{R}} = \frac{1}{R-1}\sum_{k=1}^{R}\left(\mathbf{x}_k - \mu^{mle}\right)\left(\mathbf{x}_k - \mu^{mle}\right)^T$$

Confidence intervals

We need to talk

We need to discuss how accurate we expect μ^{mle} and Σ^{mle} to be as a function of R

And we need to consider how to estimate these accuracies from data…

•Analytically *

•Non-parametrically (using randomization and bootstrapping) *

But we won't. Not yet.

*Will be discussed in future Andrew lectures…just before we need this technology.

Structural error

Actually, we need to talk about something else too..

What if we do all this analysis when the true distribution is in fact not Gaussian?

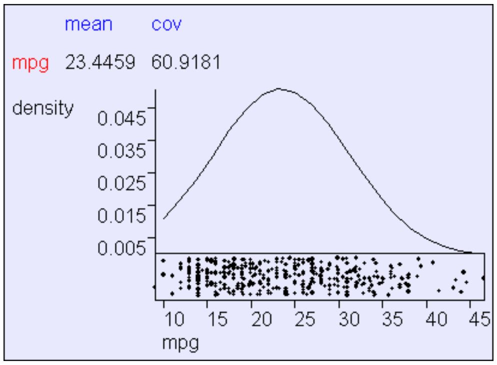
How can we tell? *

How can we survive? *

*Will be discussed in future Andrew lectures…just before we need this technology.

Gaussian MLE in action

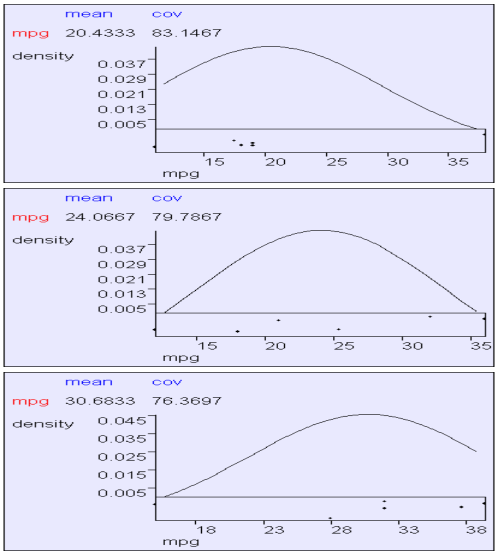
Using R=392 cars from the "MPG" UCI dataset supplied by Ross Quinlan



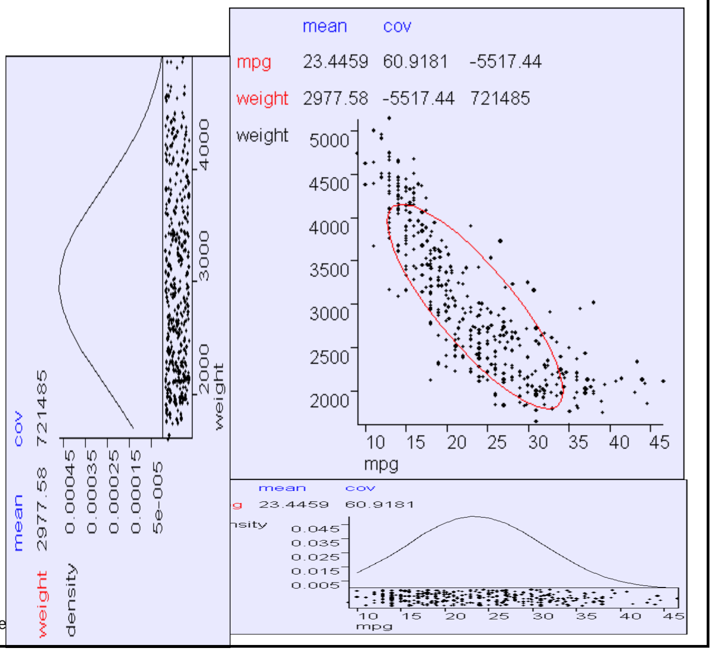
Data-starved Gaussian MLE

Using three subsets of MPG.

Each subset has 6 randomly-chosen cars.



Bivariate MLE in action

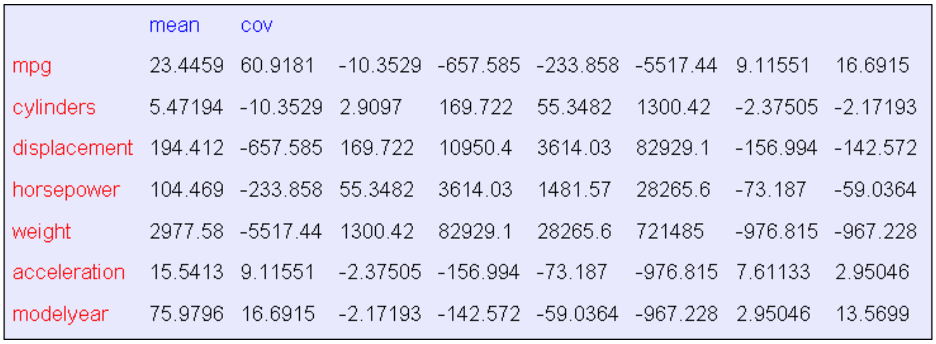


	mean	cov	
mpg	23.4459	60.9181	-5517.44
weight	2977.58	-5517.44	721485

	mean	cov
weight	2977.58	721485

	mean	cov
g	23.4459	60.9181

Multivariate MLE

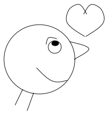


	mean	cov						
mpg	23.4459	60.9181	-10.3529	-657.585	-233.858	-5517.44	9.11551	16.6915
cylinders	5.47194	-10.3529	2.9097	169.722	55.3482	1300.42	-2.37505	-2.17193
displacement	194.412	-657.585	169.722	10950.4	3614.03	82929.1	-156.994	-142.572
horsepower	104.469	-233.858	55.3482	3614.03	1481.57	28265.6	-73.187	-59.0364
weight	2977.58	-5517.44	1300.42	82929.1	28265.6	721485	-976.815	-967.228
acceleration	15.5413	9.11551	-2.37505	-156.994	-73.187	-976.815	7.61133	2.95046
modelyear	75.9796	16.6915	-2.17193	-142.572	-59.0364	-967.228	2.95046	13.5699

Covariance matrices are not exciting to look at

Being Bayesian: MAP estimates for Gaussians

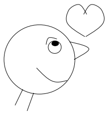
- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_R$ ~(i.i.d) $N(\mu,\Sigma)$
- But you don't know μ or Σ
- MAP: Which (μ,Σ) maximizes $p(\mu,\Sigma \mid \mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_R)$?



Step 1: Put a prior on (μ,Σ)

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_R$ ~(i.i.d) $N(\mu,\Sigma)$
- But you don't know μ or Σ
- MAP: Which (μ,Σ) maximizes $p(\mu,\Sigma \mid \mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_R)$?



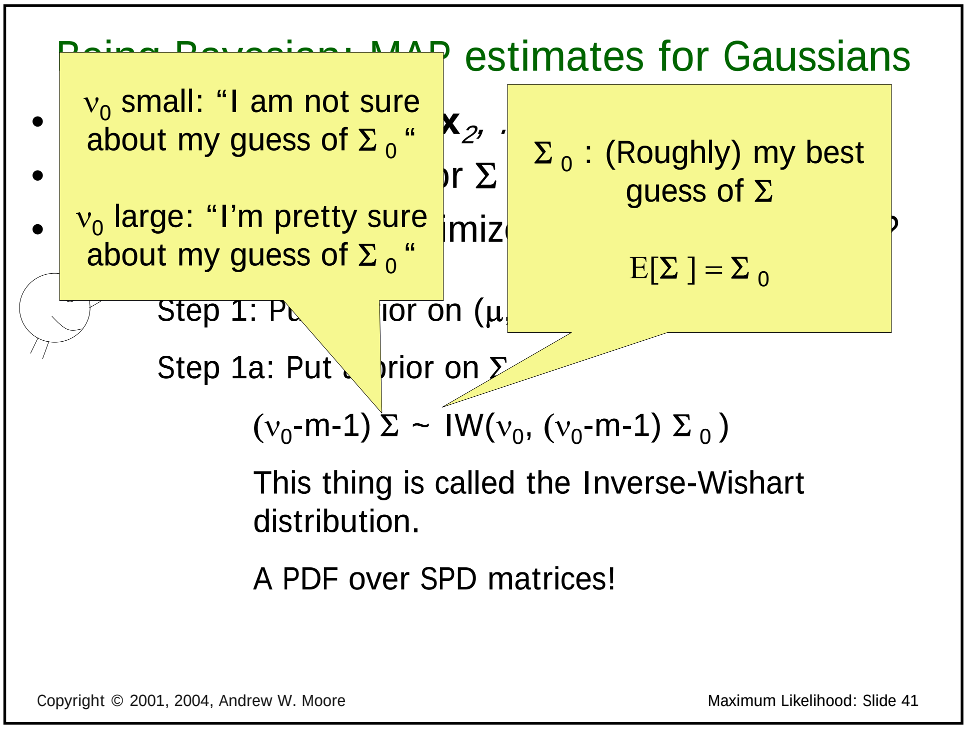
Step 1: Put a prior on (μ,Σ)

Step 1a: Put a prior on Σ

$(\nu_0 - m - 1)\,\Sigma \sim IW(\nu_0, (\nu_0 - m - 1)\,\Sigma_0)$

This thing is called the Inverse-Wishart distribution.

