Maximum Likelihood Estimation Part II

ZOL851 – October28th 2014

Goals

- Continue on our search for truth… at least approximately….
- Continue developing our understanding of MLE. Both conceptually and computationally.
- Start developing inferential approaches associated with MLE.

Remaining Readings for likelihood.

- Bolker,B. EMD book 2008. Pp 196-212, 215-221.
- If you want to read about optimization methods: Bolker Chapter 7:222-233

Initial readings for Bayesian Analysis (starting next week).

- Bolker pg 107-115, 185-187, 194-196, 212-216.

What we want to find is the parameter estimates that maximize the likelihood (biggest single number)

Sometimes we are interested in the negative log likelihood, in which case we are looking for a minimum.

Principle of Maximum Likelihood

Find an estimate for θ such that it maximizes the likelihood of observing the data that were actually observed. In other words, given a sample of observations **x** for the random variable **X**, find the solution for θ that maximizes the joint probability function $p(\mathbf{x}|\theta)$.

Eliason 1993

How do we find the MLE?

Basic calculus reminds us that we can find minima and maxima by taking the derivative of our likelihood function with respect to the parameter(s) of interest.

For a model with a single parameter

$$S(\theta) = \frac{dL(\theta)}{d\theta}$$

S is called the Score function. Evaluating S(θ) = 0 will provide the MLE for θ.

More Generally

$$\mathbf{S}(\theta) = \frac{\partial L(\theta)}{\partial \theta} = \begin{matrix} \frac{\partial L(\theta)}{\partial \theta_1} \\ \frac{\partial L(\theta)}{\partial \theta_2} \\ \vdots \\ \vdots \\ \frac{\partial L(\theta)}{\partial \theta_n} \end{matrix}$$

i.e. partial derivative of the likelihood with respect to parameter 1.

i.e. partial derivative of the likelihood with respect to parameter 2.

i.e. partial derivative of the likelihood with respect to parameter n.

Example 1. Suppose n values, $z_1 \cdots z_n$, are sampled independently from an underlying normal with unknown mean μ and unit variance ($\sigma^2 = 1$). Letting $\mathbf{z} = (z_1, z_2, \cdots, z_n)$, what is the MLE for μ given $\mathbf{z}$? Since the observations are independent, the resulting probability density function for $\mathbf{z}$ is the product of n normal density functions,

$$\begin{aligned} p(\mathbf{z}, \mu) &= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left[-(z_i - \mu)^2/2\right] \\ &= (2\pi)^{-n/2} \exp\left[-\sum_{i=1}^{n} (z_i - \mu)^2/2\right] \end{aligned} \tag{A4.3}$$

The log-likelihood (or support) becomes

$$L(\mu \,|\, \mathbf{z}) = \ln\left[\ell(\mu \,|\, \mathbf{z})\right] = -\left(\frac{n}{2}\right) \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{n} (z_i - \mu)^2 \tag{A4.4}$$

which has the score function

$$S(\mu) = \frac{\partial\, L(\mu \,|\, \mathbf{z})}{\partial \mu} = \sum_{i=1}^{n} (\, z_i - \mu\,) = n(\overline{z} - \mu) \tag{A4.5}$$

Setting the score equal to zero and solving gives the MLE, $\widehat{\mu} = \overline{z}$.

Lynch and Walsh 1997

Analytical solutions versus numerical optimization

- Many people have derived the analytical solutions for numerous complex problems.
- However, for many problems, there have not been any such derivations. Thus we use numerical optimization.
- Much easier for us, since many such algorithms are built into R!!!

Large sample properties of MLE

- Consistency: As sample size increases the MLE converges to the true parameter value.

- Asymptotic normality and efficiency: As sample size increases, the sampling distribution of the MLE converges to normal, and generally no other estimation method has smaller variance. (I.e. for large sample sizes, MLE has the smallest confidence intervals).

