Tutorials in Quantitative Methods for Psychology
2007, Vol. 3(1), p. 8-13.

DOI: 10.20982/tqmp.03.1.p008

Implementing and evaluating the nested maximum likelihood estimation technique

Denis Cousineau
Université de Montréal

Estimating parameters describing response time distributions is difficult. The most commonly used method for parameter estimation is the maximum likelihood method (ML). However, this method applied on the three-parameter Weibull distribution returns biased estimates and the amount of bias is unknown. A recent method, that we call nested maximum likelihood, was proposed by Gourdin, Hansen and Jaumard (1994). Due to its complexity, it has never been used and tested systematically. Here I compare it to the maximum likelihood method. The results shows that nested maximum likelihood is slightly better than ML. Although the gains are marginal, the method has important implications for future research in parameter estimation.

Statistical methods in psychology are dominated by the normal distribution. However, very few measures in experimental psychology follow this distribution. For example, the time to complete a task (maybe the most direct access to cognitive processes) is always asymmetrical with a long tail to the right (e. g. Cousineau and Shiffrin, 2004, Hockley, 1984, with an exception, Hopkins and Kristofferson, 1980). Hence, it is of prime importance that we move towards a description of the response times (RTs) that acknowledge this asymmetry.

The most natural such description assumes that there is a true minimum RT below which it is not possible to respond (fast-guessing notwithstanding). Then, three convenient descriptors could be: the lowest possible RT, the width of the distribution and the degree of asymmetry. See Rouder, Lu, Speckman, Sun and Jiang (2005) for reasons supporting this choice. Whereas the width is akin to standard deviation, there is nothing resembling a mean parameter in the above descriptors, highlighting the conceptual gap with the normal distribution. Figure 1 illustrates two distributions from Cousineau and Larochelle (2004) with the corresponding descriptors.

A parametric approach for quantifying parameters consists in first assuming an underlying theoretical distribution and then adjusting its parameters to the data set through best-fitting techniques.

Theoretical distributions

There exist many families of distributions that could be fit to a data set in order to get parameters (Luce, 1986, Townsend and Ashby, 1983). One of the most often-used distribution in psychology is the ExGaussian distribution (e.g. Ratcliff, 1978, Hohle, 1965). However, it has the implausible assumption that valid RTs could occurs before the stimulus (the Gumbel distribution has the same assumption). An alternative distribution is the Lognormal distribution (also called the Galton distribution; Limpton, Stahel and Appt, 2001, West and Schlesinger, 1990, Galton, 1879). However, more and more, the Weibull distribution is used (Weibull, 1952, Logan, 1992, Palmer, 1998, Tuerlinckx, 2004, and many others). Rouder, Lu, Speckman, Sun and Jiang (2005) review practical reasons to use this family of distributions. Also, Cousineau, Goodman and Shiffrin (2003) suggest that it could be a consequence of the way the human

We would like to thank Sébastien Hélie for his comments on an earlier version of this text. Request for reprint should be addressed to Denis Cousineau, Département de psychologie, Université de Montréal, C. P. 6128, succ. Centre-ville, Montréal (Québec) H3C 3J7, CANADA, or using e-mail at Denis.Cousineau@Umontreal.CA. This research was supported in part by le Conseil de la Recherche en Sciences Naturelles et en Génie du Canada.