A PDF over SPD matrices!



Being Bayesian: MAP estimates for Gaussians

- $\mathbf{x}_2$, ..
- or Σ
- imiz ?

ν_0 small: “I am not sure about my guess of Σ_0 “

ν_0 large: “I’m pretty sure about my guess of Σ_0 “

Σ_0 : (Roughly) my best guess of Σ

$E[\Sigma] = \Sigma_0$

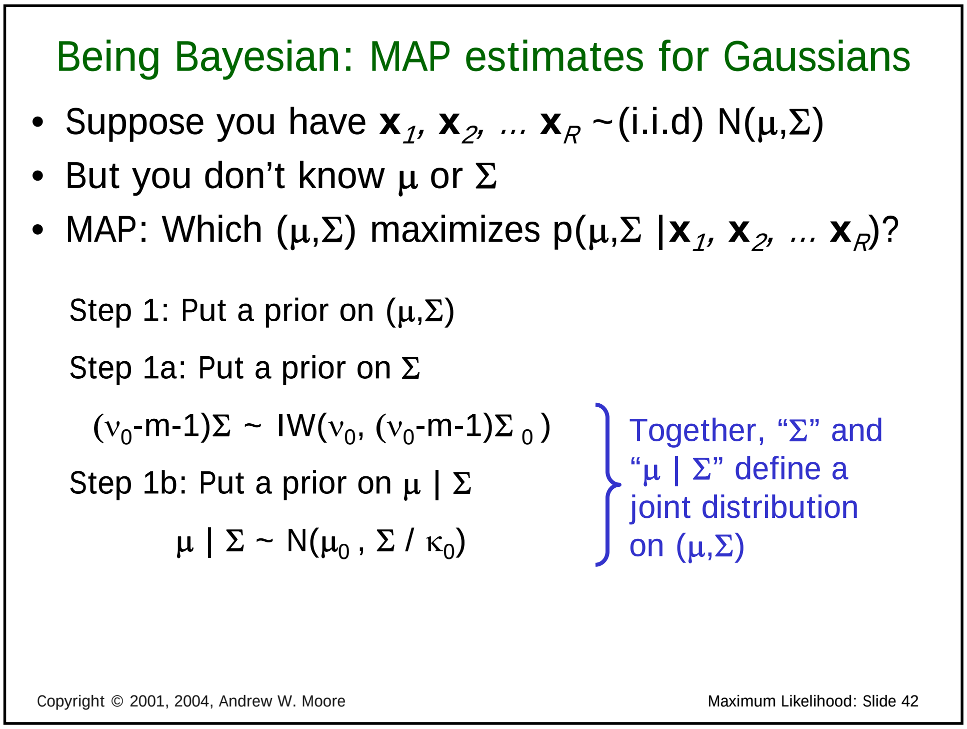
Step 1: Pu ior on $(\mu,$

Step 1a: Put prior on Σ

$(\nu_0\text{-m-1})\ \Sigma \sim \text{IW}(\nu_0, (\nu_0\text{-m-1})\ \Sigma_0)$

This thing is called the Inverse-Wishart distribution.

A PDF over SPD matrices!



Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) $N(\mu,\Sigma)$
- But you don’t know μ or Σ
- MAP: Which (μ,Σ) maximizes $p(\mu,\Sigma \mid \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R)$?

Step 1: Put a prior on (μ,Σ)

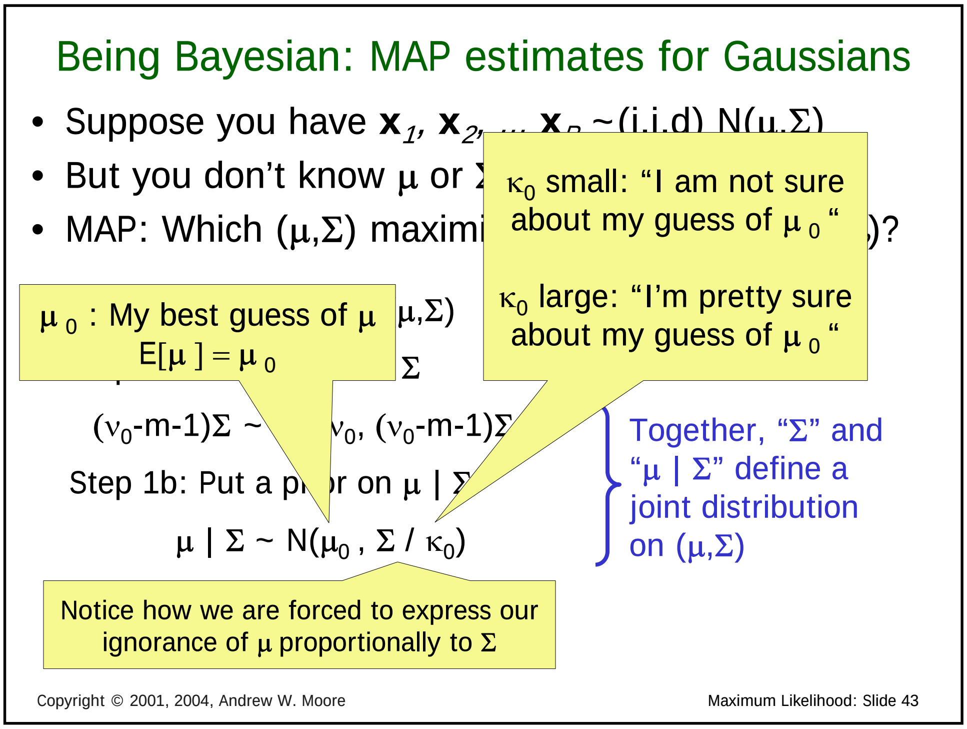
Step 1a: Put a prior on Σ

$(\nu_0\text{-m-1})\Sigma \sim \text{IW}(\nu_0, (\nu_0\text{-m-1})\Sigma_0)$

Step 1b: Put a prior on $\mu \mid \Sigma$

$\mu \mid \Sigma \sim N(\mu_0 , \Sigma / \kappa_0)$

Together, “Σ” and “$\mu \mid \Sigma$” define a joint distribution on (μ,Σ)



Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1$, $\mathbf{x}_2$, ... $\mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MAP: Which (μ,Σ) maximi)?

κ_0 small: "I am not sure about my guess of μ_0 "

κ_0 large: "I'm pretty sure about my guess of μ_0 "

μ_0 : My best guess of μ
E[μ] = μ_0

μ,Σ)

Σ

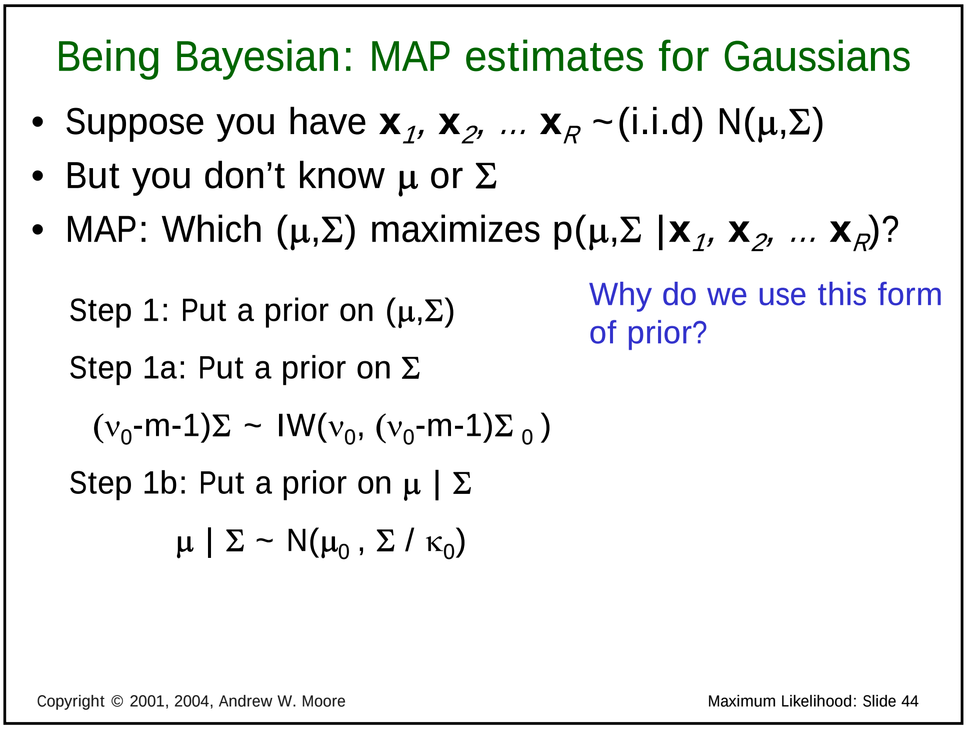
$(\nu_0\text{-m-1})\Sigma$ ~ ν_0, $(\nu_0\text{-m-1})\Sigma$

Step 1b: Put a pr on μ | Σ

$\mu \mid \Sigma \sim N(\mu_0 , \Sigma / \kappa_0)$

Together, "Σ" and "μ | Σ" define a joint distribution on (μ,Σ)

Notice how we are forced to express our ignorance of μ proportionally to Σ



Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1$, $\mathbf{x}_2$, ... $\mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MAP: Which (μ,Σ) maximizes p(μ,Σ |$\mathbf{x}_1$, $\mathbf{x}_2$, ... $\mathbf{x}_R$)?