- Variance: For large sample sizes, the variance of an MLE is (for a single parameter) ~

$$\sigma^2(\widehat{\theta}) \simeq -\left(\frac{\partial^2 L(\theta \mid \mathbf{z})}{\partial \theta^2} \bigg|_{\theta=\widehat{\theta}} \right)^{-1}$$

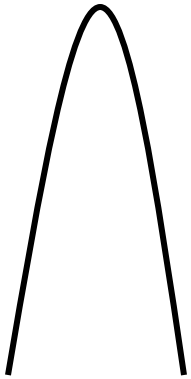
Lynch and Walsh 1997

First derivative gives us the maximum of the distribution (by giving us the point in the function where the slope = 0)

Second derivative tells us about the curvature of the function

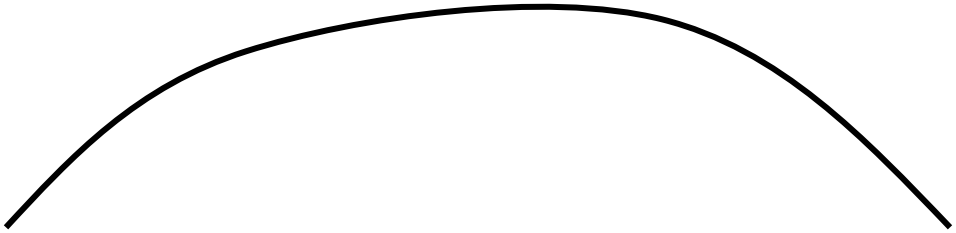
The curvature of the likelihood function provides information on uncertainty in the MLE

Clearly we have a lot of information in the curvature of our likelihood surface. I.e. if the curvature (slope) is large (high) near the MLE it suggests there is high precision in the estimate. This should lead to low SE.



High slope

High precision



Low slope. Low precision

Uncertainty in the MLE

We can use the information from the curvature to calculate the variance-covariance matrix of the parameters, θ.

$$(\mathrm{Var}(\theta))^{1/2} = \mathrm{SD}(\theta) = \mathrm{SE}(\theta)$$

To calculate the variance(θ) we need the **Hessian** Matrix

Hessian

- The Hessian (**H**) is the matrix of second partial derivates (from the Log-Likelihood).
- This includes the derivates on the same variables twice (μ & σ^2) as well as the *cross-derivates*.

$$\begin{bmatrix} \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \theta_2^2} \end{bmatrix}_{\theta=\hat{\theta}}$$

Evaluated at the MLE, theta-hat for the parameters.

Hessian

$$\begin{matrix} \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \mu^2} & \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \mu \partial \sigma^2} & \frac{\partial^2 L(\boldsymbol{\theta})}{\partial (\sigma^2)^2} \end{matrix}_{\theta = \hat{\mu}, \hat{\sigma}^2}$$

Evaluated at the MLE for the parameters.

Fisher (observed) information matrix

$\mathbf{F}(\theta) \approx - \mathbf{H}(\theta)$

Fisher information matrix is just the negative of the Hessian

When **F** is evaluated at the MLE for θ, it is called the *observed* information matrix.

The approximate Variance Covariance matrix for the parameter estimates is just:

$$VCOV(\theta) = \mathbf{F}(\theta)^{-1}$$

observed information matrix

When **F** is evaluated at the MLE for θ, it is called the *observed* information matrix.

The approximate Variance Covariance matrix for the parameter estimates is just:

$$VCOV(\theta) = \mathbf{F}(\theta)^{-1}$$

Confidence intervals/ellipses/envelopes constructed this way are quadratic approximations. Profiling the likelihood is preferred.

Thankfully R does all of the work for us.

Using either the mle() (stats4) or mle2 (bbmle) you can take your object (likelihood.x) and

vcov(likelihood.x) to get these numbers.

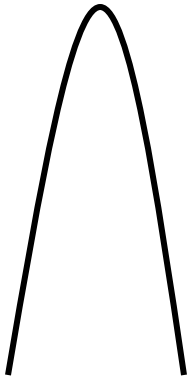
If you then take the square roots you get the standard deviations (standard errors for the parameters).

There is a bit of additional calculations that need to be done see pages 197-201.

Back to R

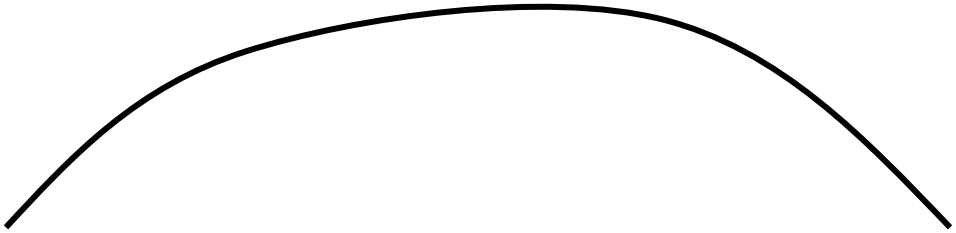
Likelihood profiles and slices

Clearly we have a lot of information in the curvature of our likelihood surface. I.e. if the curvature (slope) is large (high) near the MLE it suggests there is high precision in the estimate. This should lead to low SE.