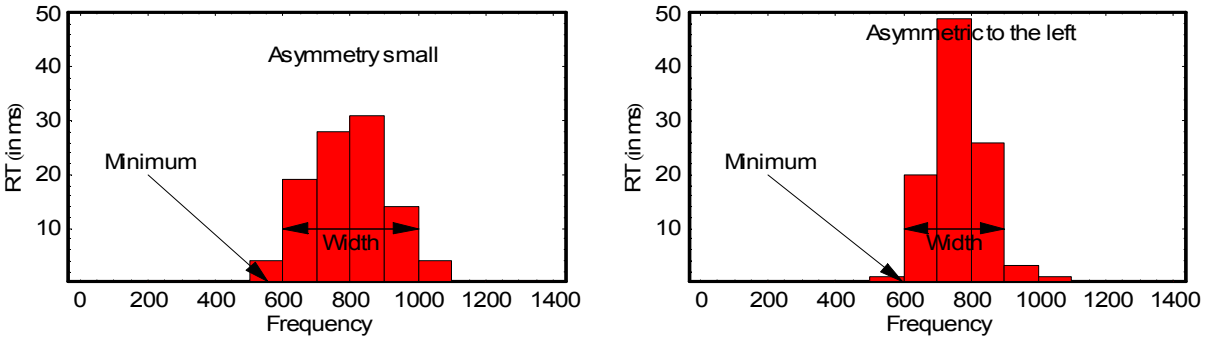

Figure 1. Two distributions of RTs with an illustration of the parameters describing them. The distribution to the right is shifted towards large RTs, has smaller width and is more asymmetrical than the one on the left.

brain works. Figure 2 illustrates some possible Weibull distributions along with their parameters. The parameters are the shift (the left-right position of the minimum), the scale (the width) and the shape (the asymmetry). They are often denoted with the greek letters α, β and γ respectively.

Fitting techniques

There exist a few families of techniques to find the best-fitting values of the parameters. One is the method of moment (e. g. Harter and Moore, 1965), another is through Bayesian estimation techniques (e. g. Rouder, Sun, Speckman, Lu and Zhou, 2003), but the most commonly used method is the maximum likelihood (ML) parameter estimation method. Refer to Myung (2003) for a tutorial or Cousineau and Larochelle (1997), Cousineau, Brown and Heathcote (2004). It requires a function computing the likelihood of one possible set of parameters given empirical RTs (noted X). This function is noted $L(\alpha, \beta, \gamma \mid X)$. This function is subjected to a maximization procedure which varies freely the parameter values until the likelihood is maximized. Often, minus the likelihood function is minimized, as minimization procedures used to be more easily available. Also, to avoid underflow on most computers, the log of the likelihood function is used. Hence, the process is to find α, β and γ such that minus the log of the likelihood is minimized, noted in short:

$$\underset{\substack{\alpha\in A\\ \beta\in B\\ \gamma\in\Gamma}}{Min} - Log\left(L(\alpha,\beta,\gamma|X)\right)$$

A, B and Γ are the domains of the parameters. For the Weibull distribution, $A = \{-\infty < \alpha < \text{Min}(X)\}$, $B = \{\beta > 0\}$, $\Gamma = \{\gamma > 0\}$. In practical application, it is preferable to use $\Gamma = \{0 < \gamma < 5\}$ as RT distributions are never asymmetrical to the right.

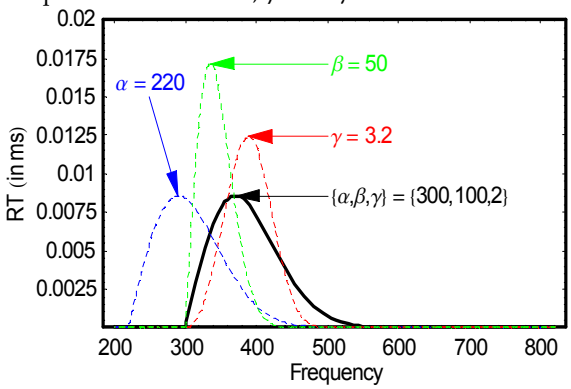

Figure 2. Examples of Weibull distributions. The thin line distributions differ from the thick line distribution by only one parameter.

Why another method?

ML is the most efficient method to estimate parameters (see next for a formal definition of efficiency). However, it is also known to be biased (Hirose, 1999): On average, the parameter estimated is not going to be equal to the true parameter of a given population. The bias can be quite large for small sample sizes. For example, for a sample of 8 RTs (taken from a simulated population), the scale parameter is underestimated on average by near 40%! For larger samples, the bias tends to disappear (asymptotically unbiased). The trouble is that the exact amount of bias is unknown. One consequence is that it is not possible to compare parameters taken from samples differing in size. Also, we don't know whether bias depends on the asymmetry or not.