Step 1: Put a prior on (μ,Σ)

Step 1a: Put a prior on Σ

$(\nu_0\text{-m-1})\Sigma \sim IW(\nu_0, (\nu_0\text{-m-1})\Sigma_0)$

Step 1b: Put a prior on μ | Σ

$\mu \mid \Sigma \sim N(\mu_0 , \Sigma / \kappa_0)$

Why do we use this form of prior?

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MAP: Which (μ,Σ) maximizes p(μ,Σ |$\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$)?

Step 1: Put a prior on (μ,Σ)

Step 1a: Put a prior on Σ

$(\nu_0\text{-m-1})\Sigma \sim \text{IW}(\nu_0, (\nu_0\text{-m-1})\Sigma_0)$

Step 1b: Put a prior on μ | Σ

$\mu \mid \Sigma \sim N(\mu_0, \Sigma / \kappa_0)$

Why do we use this form of prior?

Actually, we don't have to

But it is computationally and algebraically convenient…

…it's a *conjugate prior.*

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- MAP: Which (μ,Σ) maximizes p(μ,Σ |$\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$)?

Step 1: Prior: $(\nu_0\text{-m-1})\,\Sigma \sim \text{IW}(\nu_0, (\nu_0\text{-m-1})\,\Sigma_0)$, $\mu \mid \Sigma \sim N(\mu_0, \Sigma / \kappa_0)$

Step 2:

$$\boxed{\bar{\mathbf{x}} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_k} \quad \boxed{\boldsymbol{\mu}_R = \frac{\kappa_0\boldsymbol{\mu}_0 + R\bar{\mathbf{x}}}{\kappa_0 + R}} \quad \boxed{\nu_R = \nu_0 + R} \quad \boxed{\kappa_R = \kappa_0 + R}$$

$$(\nu_R + m - 1)\boldsymbol{\Sigma}_R = (\nu_0 + m - 1)\boldsymbol{\Sigma}_0 + \sum_{k=1}^{R}(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T + \frac{(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T}{1/\kappa_0 + 1/R}$$

Step 3: Posterior: $(\nu_R\text{+m-1})\Sigma \sim \text{IW}(\nu_R, (\nu_R\text{+m-1})\,\Sigma_R)$,

$\mu \mid \Sigma \sim N(\mu_R, \Sigma / \kappa_R)$

Result: $\mu^{map} = \mu_R$, E[Σ |$\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$]= Σ_R

Being Bayesian:

- Suppose you hav
- MAP: Which (μ,Σ)

Step 1: Prior: $(\nu_0\text{-m-1})\ \Sigma$ ~

- Look carefully at what these formulae are doing. It's all very sensible.
- Conjugate priors mean prior form and posterior form are same and characterized by "sufficient statistics" of the data.
- The marginal distribution on μ is a student-t
- One point of view: it's pretty academic if R > 30

Step 2:

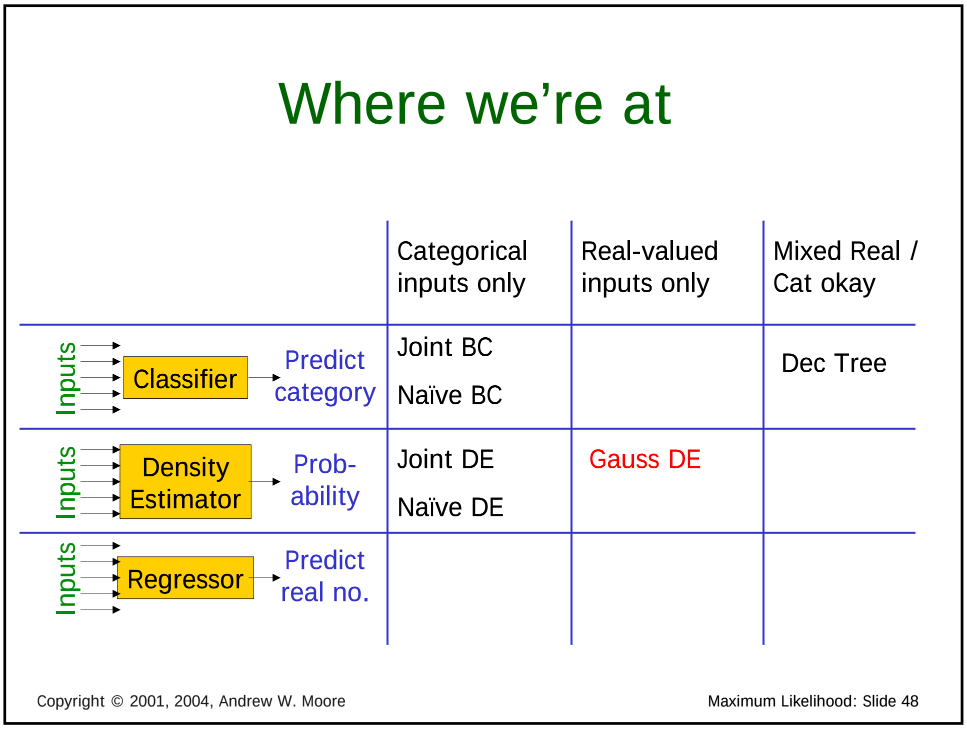
$$\overline{\mathbf{x}} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_k \qquad \boldsymbol{\mu}_R = \frac{\kappa_0\boldsymbol{\mu}_0 + R\overline{\mathbf{x}}}{\kappa_0 + R} \qquad \nu_R = \nu_0 + R \qquad \kappa_R = \kappa_0 + R$$

$$(\nu_R + m - 1)\boldsymbol{\Sigma}_R = (\nu_0 + m - 1)\boldsymbol{\Sigma}_0 + \sum_{k=1}^{R}(\mathbf{x}_k - \overline{\mathbf{x}})(\mathbf{x}_k - \overline{\mathbf{x}})^T + \frac{(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)^T}{1/\kappa_0 + 1/R}$$

Step 3: Posterior: $(\nu_R\text{+m-1})\Sigma$ ~ $\text{IW}(\nu_R, (\nu_R\text{+m-1})\ \Sigma_R)$,

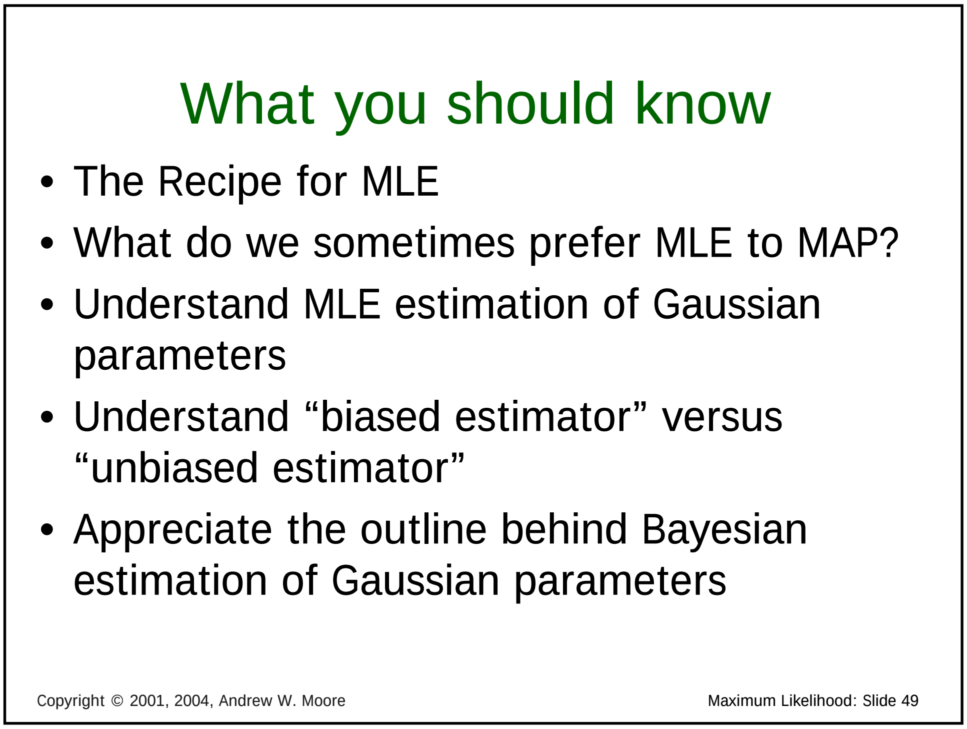
$\mu \mid \Sigma$ ~ $N(\mu_R, \Sigma / \kappa_R)$