High slope

High precision



Low slope. Low precision

Likelihood profiles and slices

But what do we do when we have more than a couple of parameters (so that it gets difficult to look at the curvature, even in contour plots).

We could do all pairwise contour plots for a small number of parameters.

But there are better ways.

Likelihood slice

- How do we look at one parameter at a time while accounting for all other parameters?

- The simplest, but “incorrect” way of doing this is by examining the likelihood “slices”.
- Fix all parameters at their MLE, except the focal parameter and let it vary.

Likelihood slice

- Can you think of what is wrong with the idea of a slice?

unrealistic to hold other parameters constant while examining just one ?

Likelihood profiles

- Instead we use likelihood profiles.
- We set the focal parameter to a range of values (near its MLE).
- At each of these new values we calculate the Maximum Likelihood across ALL other parameters.
- This generates “ridge-lines” in parameter space.

Likelihood profiles and slices

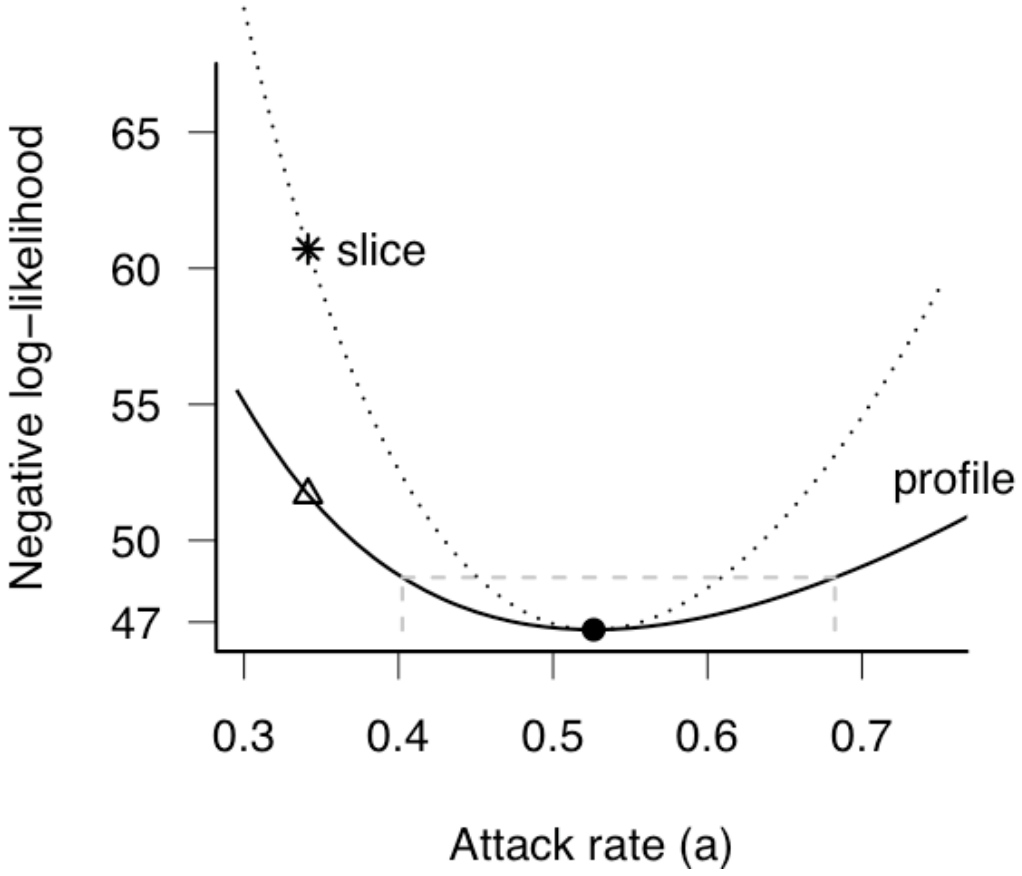


Figure 6.8

Likelihood Profiles

- When you use confint() for a likelihood object, the confidence intervals are approximate, based on the likelihood profile.
- Sometimes profiling fails numerically. You may then need to fall back on the quadratic approximation.
- You could also consider parametric or non-parametric bootstrapping to get CIs as well.

Likelihood Profiles

- Sometimes your profile will tell you it found a better MLE solution, so use those new values as starting values and refit your model.

