Model Calibration & Selection

Marisa Eisenberg
University of Michigan, Ann Arbor
Parameter Estimation

- In general—search parameter space to find optimal fit to data
- Or to characterize distribution of parameters that matches data
Parameter Estimation

- Basic idea: parameters that give model behavior that more closely matches data are ‘best’ or ‘most likely’

- Frame this from a statistical perspective (inference, regression)

- Can determine ‘most likely’ parameters or distribution, confidence intervals, etc.
How to frame this statistically?

- **Maximum Likelihood Approach**
  
  Idea: rewrite the ODE model as a statistical model, where we suppose we know the general form of the density function but not the parameter values.

  Then if we knew the parameters we could calculate probability of a particular observation/data:

  $$P(z \mid p)$$

  ![Diagram](data parameters)
Maximum Likelihood

- **Likelihood Function**

\[ P(z \mid p) = f(z, p) = L(p \mid z) \]

- Re-think the distribution as a function of the data instead of the parameters

- E.g.

\[ f(z \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left( - \frac{(z - \mu)^2}{2\sigma^2} \right) = L(\mu, \sigma^2 \mid z) \]

- Find the value of \( p \) that maximizes \( L(p \mid z) \) - this is the maximum likelihood estimate (MLE) (most likely given the data)
Likelihood Function

PDF given a parameter value
Likelihood Function

Move the parameter and the distribution shifts
Likelihood Function

Parameter value vs Data value graph.
Likelihood Function
Likelihood Function

PDF given a parameter value
Likelihood Function

Parameter value

Data value

Likelihood function given data
Maximum Likelihood

- **Consistency** - with sufficiently large number of observations $n$, it is possible to find the value of $p$ with arbitrary precision (i.e. converges in probability to $p$)

- **Normality** - as the sample size increases, the distribution of the MLE tends to a Gaussian distribution with mean and covariance matrix equal to the inverse of the Fisher information matrix

- **Efficiency** - achieves CR bound as sample size $\rightarrow \infty$ (no consistent estimator has lower asymptotic mean squared error than MLE)
Example - ODE Model with Gaussian Error

- Model:
  \[
  \dot{x} = f(x, t, p) \\
  y = g(x, t, p)
  \]

- Suppose data is taken at times \( t_1, t_2, \ldots, t_n \)

- Data at \( t_i = z_i = y(t_i) + e_i \)

- Suppose error is gaussian and unbiased, with known variance \( \sigma^2 \) (can also be considered an unknown parameter)
Example - ODE Model with Gaussian Error

• The measured data $z_i$ at time $i$ can be viewed as a sample from a Gaussian distribution with mean $y(x, t_i, p)$ and variance $\sigma^2$

• Suppose all measurements are independent (is this realistic?)
Example - ODE Model with Gaussian Error

- Then the likelihood function can be calculated as:

\[
f(z_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i - \mu)^2}{2\sigma^2}\right)
\]
Example - ODE Model with Gaussian Error

• Then the likelihood function can be calculated as:

Gaussian PDF:  \[ f(z_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i - \mu)^2}{2\sigma^2}\right) \]

Formatted for model:

\[ f(z_i \mid y(x,t_i,p),\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i - y(t_i,p))^2}{2\sigma^2}\right) \]
Example - ODE Model with Gaussian Error

- Then the likelihood function can be calculated as:

  Gaussian PDF: \[ f(z_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i - \mu)^2}{2\sigma^2}\right) \]

  Formatted for model: \[ f(z_i | y(x,t_i,p),\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i - y(t_i,p))^2}{2\sigma^2}\right) \]

  Likelihood function assuming independent observations:

  \[ L(y(t_i,p),\sigma^2 | z_1,\ldots,z_n) = f(z_1,\ldots,z_n | y(t_i,p),\sigma^2) = \prod_{i=1}^{n} f(z_i | y(t_i,p),\sigma^2) \]
Example - ODE Model with Gaussian Error

\[
L(y(t_i, p), \sigma^2 | z_1, \ldots, z_n) = f(z_1, \ldots, z_n | y(t_i, p), \sigma^2) \\
= \prod_{i=1}^{n} f(z_i | y(t_i, p), \sigma^2) \\
= \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left( - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (z_i - y(t_i, p))^2 \right)
\]
Example - ODE Model with Gaussian Error