Result: $\mu^{map} = \mu_R$, $E[\Sigma \mid \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R] = \Sigma_R$



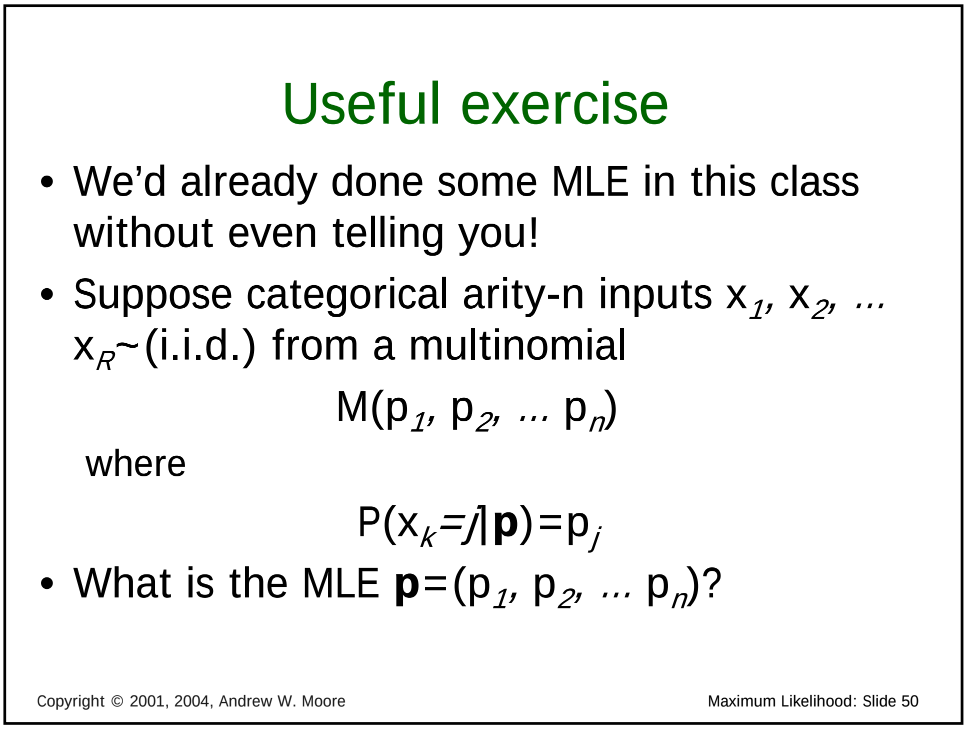
Where we're at

	Categorical inputs only	Real-valued inputs only	Mixed Real / Cat okay
Inputs → Classifier → Predict category	Joint BC Naïve BC		Dec Tree
Inputs → Density Estimator → Prob-ability	Joint DE Naïve DE	Gauss DE	
Inputs → Regressor → Predict real no.			



What you should know

- The Recipe for MLE
- What do we sometimes prefer MLE to MAP?
- Understand MLE estimation of Gaussian parameters
- Understand “biased estimator” versus “unbiased estimator”
- Appreciate the outline behind Bayesian estimation of Gaussian parameters



Useful exercise

- We’d already done some MLE in this class without even telling you!
- Suppose categorical arity-n inputs $x_1, x_2, \ldots x_R$~(i.i.d.) from a multinomial

$$M(p_1, p_2, \ldots p_n)$$

where

$$P(x_k = j | \mathbf{p}) = p_j$$

- What is the MLE $\mathbf{p}=(p_1, p_2, \ldots p_n)$?

ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Mathematical Psychology 47 (2003) 90–100

Journal of Mathematical Psychology

http://www.elsevier.com/locate/jmp

Tutorial

Tutorial on maximum likelihood estimation

In Jae Myung*

Department of Psychology, Ohio State University, 1885 Neil Avenue Mall, Columbus, OH 43210-1222, USA

Received 30 November 2001; revised 16 October 2002

Abstract

In this paper, I provide a tutorial exposition on maximum likelihood estimation (MLE). The intended audience of this tutorial are researchers who practice mathematical modeling of cognition but are unfamiliar with the estimation method. Unlike least-squares estimation which is primarily a descriptive tool, MLE is a preferred method of parameter estimation in statistics and is an indispensable tool for many statistical modeling techniques, in particular in non-linear modeling with non-normal data. The purpose of this paper is to provide a good conceptual explanation of the method with illustrative examples so the reader can have a grasp of some of the basic principles.

1. Introduction

In psychological science, we seek to uncover general laws and principles that govern the behavior under investigation. As these laws and principles are not directly observable, they are formulated in terms of hypotheses. In mathematical modeling, such hypotheses about the structure and inner working of the behavioral process of interest are stated in terms of parametric families of probability distributions called models. The goal of modeling is to deduce the form of the underlying process by testing the viability of such models.

Once a model is specified with its parameters, and data have been collected, one is in a position to evaluate its goodness of fit, that is, how well it fits the observed data. Goodness of fit is assessed by finding parameter values of a model that best fits the data—a procedure called *parameter estimation*.

There are two general methods of parameter estimation. They are least-squares estimation (LSE) and maximum likelihood estimation (MLE). The former has been a popular choice of model fitting in psychology (e.g., Rubin, Hinton, & Wenzel, 1999; Lamberts, 2000 but see Usher & McClelland, 2001) and is tied to many familiar statistical concepts such as linear regression, sum of squares error, proportion variance accounted for (i.e. r^2), and root mean squared deviation. LSE, which unlike MLE requires no or minimal distributional assumptions, is useful for obtaining a descriptive measure for the purpose of summarizing observed data, but it has no basis for testing hypotheses or constructing confidence intervals.

On the other hand, MLE is not as widely recognized among modelers in psychology, but it is a standard approach to parameter estimation and inference in statistics. MLE has many optimal properties in estimation: sufficiency (complete information about the parameter of interest contained in its MLE estimator); consistency (true parameter value that generated the data recovered asymptotically, i.e. for data of sufficiently large samples); efficiency (lowest-possible variance of parameter estimates achieved asymptotically); and parameterization invariance (same MLE solution obtained independent of the parametrization used). In contrast, no such things can be said about LSE. As such, most statisticians would not view LSE as a general method for parameter estimation, but rather as an approach that is primarily used with linear regression models. Further, many of the inference methods in statistics are developed based on MLE. For example, MLE is a prerequisite for the chi-square test, the G-square test, Bayesian methods, inference with missing data, modeling of random effects, and many model selection criteria such as the Akaike information criterion (Akaike, 1973) and the Bayesian information criteria (Schwarz, 1978).

*Fax: +614-292-5601.

E-mail address: myung.1@osu.edu.

0022-2496/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.

doi:10.1016/S0022-2496(02)00028-7

In this tutorial paper, I introduce the maximum likelihood estimation method for mathematical modeling. The paper is written for researchers who are primarily involved in empirical work and publish in experimental journals (e.g. *Journal of Experimental Psychology*) but do modeling. The paper is intended to serve as a stepping stone for the modeler to move beyond the current practice of using LSE to more informed modeling analyses, thereby expanding his or her repertoire of statistical instruments, especially in non-linear modeling. The purpose of the paper is to provide a good conceptual understanding of the method with concrete examples. For in-depth, technically more rigorous treatment of the topic, the reader is directed to other sources (e.g., Bickel & Doksum, 1977, Chap. 3; Casella & Berger, 2002, Chap. 7; DeGroot & Schervish, 2002, Chap. 6; Spanos, 1999, Chap. 13).

2. Model specification

2.1. Probability density function

From a statistical standpoint, the data vector $y = (y_1, \ldots, y_m)$ is a random sample from an unknown population. The goal of data analysis is to identify the population that is most likely to have generated the sample. In statistics, each population is identified by a corresponding probability distribution. Associated with each probability distribution is a unique value of the model's parameter. As the parameter changes in value, different probability distributions are generated. Formally, a model is defined as the family of probability distributions indexed by the model's parameters.