Deviance

$$D(y) = -2[\log\{p(y \mid \hat{\theta}_M)\} - \log\{p(y \mid \hat{\theta}_s)\}]$$

$$= -2[L(\hat{\theta}_M \mid \mathrm{y}) - L(\hat{\theta}_s \mid \mathrm{y})]$$

$\hat{\theta}_M$ = MLE for parameters of model of interest

$\hat{\theta}_s$ = MLE for parameters of the ***saturated*** model

The saturated model is the model with a parameter estimated for every observation in the data set.

Deviance

Since the saturated model will be a constant you can also use this to "compare" between two (nested) model.

$$D(y) = -2[\log\{p(y \mid \hat{\theta}_{M1})\} - \log\{p(y \mid \hat{\theta}_{M2})\}]$$

$$= -2[L(\hat{\theta}_{M1} \mid y) - L(\hat{\theta}_{M2} \mid y)]$$

$\hat{\theta}_{M1}$ = MLE for parameters of reduced/restricted model

$\hat{\theta}_{M2}$ = MLE for parameters of the full model

Likelihood Ratio Test and confidence intervals

Edwards (1992) suggests can reasonably set confidence regions by including all parameter values within 2 log likelihood units of the maximum log-likelihood. (or e^2 ~7.4 without log transformation).

No Frequentist interpretation (I.e. no p value).

As an alternative we can also use the Likelihood ratio test (LRT).

Likelihood Ratio tests

- Likelihood ratio provides a useful and flexible approach for hypothesis testing and estimation of Confidence intervals.

$$LR = 2\ln\left(\frac{\ell(\widehat{\boldsymbol{\Theta}}\,|\,\mathbf{z})}{\ell(\widehat{\boldsymbol{\Theta}}_r\,|\,\mathbf{z})}\right) = -2\ln\left(\frac{\ell(\widehat{\boldsymbol{\Theta}}_r\,|\,\mathbf{z})}{\ell(\widehat{\boldsymbol{\Theta}}\,|\,\mathbf{z})}\right) = -2\left[\,L(\widehat{\boldsymbol{\Theta}}_r\,|\,\mathbf{z}) - L(\widehat{\boldsymbol{\Theta}}\,|\,\mathbf{z})\right]$$

This is sometimes also called the deviance… let's think about this…..

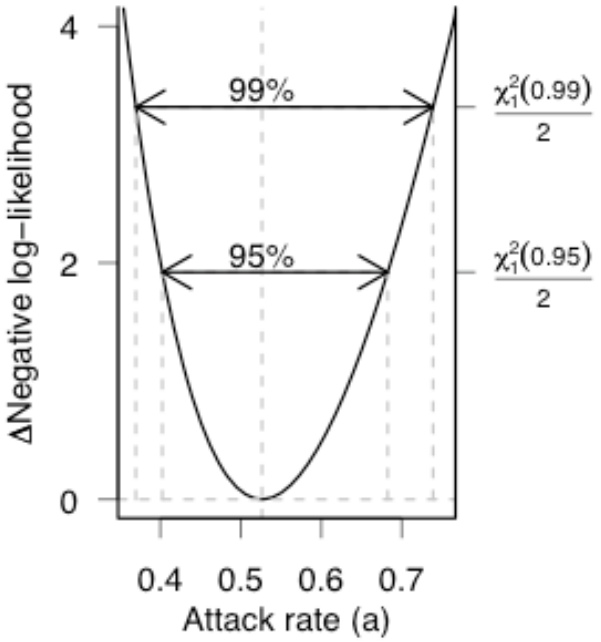
Where $L(\hat{\theta}|z)$ is the likelihood of the full model & $L(\hat{\theta}_r|z)$ is the "reduced" or "restricted" model.

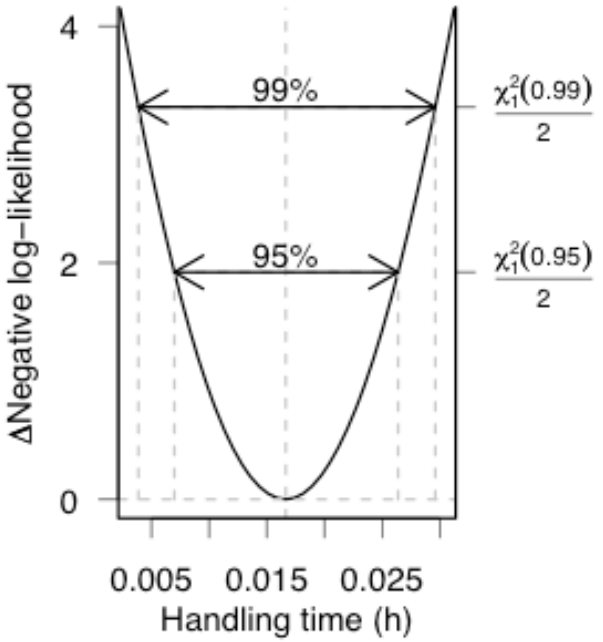
What does "reduced" or "restricted" mean?