This is why new techniques may potentially be important: They may find estimates with smaller bias. Previous variations on the ML methods are MPS (Cheng and Amin, 1983), QMP (Heathcote, Brown and Cousineau, 2004) and prior-informed ML (Cousineau, submitted).

The nested maximum likelihood technique

This technique was proposed by Gourdin, Hansen and Jaumard a decade ago (1994). However, due to the complexity of implementing this method within traditional computer languages, it has never been used. Further, the authors never tested their method on samples taken from simulated populations with known parameters so that the amount of bias be estimated.

The method differs from regular ML in that it does not

fit all three parameters simultaneously. Instead, it explores one parameter, adjusting the other two so that this parameter yields the best fit possible. To adjust the two "inner" parameters, it just acts likewise: exploring one parameter, the last one being best-fitted accordingly. Hence, this technique replaces one difficult minimization problem with three simpler (one-dimensional) nested problems.

Replacing $-\text{Log}(L(\theta \mid X))$ with $\text{LL}(\theta)$ for short, the method is formally given by:

$$\text{Best fitting } L = \underset{\alpha \in \text{A}}{Min}\, LL_3(\alpha)$$
$$\text{in which } \text{LL}_3(\alpha) = \underset{\gamma \in \Gamma}{Min}\, LL_2(\alpha, \gamma)$$
$$\text{in which } \text{LL}_2(\alpha, \gamma) = \underset{\beta \in \text{B}}{Min}\, LL_1(\alpha, \beta, \gamma)$$

Every time a new α value is tried at the top level (3), it starts a cascade of computations going down to the bottom level (1).

Gourdin et al. (1994) were able to demonstrate that this nested ML method has all the properties of the regular ML method (efficiency, asymptotically unbiased when n is large). However, for a small n, we don't know if bias is smaller with this technique compared to other techniques.

The pseudo-code of the implementation proposed by Gourdin et al. spans two whole pages; the finished program certainly required over 40 pages in C. Here, we propose the same procedure implemented using *Mathematica* which takes only one page (Wolfram, 1996). See Listing 1 for the full program.

Testing nested ML estimates against regular ML estimates

To assess the capabilities of nested ML to estimate correctly the parameters, we ran a series of Monte Carlo simulations. The samples are taken from a simulated population with known parameters and then the parameters were estimated using both methods. Because α and β are scale parameters, they were fixed at $\alpha = 300$ and $\beta = 100$. the shape, being one possible cause of bias, was varied ($\gamma = 1.0$, 1.5 or 2.0) as well as the sample size (n = 8, 16, 32, 64, 128). Simulations for each combination of γ and n were replicated a thousand times. We measured:

1- Bias: the difference between the true parameters and the mean of the estimates. Formally, $Bias = E(\theta_i - \theta_T) = \overline{\theta} - \theta_T$ where θ is one of the parameter, θ_T is the true parameter value, θ_i is the estimated parameter on simulation i, and E is the mean.

2- Efficiency: the amount of variability in the estimates around their average. Bad methods have very large efficiency so that fitting one data set can result in wildly differing estimates. Formally, $Eff = SD(\theta_i - \overline{\theta}) = SD(\theta_i)$ where SD is the standard deviation.