- It is often more convenient to minimize the Negative Log Likelihood (-LL) instead of maximizing the Likelihood

- Log is well behaved, minimization algorithms common

\[
-LL = -\ln \left( \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left( -\frac{\sum_{i=1}^{n} (z_i - y(t_i, p))^2}{2\sigma^2} \right) \right)
\]
Example - ODE Model with Gaussian Error

\[-LL = -\ln \left( \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left( -\frac{\sum_{i=1}^{n} (z_i - y(t_i, p))^2}{2\sigma^2} \right) \right) \]

\[-LL = \left\{ -\frac{n}{2} \ln(2\pi) - n \ln(\sigma) - \frac{\sum_{i=1}^{n} (z_i - y(t_i, p))^2}{2\sigma^2} \right\} \]
Example - ODE Model with Gaussian Error

\[-LL = \frac{n}{2} \ln(2\pi) + n \ln(\sigma) + \frac{\sum_{i=1}^{n}(z_i - y(t_i, p))^2}{2\sigma^2}\]

If \(\sigma\) is known, then first two terms are constants & will not be changed as \(p\) is varied—so we can minimize only the 3rd term and get the same answer

\[
\min_p (-LL) = \min_p \left( \frac{\sum_{i=1}^{n}(z_i - y(t_i, p))^2}{2\sigma^2} \right)
\]
Example - ODE Model with Gaussian Error

- Similarly for denominator:

\[
\min_p (-LL) = \min_p \left( \frac{\sum_{i=1}^{n} (z_i - y(t_i, p))^2}{2\sigma^2} \right) = \min_p \left( \sum_{i=1}^{n} (z_i - y(t_i, p))^2 \right)
\]

- This is just least squares!

- So, least squares is equivalent to the ML estimator when we assume a constant known variance.
Maximum Likelihood Summary for ODEs

• Can calculate other ML estimators for different distributions

• Not always least squares-ish! (mostly not)

• Although surprisingly, least squares does fairly decently a lot of the time
Example - Poisson ML

• For count data (e.g. incidence data), the Poisson distribution is often more realistic than Gaussian

• Likelihood function?
Example - Poisson ML

- Model:
  \[ \dot{x} = f(x, t, p) \]
  \[ y = g(x, t, p) \]

- Data \( z_i \) is assumed to be Poisson with mean \( y(t_i) \)

- Assume all data points are independent

- Poisson PMF:
  \[ f(z_i \mid y(t_i)) = \frac{y(t_i)^{z_i} e^{-y(t_i)}}{z_i!} \]
Example - Poisson ML

- Likelihood function:

\[
L(y(t,p) \mid z_1, \ldots, z_n) = f(z_1, \ldots, z_n \mid y(t,p))
= \prod_{i=1}^{n} f(z_i \mid y(t,p))
= \prod_{i=1}^{n} \frac{y(t_i)^{z_i} e^{-y(t_i)}}{z_i!}
\]
Poisson ML

• Negative log likelihood:

\[-LL = - \ln \left( \prod_{i=1}^{n} \frac{y(t_i)^{z_i} e^{-y(t_i)}}{z_i!} \right)\]

\[= - \sum_{i=1}^{n} \ln \left( \frac{y(t_i)^{z_i} e^{-y(t_i)}}{z_i!} \right)\]

\[= - \sum_{i=1}^{n} z_i \ln(y(t_i)) + \sum_{i=1}^{n} y(t_i) + \sum_{i=1}^{n} \ln(z_i)\]

• Last term is constant
Example - Poisson ML

- Poisson ML Estimator:

\[
\min_p (-LL) = \min_p \left( -\sum_{i=1}^{n} z_i \ln(y(t_i)) + \sum_{i=1}^{n} y(t_i) \right)
\]