LRT

- For large sample sizes the deviance is is ~ χ^2 distributed with *r* df.
- *r* is the number of restricted parameters (that are fixed) in the reduced model
- Put another way *r* is the number of "free" parameters for the model of interest in comparison to the reduced model.

LRT and confidence intervals





α	$\frac{\chi^2_1(\alpha)}{2}$	$-L + \frac{\chi^2_1(\alpha)}{2}$
0.95	1.92	
0.99	3.32	

LRT

- Clearly the LRT can also be used to test explicit hypotheses, against particular sets of reduced (fewer parameters) or null models. This is the general framework used for generalized linear models (and can be used for linear models as well).
- LRT can only be used to compare nested models.

LRT and Information criteria

- LRT can only be used to compare nested models (what do we mean by nested).

- To compare non-nested models we will start to learn about information criteria such as AIC & BIC.
- AIC = -2L + 2k = deviance + 2k.

L is the log-likelihood and k is the number of parameters in the model. Much more on this to come…For now smaller values of AIC/BIC indicate "*better*" fits…

LRT vs. Wald test

Tests of whether an effect “exists” or not in the NHST framework for MLE.

LRT

- Compares deviance
- Only an approximation/assumption is chi square distributed.
- The approximation fails when the MLE for the parameter is near the value under the restricted model (usually zero).
- It also works poorly when the MLE is near a boundary condition.

Wald test

- Two approximations/assumptions relevant.

Chi square distribution, and the approximation of the SE (from the MLE)

$$\frac{\left(\hat{\theta} - \theta_0\right)^2}{Var\left(\hat{\theta}\right)}$$

Which is ~ chi2 distributed.

Theta hat is the MLE while theta-not is the “pre-determined” parameter value (often 0).

Wald test

- Can also be formulated for normal distribution

$$\frac{\left(\hat{\theta} - \theta_0\right)}{se\left(\hat{\theta}\right)}$$

Which is ~ normally distributed

LRT vs. Wald test: Comment from Douglas Bates on October 13th 2008 to R-Mixed-Models group (in response to a slightly tongue in cheek query by Ben Bolker).

My reasoning, based on my experiences with nonlinear regression models and other nonlinear models, is that a test that involves fitting the alternative model and the null model then comparing the quality of the fit will give more realistic results than a test that only involves fitting the alternative model and using that fit to extrapolate to what the null model fit should be like.

We will always use approximations in statistics but as we get more powerful computing facilities some of the approximations that we needed to use in the past can be avoided. I view Wald tests as an approximation to the quantity that we want to use to compare models, which is some measure of the comparative fit. The likelihood ratio or the change in the deviance seems to be a reasonable way of comparing the fits of two nested models. There may be problems with calibrating that quantity (i.e. converting it to a p-value) in which case we may want to use a bootstrap or some other simulation-based method like MCMC. However, I don't think this difficulty would cause me to say that it is better to use an approximation to the model fit under the null hypothesis than to go ahead and fit it.

Ben Bolker's response to Douglas Bates (same date)

Doug's response makes perfect sense to me.