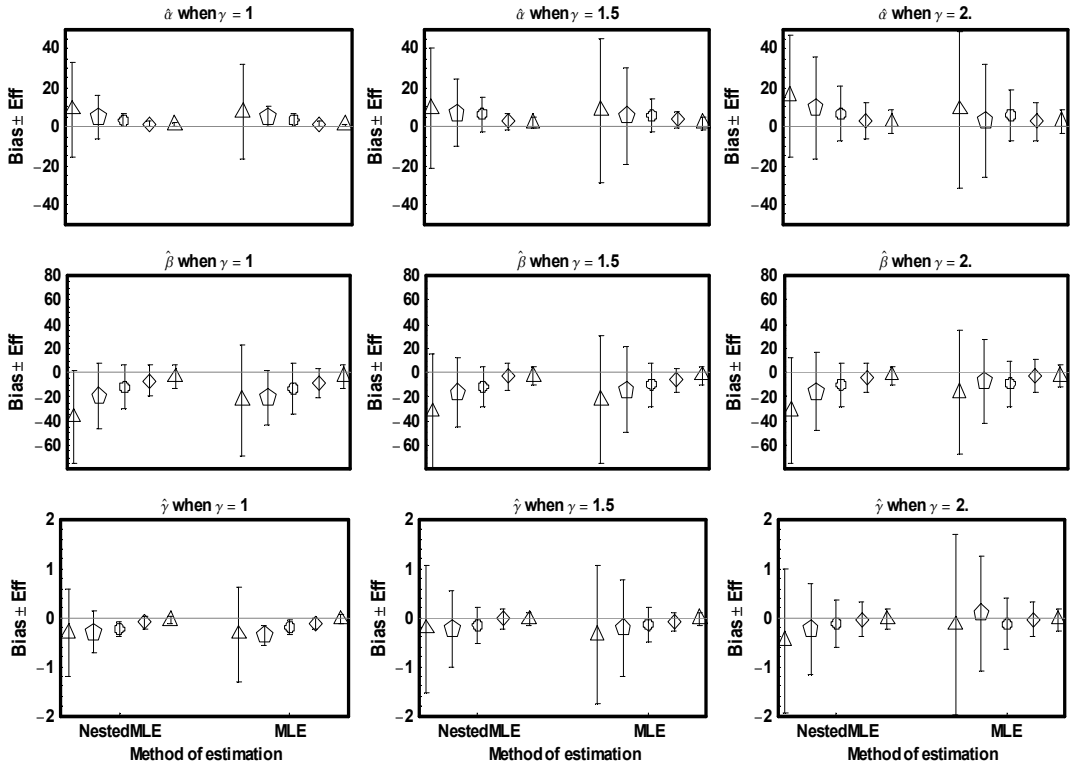

Figure 3. Bias and efficiency for all the combinations of Γ (from left to right) of the estimated parameters (from top to bottom). In each graph, the left part is the nested ML estimation technique and the right part is the regular ML estimation technique. The symbols denotes the various samples sizes (from left to right: 8, 16, 32, 64 and 128).