Let $f(y|w)$ denote the *probability density function* (PDF) that specifies the probability of observing data vector y given the parameter w. Throughout this paper we will use a plain letter for a vector (e.g. y) and a letter with a subscript for a vector element (e.g. y_i). The parameter $w = (w_1, \ldots, w_k)$ is a vector defined on a multi-dimensional parameter space. If individual observations, y_i's, are statistically independent of one another, then according to the theory of probability, the PDF for the data $y = (y_1, \ldots, y_m)$ given the parameter vector w can be expressed as a multiplication of PDFs for individual observations,

$$f(y = (y_1, y_2, \ldots, y_n) \mid w) = f_1(y_1 \mid w) f_2(y_2 \mid w) \cdots f_n(y_m \mid w). \quad (1)$$

To illustrate the idea of a PDF, consider the simplest case with one observation and one parameter, that is, $m = k = 1$. Suppose that the data y represents the number of successes in a sequence of 10 Bernoulli trials (e.g. tossing a coin 10 times) and that the probability of a success on any one trial, represented by the parameter w, is 0.2. The PDF in this case is given by

$$f(y \mid n = 10, w = 0.2) = \frac{10!}{y!(10-y)!}(0.2)^y(0.8)^{10-y} \quad (y = 0, 1, \ldots, 10) \quad (2)$$

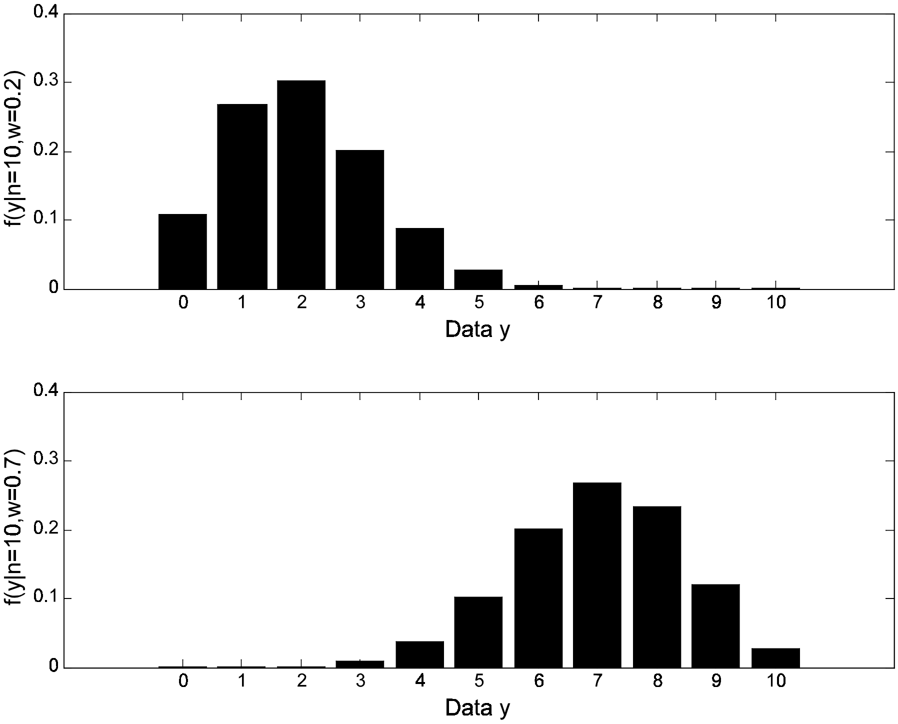


Fig. 1. Binomial probability distributions of sample size $n = 10$ and probability parameter $w = 0.2$ (top) and $w = 0.7$ (bottom).

which is known as the binomial distribution with parameters $n = 10$, $w = 0.2$. Note that the number of trials (n) is considered as a parameter. The shape of this PDF is shown in the top panel of Fig. 1. If the parameter value is changed to say $w = 0.7$, a new PDF is obtained as

$$f(y \mid n = 10, w = 0.7) = \frac{10!}{y!(10-y)!}(0.7)^y(0.3)^{10-y} \quad (y = 0, 1, \ldots, 10) \tag{3}$$

whose shape is shown in the bottom panel of Fig. 1. The following is the general expression of the PDF of the binomial distribution for arbitrary values of w and n:

$$f(y|n, w) = \frac{n!}{y!(n-y)!}w^y(1-w)^{n-y} \quad (0 \leqslant w \leqslant 1;\ \ y = 0, 1, \ldots, n) \tag{4}$$

which as a function of y specifies the probability of data y for a given value of n and w. The collection of all such PDFs generated by varying the parameter across its range (0–1 in this case for w, $n \geqslant 1$) defines a model.

2.2. *Likelihood function*

Given a set of parameter values, the corresponding PDF will show that some data are more probable than other data. In the previous example, the PDF with $w = 0.2$, $y = 2$ is more likely to occur than $y = 5$ (0.302 vs. 0.026). In reality, however, we have already observed the data. Accordingly, we are faced with an inverse problem: Given the observed data and a model of interest, find the one PDF, among all the probability densities that the model prescribes, that is most likely to have produced the data. To solve this inverse problem, we define the *likelihood function* by reversing the roles of the data vector y and the parameter vector w in $f(y|w)$, i.e.

$$L(w|y) = f(y|w). \tag{5}$$

Thus $L(w|y)$ represents the likelihood of the parameter w given the observed data y, and as such is a function of w. For the one-parameter binomial example in Eq. (4), the likelihood function for $y = 7$ and $n = 10$ is given by

$$L(w \mid n = 10, y = 7) = f(y = 7 \mid n = 10, w) = \frac{10!}{7!3!}w^7(1-w)^3 \quad (0 \leqslant w \leqslant 1). \tag{6}$$

The shape of this likelihood function is shown in Fig. 2.

There exist an important difference between the PDF $f(y|w)$ and the likelihood function $L(w|y)$. As illustrated in Figs. 1 and 2, the two functions are defined on different axes, and therefore are not directly comparable to each other. Specifically, the PDF in Fig. 1 is a function of the data given a particular set of parameter values, defined on the *data scale*. On the other hand, the likelihood function is a function of the parameter given a particular set of observed data, defined on the *parameter scale*. In short, Fig. 1 tells us the probability of a particular data value for a fixed parameter, whereas Fig. 2 tells us the likelihood ("unnormalized probability") of a particular parameter value for a fixed data set. Note that the likelihood function in this figure is a curve

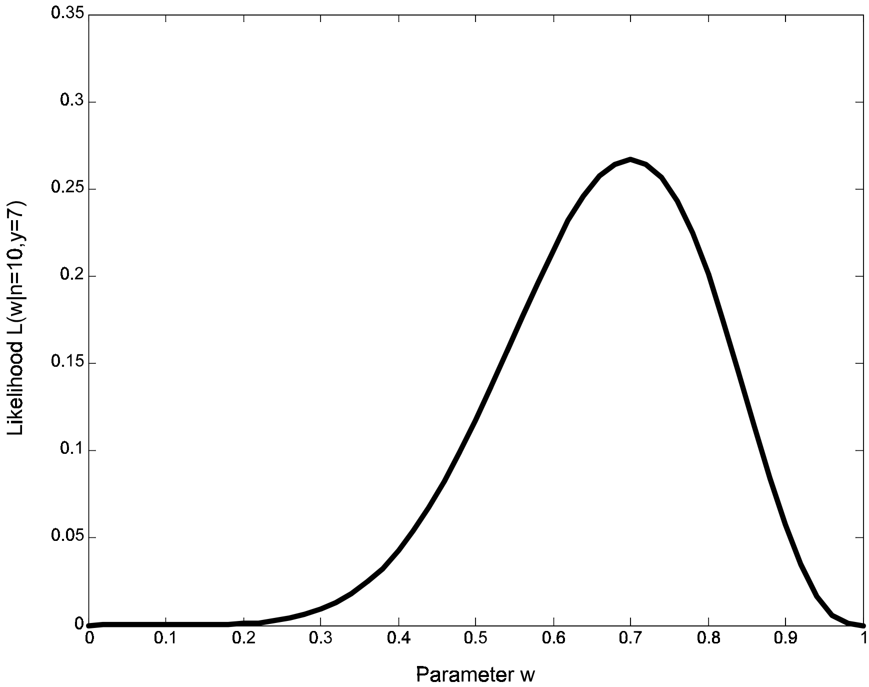


Fig. 2. The likelihood function given observed data $y = 7$ and sample size $n = 10$ for the one-parameter model described in the text.

because there is only one parameter beside n, which is assumed to be known. If the model has two parameters, the likelihood function will be a surface sitting above the parameter space. In general, for a model with k parameters, the likelihood function $L(w|y)$ takes the shape of a k-dim geometrical "surface" sitting above a k-dim hyperplane spanned by the parameter vector $w = (w_1, \ldots, w_k)$.

3. Maximum likelihood estimation

Once data have been collected and the likelihood function of a model given the data is determined, one is in a position to make statistical inferences about the population, that is, the probability distribution that underlies the data. Given that different parameter values index different probability distributions (Fig. 1), we are interested in finding the parameter value that corresponds to the desired probability distribution.

The principle of *maximum likelihood estimation* (MLE), originally developed by R.A. Fisher in the 1920s, states that the desired probability distribution is the one that makes the observed data "most likely," which means that one must seek the value of the parameter vector that maximizes the likelihood function $L(w|y)$. The resulting parameter vector, which is sought by searching the multi-dimensional parameter space, is called the *MLE estimate*, and is denoted by $w_{\text{MLE}} = (w_{1,\text{MLE}}, \ldots, w_{k,\text{MLE}})$. For example, in Fig. 2, the MLE estimate is $w_{\text{MLE}} = 0.7$ for which the maximized likelihood value is $L(w_{\text{MLE}} = 0.7|n = 10, y = 7) = 0.267$. The probability distribution corresponding to this MLE estimate is shown in the bottom panel of Fig. 1. According to the MLE principle, this is the population that is most likely to have generated the observed data of $y = 7$. To summarize, maximum likelihood estimation is a method to seek the probability distribution that makes the observed data most likely.