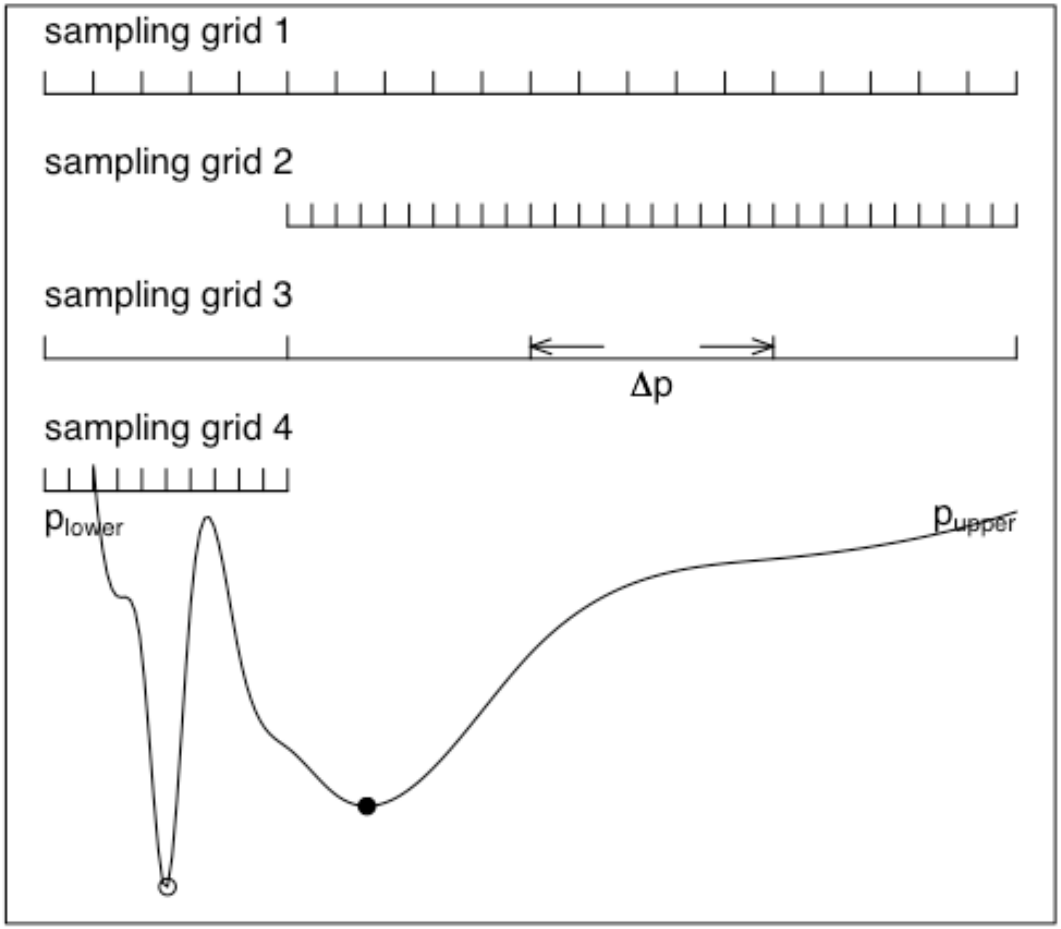
However, from the on-the-ground, what-do-I-say-about-my-data-now point of view, it seems that this is really an empirical question. I would guess (wildly) that both the LRT and the Wald test would converge asymptotically on the right answer. **For classical ML problems**, I have the feeling (unsupported by evidence!) that LRT converges faster/is less wrong at any given value of N than Wald tests (which, as you say, represent a second level of approximation). I have no idea if this is true for GLMMs. Really the only reason that I spoke against LRTs was that it is well known (as shown e.g. in PB2000) that they are dicey for LMMs, while the situation for Wald tests is relatively unknown. In the absence of data, which is stronger: our prior belief that Wald tests are bad because they're less reliable than LRT in some other contexts, or our optimism that Wald tests aren't bad because they haven't been shown to be so?

If it really hasn't been done (and while I'm far from omniscient I did *try* to review the literature on this topic, and have yet to find an answer, or to have anyone on this list provide an answer), I guess it's time to crank up the old simulation engine and have a look ...

Optimization

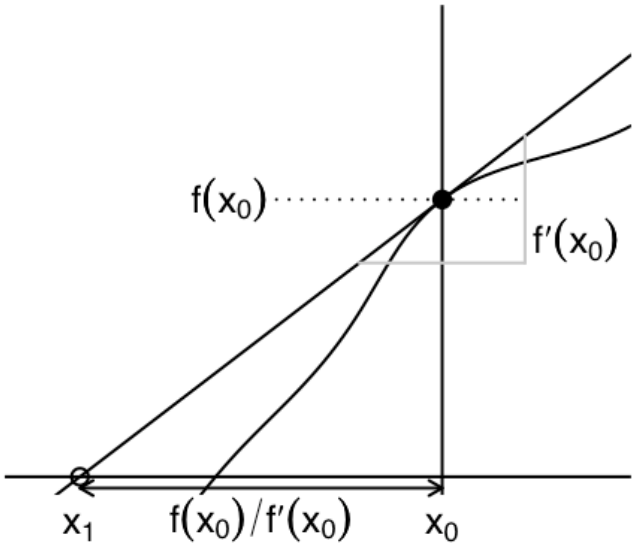
- Brute force/grid search
- Derivate based methods (Newton-Raphson, BFGS).
- Nelder-Mead Simplex
- Stochastic global optimization: Simulated annealing/ Metropolis algorithm.
- Expectation Maximization