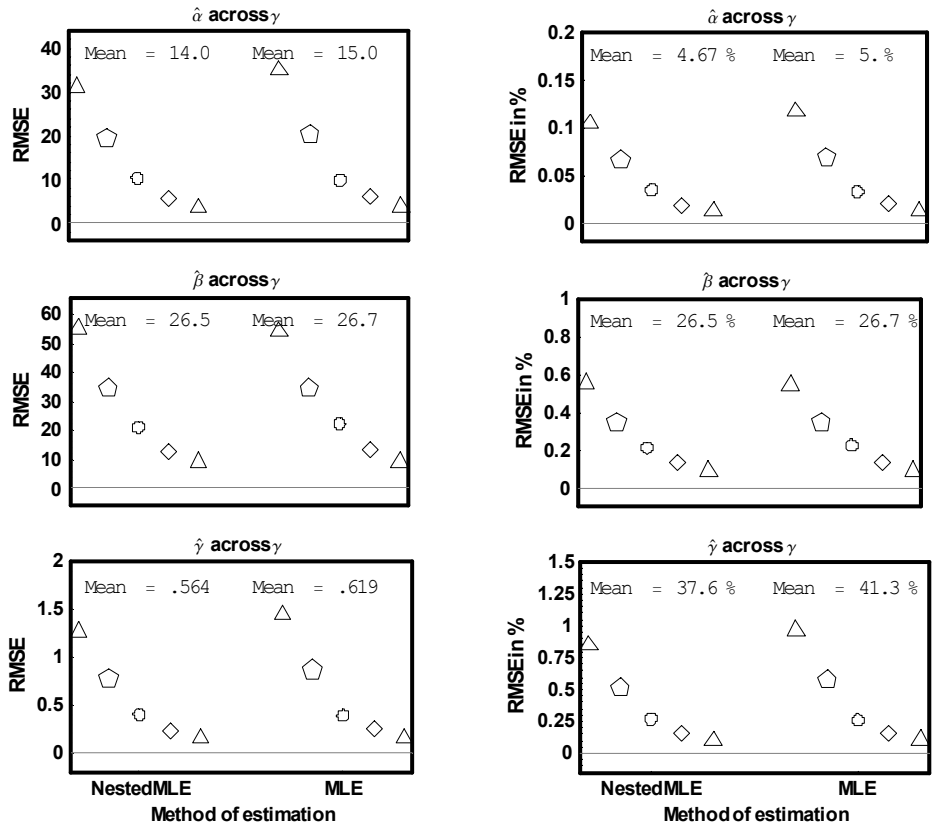

Figure 4. RMSE for the three parameters across Υ. Format follows that of Figure 4. On the left is RMSE, on the right, RMSE expressed in percentage of the true parameter value.

Both methods used the gradient descent minimization algorithm (Chandler, 1965) and both had the same parameter domains: A = {0 < α < Min(*X*) }, B = {β > 0}, Γ= {0 < γ < 5}.

The results are seen in Figure 3. As seen, both methods are biased in the same directions (shift is underestimated, scale and shape are overestimated) and by the same magnitude. Whereas bias is little dependant on the shape parameter, it seems that efficiency is, being worst for γ = 2.

The same results are presented in Figure 4 collapsed across the γ values. In this figure, we used the root mean square error of estimation (*RMSE*) where

$$RMSE = E\{ (\theta - \theta_T)^2 \}$$

It can be shown that $RMSE^2 = Bias^2 + Eff^2$. It is a good measure when both bias and efficiency are equally important. As seen, the parameter α is the best estimated parameter (in percent, RMSE are 4.7% and 5.0% across sample sizes for nested ML and ML respectively). The other two parameters are poorly estimated (in percent, near 27% and 39% for both methods). There is systematically a small advantage of nested MLE over ML, but the gain is very small.

Discussion

Nested ML is definitely not the solution to adopt for practical applications. So why bother? The demonstrations that accompany the method have profound implications for future research. It shows that the problem of parameter estimation can be broken down in encapsulated problems that can be attacked independently. Among other things, it opens the door to mixed solutions. For example, one parameter might be estimated using another technique than ML. So doing will not yield an efficient method (as ML are generally the most efficient) but if bias can be reduced so that global *RMSE* will not deteriorate, it is going to be an important progress.

References

Chandler, P. J. (1965). Subroutine STEPIT: An algorithm that finds the values of the parameters which minimize a given continuous function [computer program]. *Bloomington: Indiana University, Quantum chemistry*.

Cheng, R. C. H. & Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society B, 45*, 394-403.

Cousineau, D. & Larochelle, S. (1997). PASTIS: A Program for Curve and Distribution Analyses. *Behavior Research Methods, Instruments, & Computers, 29*, 542-548.

Cousineau, D. & Larochelle, S. (2004). *Visual-Memory search: An integrative perspective. Psychological Research, 69*, 77-

105.

Cousineau, D., & Shiffrin, R. M. (2004). *Termination of a visual search with large display size effect*. *Spatial Vision, 17*, 327-352.

Cousineau, D., Brown, S., & Heathcote, A. (2004). Fitting distributions using maximum likelihood: Methods and packages. Behavior *Research Methods, Instruments, & Computers, 36*, 742-756.

Cousineau, D., Goodman, V. & Shiffrin, R. M. (2002). Extending statistics of extremes to distributions varying on position and scale, and implication for race models. *Journal of Mathematical Psychology, 46*, 431-454.

Galton, F. (1879). The geometric mean in vital and social statistics. *Proceedings of the Royal Society of London, 29*, 365-367.

Gourdin, E., Hansen, P., & Jaumard, B. (1994). Finding maximum likelihood estimators for the three-parameter Weibull distribution. *Journal of Global Optimization, 5*, 373-397.

Harter, H. L. & Moore, H. (1965). Maximum likelihood estimation of the parameters of Gamma and Weibull populations from complete and from censored samples. *Technometrics, 7*, 639-643.

Heathcote, A., Brown, S., & Cousineau, D. (2004). QMPE: Estimating Lognormal, Wald and Weibull RT distributions with a parameter dependent lower bound. *Behavior Research Methods, Instruments, & Computers, 36*, 277-290.