3.1. Likelihood equation

MLE estimates need not exist nor be unique. In this section, we show how to compute MLE estimates when they exist and are unique. For computational convenience, the MLE estimate is obtained by maximizing the log-likelihood function, $\ln L(w|y)$. This is because the two functions, $\ln L(w|y)$ and $L(w|y)$, are monotonically related to each other so the same MLE estimate is obtained by maximizing either one. Assuming that the log-likelihood function, $\ln L(w|y)$, is differentiable, if w_{MLE} exists, it must satisfy the following partial differential equation known as the *likelihood equation*:

$$\frac{\partial \ln L(w|y)}{\partial w_i} = 0 \tag{7}$$

at $w_i = w_{i,\text{MLE}}$ for all $i = 1, \ldots, k$. This is because the definition of maximum or minimum of a continuous differentiable function implies that its first derivatives vanish at such points.

The likelihood equation represents a necessary condition for the existence of an MLE estimate. An additional condition must also be satisfied to ensure that $\ln L(w|y)$ is a maximum and not a minimum, since the first derivative cannot reveal this. To be a maximum, the shape of the log-likelihood function should be convex (it must represent a peak, not a valley) in the neighborhood of w_{MLE}. This can be checked by calculating the second derivatives of the log-likelihoods and showing whether they are all negative at $w_i = w_{i,\text{MLE}}$ for $i = 1, \ldots, k$,[1]

$$\frac{\partial^2 \ln L(w|y)}{\partial w_i^2} < 0. \tag{8}$$

To illustrate the MLE procedure, let us again consider the previous one-parameter binomial example given a fixed value of n. First, by taking the logarithm of the likelihood function $L(w|n = 10, y = 7)$ in Eq. (6), we obtain the log-likelihood as

$$\ln L(w \mid n = 10, y = 7) = \ln \frac{10!}{7!3!} + 7 \ln w + 3 \ln(1 - w) \tag{9}$$

Next, the first derivative of the log-likelihood is calculated as

$$\frac{d \ln L(w \mid n = 10, y = 7)}{dw} = \frac{7}{w} - \frac{3}{1 - w} = \frac{7 - 10w}{w(1 - w)}. \tag{10}$$

By requiring this equation to be zero, the desired MLE estimate is obtained as $w_{\text{MLE}} = 0.7$. To make sure that the solution represents a maximum, not a minimum, the second derivative of the log-likelihood is calculated and evaluated at $w = w_{\text{MLE}}$,

$$\begin{aligned} \frac{d^2 \ln L(w \mid n = 10, y = 7)}{dw^2} &= -\frac{7}{w^2} - \frac{3}{(1 - w)^2} \\ &= -47.62 < 0 \end{aligned} \tag{11}$$

which is negative, as desired.

In practice, however, it is usually not possible to obtain an analytic form solution for the MLE estimate, especially when the model involves many parameters and its PDF is highly non-linear. In such situations, the MLE estimate must be sought numerically using nonlinear optimization algorithms. The basic idea of nonlinear optimization is to quickly find optimal parameters that maximize the log-likelihood. This is done by

[1] Consider the Hessian matrix $H(w)$ defined as $H_{ij}(w) = \frac{\partial^2 \ln L(w)}{\partial w_i \partial w_j}$ $(i, j = 1, \ldots, k)$. Then a more accurate test of the convexity condition requires that the determinant of $H(w)$ be *negative definite*, that is, $z' H(w = w_{\text{MLE}}) z < 0$ for any $kx1$ real-numbered vector z, where z' denotes the transpose of z.

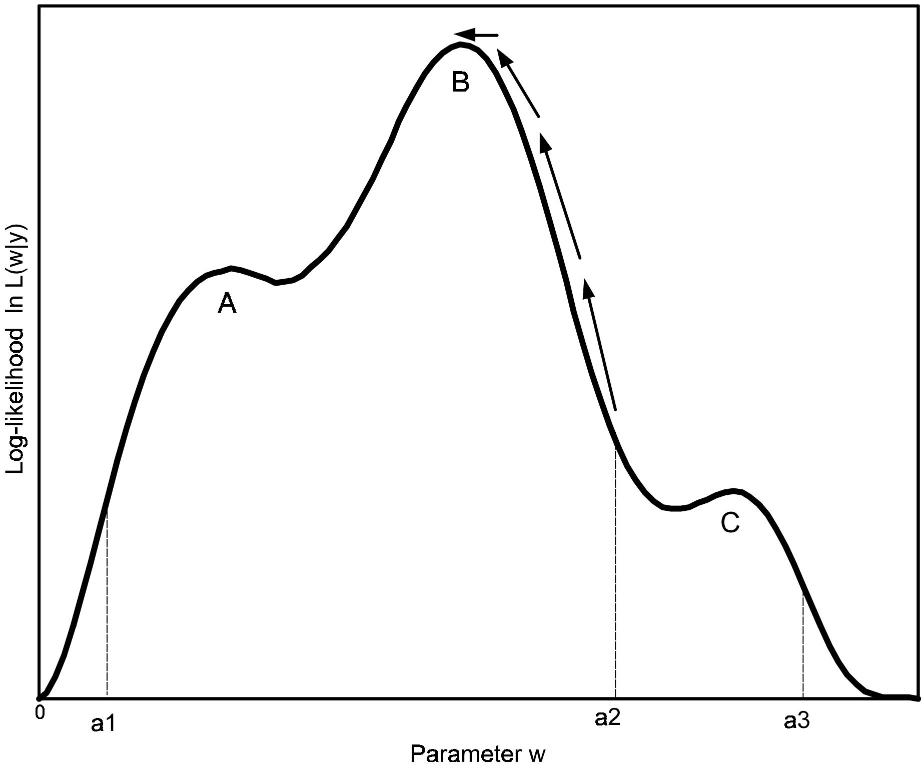


Fig. 3. A schematic plot of the log-likelihood function for a fictitious one-parameter model. Point B is the global maximum whereas points A and C are two local maxima. The series of arrows depicts an iterative optimization process.

searching much smaller sub-sets of the multi-dimensional parameter space rather than exhaustively searching the whole parameter space, which becomes intractable as the number of parameters increases. The "intelligent" search proceeds by trial and error over the course of a series of iterative steps. Specifically, on each iteration, by taking into account the results from the previous iteration, a new set of parameter values is obtained by adding small changes to the previous parameters in such a way that the new parameters are likely to lead to improved performance. Different optimization algorithms differ in how this updating routine is conducted. The iterative process, as shown by a series of arrows in Fig. 3, continues until the parameters are judged to have converged (i.e., point B in Fig. 3) on the optimal set of parameters on an appropriately predefined criterion. Examples of the stopping criterion include the maximum number of iterations allowed or the minimum amount of change in parameter values between two successive iterations.

3.2. *Local maxima*

It is worth noting that the optimization algorithm does not necessarily guarantee that a set of parameter values that uniquely maximizes the log-likelihood will be found. Finding optimum parameters is essentially a heuristic process in which the optimization algorithm tries to improve upon an initial set of parameters that is supplied by the user. Initial parameter values are chosen either at random or by guessing. Depending upon the choice of the initial parameter values, the algorithm could prematurely stop and return a sub-optimal set of parameter values. This is called the *local maxima* problem. As an example, in Fig. 3 note that although the starting parameter value at point a2 will lead to the optimal point B called the *global maximum*, the starting parameter value at point a1 will lead to point A, which is a sub-optimal solution. Similarly, the starting parameter value at a3 will lead to another sub-optimal solution at point C.

Unfortunately, there exists no general solution to the local maximum problem. Instead, a variety of techniques have been developed in an attempt to avoid the problem, though there is no guarantee of their effectiveness. For example, one may choose different starting values over multiple runs of the iteration procedure and then examine the results to see whether the same solution is obtained repeatedly. When that happens, one can conclude with some confidence that a global maximum has been found.[2]

[2] A stochastic optimization algorithm known as simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983) can overcome the local maxima problem, at least in theory, though the algorithm may not be a feasible option in practice as it may take an realistically long time to find the solution.

3.3. Relation to least-squares estimation

Recall that in MLE we seek the parameter values that are *most likely* to have produced the data. In LSE, on the other hand, we seek the parameter values that provide the *most accurate* description of the data, measured in terms of how closely the model fits the data under the square-loss function. Formally, in LSE, the *sum of squares error* (SSE) between observations and predictions is minimized:

$$SSE(w) = \sum_{i=1}^{m} (y_i - prd_i(w))^2, \tag{12}$$

where $prd_i(w)$ denotes the model's prediction for the ith observation. Note that $SSE(w)$ is a function of the parameter vector $w = (w_1, \ldots, w_k)$.