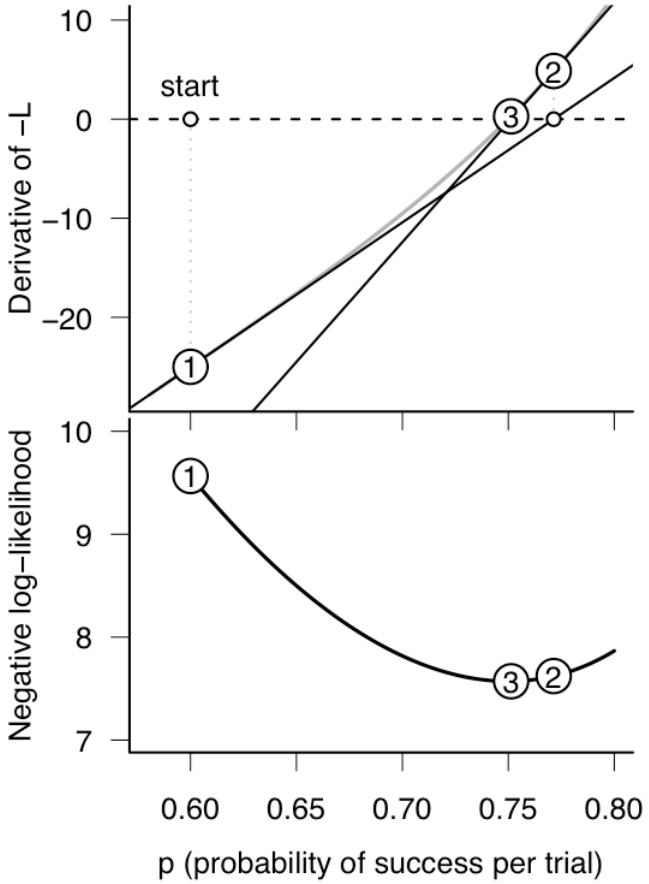
Brute Force



Slow (-)
May miss target region (-)
Not sensitive to strange curvature (+)