Hirose, H. (1999). Bias correction for the maximum-likelihood estimates in the two-parameter Weibull distribution. *IEEE Transactions on Dielectrics and Electrical Insulation, 6*, 66-68.

Hockley, W. E. (1984). Analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*, 598-615.

Hohle, R. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology, 69*, 382-386.

Hopkins, G. W. & Kristofferson, A. B. (1980). Ultrastable stimulus-reponse latencies: Acquisition and stimulus control. *Perception and Psychophysics, 27*, 241-250.

Limpert, E., Stahel, W. A. & Abbt, M. (2001). Log-normal distributions across the sciences: *Keys and clues. BioScience, 51*, 341-352.

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: a test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*, 883-914.

Luce, R. D. (1986). Response times, their role in inferring elementary mental organization. *New York: Oxford University Press*.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47*, 90-100.

Palmer, J. (1998). Attentional effects in visual search: relating search accuracy and search time, in *Richard D. Wright (eds.). Visual attention* (pp. 348-388). *New York: Oxford University Press*.

Ratcliff, R. (1979). Group Reaction Time Distributions and an Analysis of Distribution Statistics. *Psychological Bulletin, 86*, 446-461.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review, 12*, 195-223.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical bayesian statistical framework for response time distributions. *Psychometrika, 68*, 589-606.

Townsend, J. T. & Ashby, F. G. (1983). Stochastic Modeling of Elementary Psychological Processes. *Cambridge, England: Cambridge University Press*.

Tuerlinckx, F. (2004). A multivariate counting process with Weibull-distributed first-arrival times. *Journal of Mathematical Psychology, 48*, 65-79.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanic, 18*, 292-297.

West, J. & Shlesinger, M. (1990). The noise in natural phenomena. *American Scientist, 78*, 40-45.

Wolfram, S. (1996). *The Mathematica Book (third edition). New York: Cambridge University Press*.

Manuscript received November 15th, 2006

Listing 1 on next page

1- Define the pdf and the loglikelihood functions

```
f[{γ_, β_, α_}, x_] = e^-((x-α)/β)^γ (x - α)^(-1+γ) β^-γ γ;

ll[{γ_, β_, α_}, data_] := - Sum[Log[f[{γ, β, α}, data[[i]]]], {i, 1, Length[data]}]

BadFit = 10^6;
```

2- Define the nested ML method with its three nested levels

```
Fit1[x_, {c_, a_}] := Module[{b1, res1},
  res1 = FindMinimum[
     (If[b1 > 0, ll[{c, b1, a}, x], BadFit]),
     {b1, 80, 120},
     MaxIterations → 500
    ][[2]];
  {b → b1 /. res1}
 ]

Fit2[x_, {a_}] := Module[{c2, res2},
  res2 = FindMinimum[
     ( If[0 < c2 < 5, ll[{c2, b /. Fit1[x, {c2, a}], a}, x], BadFit]),
     {c2, 0.5, 2.5},
     MaxIterations → 500
    ][[2]];
  {c → c2 /. res2, b → (b /. Fit1[x, {c2 /. res2, a}])}
 ]

FitNestedMLE[x_] := Module[{a3, res3},
  res3 = FindMinimum[
     (If[0 < a3 < Min[x], ll[{c /. Fit2[x, {a3}], b /. Fit2[x, {a3}], a3}, x], BadFit]),
     {a3, 280, 300},
     MaxIterations → 500
    ][[2]];
  sol = {c → (c /. Fit2[x, {a3 /. res3}]), b → (b /. Fit2[x, {a3 /. res3}]), a → a3 /. res3};
  {ll[{c, b, a} /. sol, x], sol}
 ]
```

3- One test: Import and fit a data set

```
mypath = "C:\\Documents and Settings\\cousined\\Bureau\\34-NestedMLE\\Listing\\";
sample = Import[mypath <> "data.dat", "List"];
FitNestedMLE[sample]
```

```
{334.562, {c → 1.57777, b → 87.3595, a → 311.768}}
```