As in MLE, finding the parameter values that minimize SSE generally requires use of a non-linear optimization algorithm. Minimization of LSE is also subject to the local minima problem, especially when the model is non-linear with respect to its parameters. The choice between the two methods of estimation can have non-trivial consequences. In general, LSE estimates tend to differ from MLE estimates, especially for data that are not normally distributed such as proportion correct and response time. An implication is that one might possibly arrive at different conclusions about the same data set depending upon which method of estimation is employed in analyzing the data. When this occurs, MLE should be preferred to LSE, unless the probability density function is unknown or difficult to obtain in an easily computable form, for instance, for the diffusion model of recognition memory (Ratcliff, 1978).[3] There is a situation, however, in which the two methods intersect. This is when observations are independent of one another and are normally distributed with a constant variance. In this case, maximization of the log-likelihood is equivalent to minimization of SSE, and therefore, the same parameter values are obtained under either MLE or LSE.

4. Illustrative example

In this section, I present an application example of maximum likelihood estimation. To illustrate the method, I chose forgetting data given the recent surge of interest in this topic (e.g. Rubin & Wenzel, 1996; Wickens, 1998; Wixted & Ebbesen, 1991).

Among a half-dozen retention functions that have been proposed and tested in the past, I provide an example of MLE for the two functions, power and exponential. Let $w = (w_{1,}w_2)$ be the parameter vector, t time, and $p(w,t)$ the model's prediction of the probability of correct recall at time t. The two models are defined as

$$\begin{aligned} &\text{power model}: \quad p(w,t) = w_1 t^{-w_2} \quad (w_1, w_2 > 0), \\ &\text{exponential model}: \quad p(w,t) = w_1 \exp(-w_2 t) \\ &\qquad\qquad\qquad (w_1, w_2 > 0). \end{aligned} \tag{13}$$

Suppose that data $y = (y_1, \ldots, y_m)$ consists of m observations in which $y_i (0 \leqslant y_i \leqslant 1)$ represents an observed proportion of correct recall at time t_i $(i = 1, \ldots, m)$. We are interested in testing the viability of these models. We do this by fitting each to observed data and examining its goodness of fit.

Application of MLE requires specification of the PDF $f(y|w)$ of the data *under each model*. To do this, first we note that each observed proportion y_i is obtained by dividing the number of correct responses (x_i) by the total number of independent trials $(n), y_i = x_i/n$ $(0 \leqslant y_i \leqslant 1)$ We then note that each x_i is binomially distributed with probability $p(w,t)$ so that the PDFs for the power model and the exponential model are obtained as

$$\begin{aligned} &\text{power}: \quad f(x_i \mid n, w) = \frac{n!}{(n-x_i)!x_i!} \\ &\qquad (w_1 t_i^{-w_2})^{x_i} (1 - w_1 t_i^{-w_2})^{n-x_i}, \\ &\text{exponential}: \quad f(x_i \mid n, w) = \frac{n!}{(n-x_i)!x_i!} \\ &\qquad (w_1 \exp(-w_2 t_i))^{x_i} \\ &\qquad (1 - w_1 \exp(-w_2 t_i))^{n-x_i}, \end{aligned} \tag{14}$$

where $x_i = 0, 1, \ldots, n,\ \ i = 1, \ldots, m.$

There are two points to be made regarding the PDFs in the above equation. First, the probability parameter of a binomial probability distribution (i.e. w in Eq. (4)) is being modeled. Therefore, the PDF for each model in Eq. (14) is obtained by simply replacing the probability parameter w in Eq. (4) with the model equation, $p(w,t)$, in Eq. (13). Second, note that y_i is related to x_i by a fixed scaling constant, $1/n$. As such, any statistical conclusion regarding x_i is applicable directly to y_i, except for the scale transformation. In particular, the PDF for y_i, $f(y_i|n,w)$, is obtained by simply replacing x_i in $f(x_i|n,w)$ with ny_i.

Now, assuming that x_i's are statistically independent of one another, the desired log-likelihood function for the power model is given by

$$\begin{aligned} \ln L(w = (w_1, w_2)|n, x) &= \ln(f(x_1|n,w) \cdot f(x_2 \mid n, w) \cdots f(x_m \mid n, w)) \\ &= \sum_{i=1}^{m} \ln f(x_i|n,w) \\ &= \sum_{i=1}^{m} (x_i \ln(w_1 t_i^{-w_2}) + (n - x_i) \ln(1 - w_1 t_i^{-w_2}) \\ &\quad + \ln n! - \ln(n-x_i)! - \ln x_i!). \end{aligned} \tag{15}$$

[3] For this model, the PDF is expressed as an infinite sum of transcendental functions.

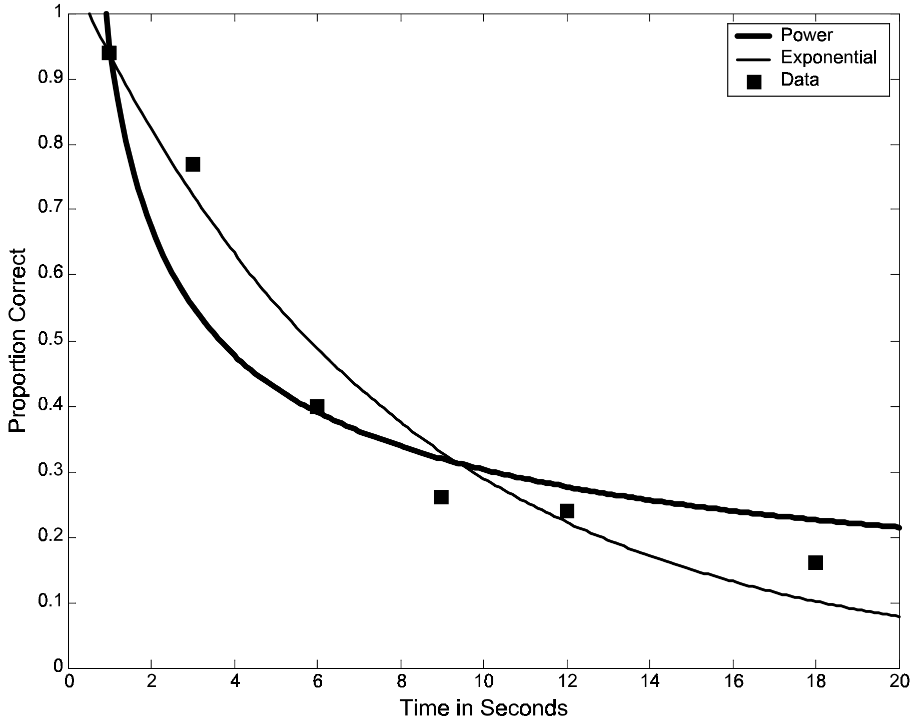


Fig. 4. Modeling forgetting data. Squares represent the data in Murdock (1961). The thick (respectively, thin) curves are best fits by the power (respectively, exponential) models.

Table 1
Summary fits of Murdock (1961) data for the power and exponential models under the maximum likelihood estimation (MLE) method and the least-squares estimation (LSE) method.

	MLE		LSE	
	Power	Exponential	Power	Exponential
Loglik/SSE (r^2)	−313.37 (0.886)	−305.31 (0.963)	0.0540 (0.894)	0.0169 (0.967)
Parameter w_1	0.953	1.070	1.003	1.092
Parameter w_2	0.498	0.131	0.511	0.141

Note: For each model fitted, the first row shows the maximized log-likelihood value for MLE and the minimized sum of squares error value for LSE. Each number in the parenthesis is the proportion of variance accounted for (i.e. r^2) in that case. The second and third rows show MLE and LSE parameter estimates for each of w_1 and w_2. The above results were obtained using Matlab code described in the appendix.

This quantity is to be maximized with respect to the two parameters, w_1 and w_2. It is worth noting that the last three terms of the final expression in the above equation (i.e., $\ln n! - \ln(n - x_i)! - \ln x_i!$) do not depend upon the parameter vector, thereby do not affecting the MLE results. Accordingly, these terms can be ignored, and their values are often omitted in the calculation of the log-likelihood. Similarly, for the exponential model, its log-likelihood function can be obtained from Eq. (15) by substituting $w_1 \exp(-w_2 t_i)$ for $w_1 t_i^{-w_2}$.

In illustrating MLE, I used a data set from Murdock (1961). In this experiment subjects were presented with a set of words or letters and were asked to recall the items after six different retention intervals, $(t_1, \ldots, t_6) = (1, 3, 6, 9, 12, 18)$ in seconds and thus, $m = 6$. The proportion recall at each retention interval was calculated based on 100 independent trials (i.e. $n = 100$) to yield the observed data $(y_1, \ldots, y_6) = (0.94, 0.77, 0.40, 0.26, 0.24, 0.16)$, from which the number of correct responses, x_i, is obtained as $100y_i$, $i = 1, \ldots, 6$. In Fig. 4, the proportion recall data are shown as squares.

The curves in Fig. 4 are best fits obtained under MLE. Table 1 summarizes the MLE results, including fit measures and parameter estimates, and also include the LSE results, for comparison. Matlab code used for the calculations is included in the appendix.

The results in Table 1 indicate that under either method of estimation, the exponential model fit better than the power model. That is, for the former, the log-likelihood was larger and the SSE smaller than for the latter. The same conclusion can be drawn even in terms of r^2. Also note the appreciable discrepancies in parameter estimate between MLE and LSE. These differences are not unexpected and are due to the fact that the proportion data are binomially distributed, not normally distributed. Further, the constant variance assumption required for the equivalence between MLE and LSE does not hold for binomial data for which the variance, $\sigma^2 = np(1-p)$, depends upon proportion correct p.