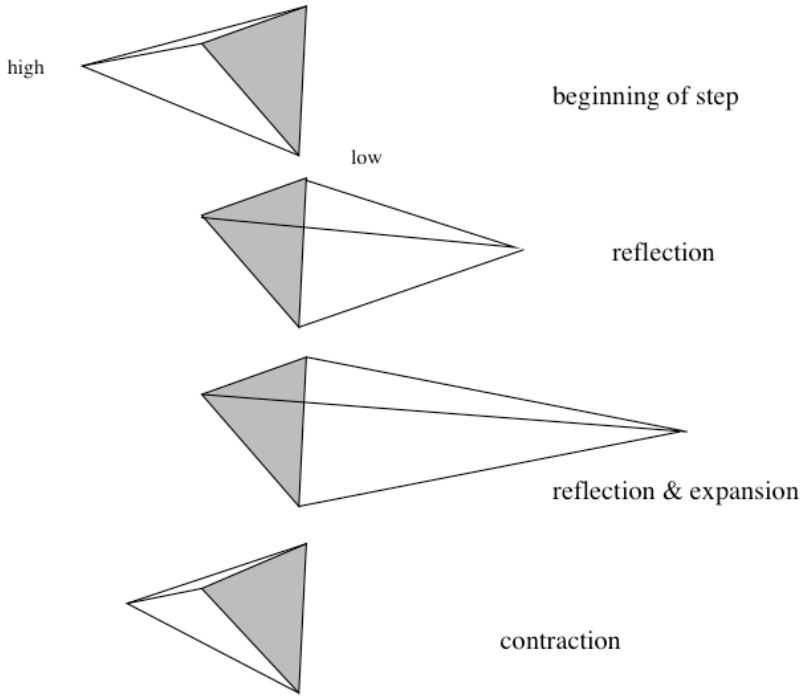
Newton-Raphson

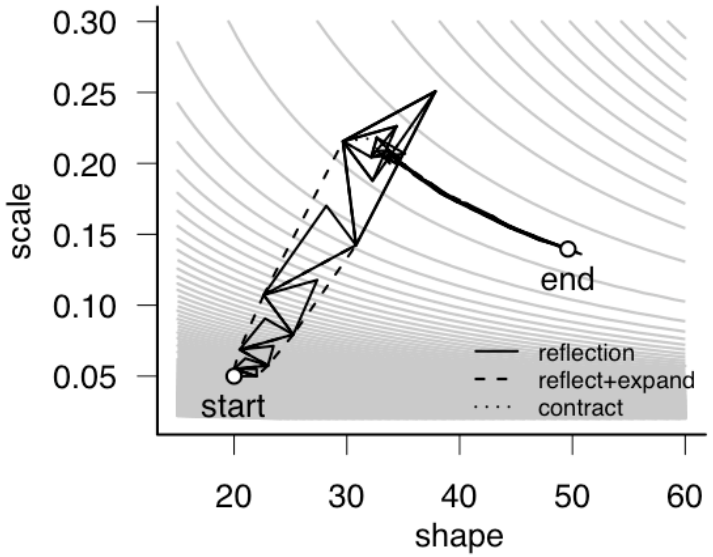




Faster (+), makes assumptions about curvature (-)
Looks for flat regions (slope near zero), this can happen for very bad fits as well.

Nelder-Mead Simplex



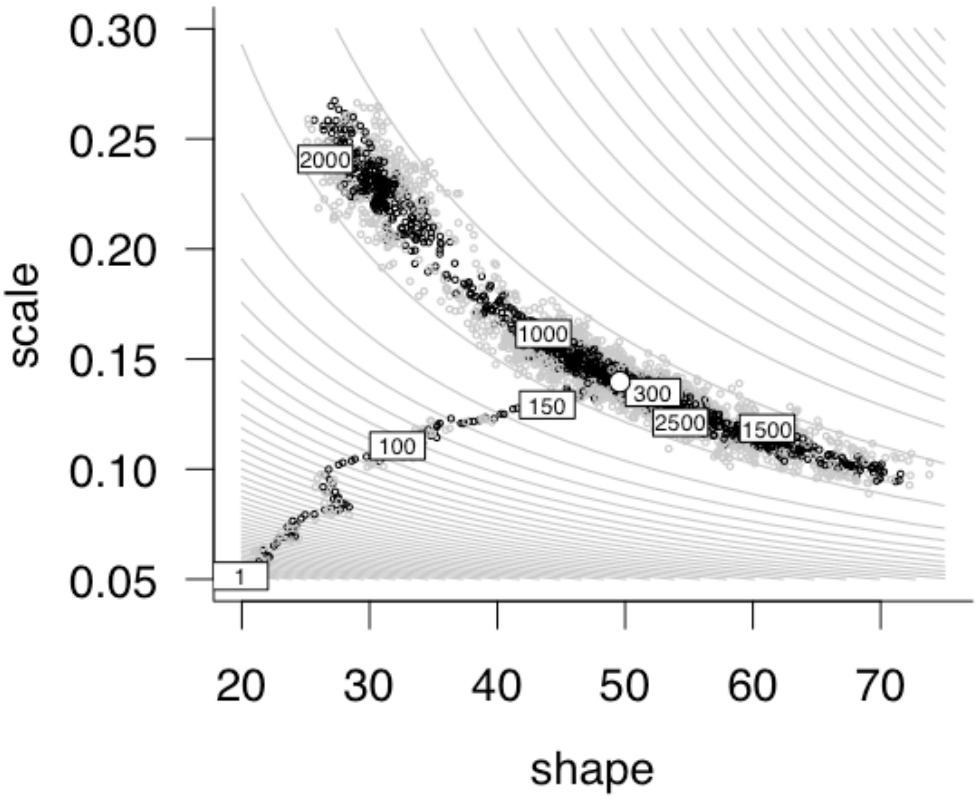


Robust (+), slow (-), does not work for one dimensional problems.

Metropolis/Simulated annealing (SANN): A stochastic global optimization method

1. Pick starting point, calculate -L
2. Pick new point at random, "near" old point. Calculate -L
3. If $-L_{new} < -L_{old}$, accept new point. If $-L_{new} > -Low_{old}$, calculate Δ-L. Pick random number between 0-1, accept new point if random number is $< e^{-\Delta(-L)/k}$, where k is a constant (temperature). Otherwise, reject new point, and keep old point.
4. Return to 2 and repeat. Periodically lower k.

Metropolis



Excellent for exploring complex likelihood surfaces (++)
SLOW (--).
Gets close quickly, but then slow to converge on the MLE.
Good way of getting better starting values. Then switch back to methods based on the curvature.

REML

- Fit model (estimate parameters) for all (and only) fixed effects (usually using OLS). Then perform MLE on the residuals of the model (i.e now the expected values for all parameters are zero)

- This is an example of using the “marginal likelihood”.

- REML estimates for variances are unbiased