- Other common distributions - negative binomial (overdispersion), zero-inflated poisson or negative binomial, etc.
Maximum Likelihood Summary for ODEs

- Basic approach - suppose only measurement error
- Data is given by distribution where model output is the mean
- Suppose each time point of data is independent
- Use PDF/PMF to calculate the likelihood
- Take the negative log likelihood, minimize this over the parameter space
Very (very!) brief intro to Bayesian Approaches to Parameter Estimation

- Allows one to account for prior information about the parameters
  - E.g. previous studies in a similar population
- Update parameter information based on new data
- Recall Bayes’ Theorem:

\[
P(p \mid z) = P(params \mid data) = \frac{P(z \mid p) \cdot P(p)}{P(z)}
\]
Very (very!) brief intro to Bayesian Approaches to Parameter Estimation

- Allows one to account for prior information about the parameters
  - E.g. previous studies in a similar population
- Update parameter information based on new data
- Recall Bayes’ Theorem:

$$P(p | z) = P(params | data) = \frac{P(z | p) \cdot P(p)}{P(z)}$$

Normalizing constant (can be difficult to calculate!)
Bayesian Parameter Estimation

- From prior distribution & likelihood distribution, determine the posterior distribution of the parameter.

- Can repeat this process as new data is available.
Bayesian Parameter Estimation

- Treats the parameters inherently as distributions (belief)
- Philosophical battle between Bayesian & frequentist perspectives
- Word of caution on choosing your priors
- Denominator issues - MAP Approach
DID THE SUN JUST EXPLODE?
(ITS NIGHT, SO WE'RE NOT SURE)

THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.

THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.

LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?

ROLL

YES.

FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{6}=0.027$.

SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.

BAYESIAN STATISTICIAN:

BET YOU $50 IT HASN'T.

from XKCD:
http://xkcd.com/1132/
Model Comparison
Examining the Model Fit to the Data

- The Eyeball Test!
- Negative log likelihood/RSS/etc.
- Parameter uncertainties & correlations
- Distribution of residuals - should make sense based on data assumptions
Examining the Model Fit to the Data

- Correlation of residuals (e.g. serial correlation coefficient)
- Wald-Wolfowitz Runs Test
Model comparison

• ‘As simple as possible, but not simpler’ (Einstein)

• How simple to make the model? How to balance goodness-of-fit with parsimony?

• Often significant structural uncertainty—how to choose between candidate models or mechanisms?

• Touches on larger issues of model misspecification and structural uncertainty/sensitivity
Model Comparison

- Many methods - F-test, likelihood ratio tests, simply comparing goodness of fit, etc.

- One of the most common/popular —

- **Akaike Information Criterion** (AIC)
AIC

• More parameters - more degrees of freedom

• Expect models with more parameters may be able to fit data better

• Danger of overfitting - parsimony

• AIC accounts for goodness of fit & overparameterization
AIC

\[ \text{AIC} = -2 \ln(\max(L)) + 2k \]
\[ = 2 \min(-LL) + 2k \]

- where \( k \) is the number of parameters
- smaller AIC is better
- AIC = -LL + penalty term for parameters
AIC

• AIC can be derived from information theory - “information loss” when using one model versus another

• One AIC has no real meaning by itself—generally need to compare AICs of competing models
Some rough rules of thumb when comparing AICs

- $\Delta_i$ (difference in AIC) values less than 2 are often considered similarly good
- $\Delta_i \leq 6$ also may be considered
- “$\Delta_i$ values greater than 10 are sufficiently poorer than the best AIC model as to be considered implausible” (Symonds & Moussalli, Behav Ecol Sociobiol (2011) 65:13–21)
Other variations

- Many alternatives!
- BIC - stronger penalty on parameters

\[ BIC = \ln(n)k - 2\ln(ML) \]

- cAIC - correction for small data sets

\[ AIC_c = AIC + \frac{2k(k+1)}{n-k-1} \]