4.1. MLE interpretation

What does it mean when one model fits the data better than does a competitor model? It is important not to jump to the conclusion that the former model does a better job of capturing the underlying process and therefore represents a closer approximation to the true model that generated the data. A good fit is a necessary, but not a sufficient, condition for such a conclusion. A superior fit (i.e., higher value of the maximized log-likelihood) merely puts the model in a list of candidate models for further consideration. This is because a model can achieve a superior fit to its competitors for reasons that have nothing to do with the model's fidelity to the underlying process. For example, it is well established in statistics that a complex model with many parameters fits data better than a simple model with few parameters, even if it is the latter that generated the data. The central question is then how one should decide among a set of competing models. A short answer is that a model should be selected based on its generalizability, which is defined as a model's ability to fit current data but also to predict future data. For a thorough treatment of this and related issues in model selection, the reader is referred elsewhere (e.g. Linhart & Zucchini, 1986; Myung, Forster, & Browne, 2000; Pitt, Myung, & Zhang, 2002).

5. Concluding remarks

This article provides a tutorial exposition of maximum likelihood estimation. MLE is of fundamental importance in the theory of inference and is a basis of many inferential techniques in statistics, unlike LSE, which is primarily a descriptive tool. In this paper, I provide a simple, intuitive explanation of the method so that the reader can have a grasp of some of the basic principles. I hope the reader will apply the method in his or her mathematical modeling efforts so a plethora of widely available MLE-based analyses (e.g. Batchelder & Crowther, 1997; Van Zandt, 2000) can be performed on data, thereby extracting as much information and insight as possible into the underlying mental process under investigation.

Acknowledgments

This work was supported by research Grant R01 MH57472 from the National Institute of Mental Health. The author thanks Mark Pitt, Richard Schweickert, and two anonymous reviewers for valuable comments on earlier versions of this paper.

Appendix

This appendix presents Matlab code that performs MLE and LSE analyses for the example described in the text.

Matlab Code for MLE

```
%  This is the main program that finds MLE estimates. Given a model, it
%  takes sample size (n), time intervals (t) and observed proportion correct
%  (y) as inputs. It returns the parameter values that maximize the log-
% likelihood function
global n t x; % define global variables
opts = optimset (‘DerivativeCheck’,‘off’,’Display’,‘off’,‘TolX’,1e-6,‘TolFun’,1e-6,
‘Diagnostics’,‘off’,‘MaxIter’,200,LargeScale’,‘off’);
  % option settings for optimization algorithm
n = 100 ;% number of independent Bernoulli trials (i.e., sample size)
t = [1 3 6 9 12 18]';% time intervals as a column vector
y = [.94 .77 .40 .26 .24 .16]';% observed proportion correct as a column vector
x = n*y;% number of correct responses

init_w = rand(2,1);% starting parameter values
low_w = zeros(2,1);% parameter lower bounds
up_w = 100*ones(2,1);% parameter upper bounds

while 1,
[w1,lik1,exit1] = fmincon (‘power_mle’,init_w,[],[],[],[],low_w,up_w,[],opts);
```

```
   % optimization for power model that minimizes minus log-likelihood (note that minimization of
   minus log-likelihood is equivalent to maximization of log-likelihood)
   % w1: MLE parameter estimates
   % lik1: maximized log-likelihood value
   % exit1: optimization has converged if exit1 >0 or not otherwise
[w2, lik2, exit2] = FMINCON(‘EXPO_MLE’, INIT_W,[], [], [], [], LOW_W, UP_W, [], OPTS);
   % optimization for exponential model that minimizes minus log-likelihood
prd1 = w1(1, 1)*t.^(-w1(2, 1));% best fit prediction by power model
r2(1, 1) = 1-sum((prd1-y).^2)/sum((y-mean(y)).^2); % r^2 for power model
prd2 = w2(1, 1)*exp(-w2(2, 1)*t);% best fit prediction by exponential model
r2(2, 1) = 1-sum((prd2-y).^2)/sum((y-mean(y)).^2); %r^2 for exponential model

if sum(r2>0) == 2
        break;
else
     init_w = rand(2, 1);
end;
end;

format long;
disp(num2str([w1 w2 r2],5));% display results
disp(num2str([lik1 lik2 exit1 exit2],5));% display results
end % end of the main program

function loglik = power_mle(w)
% POWER_MLE The log-likelihood function of the power model
     global n t x;
   p = w(1, 1)*t.^(-w(2, 1));% power model prediction given parameter
p = p + (p == zeros(6, 1))*1e-5 − (p == ones(6, 1))*1e-5; % ensure 0<p<1
  loglik = (−1)*(x.* log(p) + (n-x).* log(1-p));
   % minus log-likelihood for individual observations
  loglik = sum(loglik);% overall minus log-likelihood being minimized

function loglik = expo_mle(w)
% EXPO_MLE The log-likelihood function of the exponential model
   global n t x;
   p = w(1, 1)*exp(-w(2, 1)*t);% exponential model prediction
   p = p + (p == zeros(6, 1))*1e-5 − (p == ones(6, 1))*1e-5; % ensure 0<p<1
   loglik = (−1)*(x.* log(p) + (n-x).* log(1p));
        % minus log-likelihood for individual observations
loglik = sum(loglik);% overall minus log-likelihood being minimized
```

Matlab Code for LSE

```
% This is the main program that finds LSE estimates. Given a model, it
% takes sample size (n), time intervals (t) and observed proportion correct
% (y) as inputs. It returns the parameter values that minimize the sum of
% squares error
global t; % define global variable
opts = optimset(‘DerivativeCheck’,‘off’,‘Display’,‘off’,‘TolX’,1e-6,‘TolFun’,1e-6, ‘Diagnostic-
s’,‘off’,‘MaxIter’,200,‘LargeScale’,‘off’);
             % option settings for optimization algorithm
n = 100; % number of independent binomial trials (i.e., sample size)
t = [1 3 6 9 12 18]′;% time intervals as a column vector
```

```
y = [.94 .77 .40 .26 .24 .16]′;% observed proportion correct as a column vector

init_w = rand(2, 1);% starting parameter values
low_w = zeros(2, 1);% parameter lower bounds
up_w = 100*ones(2, 1);% parameter upper bounds
[w1, sse1, res1, exit1] = lsqnonlin(‘power_lse′, init_w, low_w, up_w, opts, y);
             % optimization for power model
             % w1: LSE estimates
             % sse1: minimized SSE value
             % res1: value of the residual at the solution
             % exit1: optimization has converged if exit1 >0 or not otherwise
[w2, sse2, res2, exit2] = lsqnonlin(‘expo_lse′, init_w, low_w, up_w, opts, y);
             % optimization for exponential model

r2(1, 1) = 1-sse1/sum((y-mean(y)).^2);% r^2 for power model
r2(2, 1) = 1-sse2/sum((y-mean(y)).^2);% r^2 for exponential model

format long;
disp(num2str([w1 w2 r2],5));% display out results
disp(num2str([sse1 sse2 exit1 exi2],5));% display out results
end % end of the main program

function dev = power_lse(w, y)
% POWER_LSE The deviation between observation and prediction of the power
% model
   global t;
   p = w(1, 1)*t.^(-w(2, 1));% power model prediction
   dev = p − y;
                % deviation between prediction and observation, the square of which is
                     being minimized

function dev = expo_lse(w, y)
% EXPO_LSE The deviation between observation and prediction of the
% exponential model
     global t;
     p = w(1, 1)*exp(−w(2, 1)*t);% exponential model prediction
     dev = p − y;
             % deviation between prediction and observation, the square of which is
             being minimized
```

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrox, B.N., & Caski, F. *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Batchelder, W. H., & Crowther, C. S. (1997). Multinomial processing tree models of factorial categorization. *Journal of Mathematical Psychology*, *41*, 45–55.

Bickel, P. J., & Doksum, K. A. (1977). *Mathematical statistics.* Oakland, CA: Holden-day, Inc.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxberry.

DeGroot, M. H., & Schervish, M. J. (2002). *Probability and statistics* (3rd ed.). Boston, MA: Addison-Wesley.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.

Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review*, *107*(2), 227–260.

Linhart, H., & Zucchini, W. (1986). *Model selection*. New York, NY: Wiley.

Murdock Jr., B. B. (1961). The retention of individual items. *Journal of Experimental Psychology*, *62*, 618–625.

Myung, I. J., Forster, M., & Browne, M. W. (2000). Special issue on model selection. *Journal of Mathematical Psychology*, *44*, 1–2.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Rubin, D. C., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1161–1176.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*, 734–760.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Spanos, A. (1999). *Probability theory and statistical inference*. Cambridge, UK: Cambridge University Press.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice; The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*(3), 424–465.

Wickens, T. D. (1998). On the form of the retention function: Comment on Rubin and Wenzel (1996): A quantitative description of retention. *Psychological Review*, *105*, 379–386.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, *2*, 409–415.