

# Fitting distributions using maximum likelihood: Methods and packages

DENIS COUSINEAU

*Université de Montréal, Montréal, Québec, Canada*

SCOTT BROWN

*University of California, Irvine, California*

and

ANDREW HEATHCOTE

*University of Newcastle, Callaghan, New South Wales, Australia*

The most powerful tests of response time (RT) models often involve the whole shape of the RT distribution, thus avoiding mimicking that can occur at the level of RT means and variances. Nonparametric distribution estimation is, in principle, the most appropriate approach, but such estimators are sometimes difficult to obtain. On the other hand, distribution fitting, given an algebraic function, is both easy and compact. We review the general approach to performing distribution fitting with maximum likelihood (ML) and a method based on quantiles (quantile maximum probability, QMP). We show that QMP has both small bias and good efficiency when used with common distribution functions (the ex-Gaussian, Gumbel, lognormal, Wald, and Weibull distributions). In addition, we review some software packages performing ML (PASTIS, QMPE, DISFIT, and MATHEMATICA) and compare their results. In general, the differences between packages have little influence on the optimal solution found, but the form of the distribution function has: Both the lognormal and the Wald distributions have nonlinear dependencies between the parameter estimates that tend to increase the overall bias in parameter recovery and to decrease efficiency. We conclude by laying out a few pointers on how to relate descriptive models of RT to cognitive models of RT. A program that generated the random deviates used in our studies may be downloaded from [www.psychonomic.org/archive/](http://www.psychonomic.org/archive/).

Since the seminal work of Townsend and Ashby (1983), it has been known that fitting or testing a model with mean response times (RT) alone has very poor diagnostic power. Often, models can mimic each other at the level of predicted means, even when their fundamental assumptions are diametrically opposed (e.g., a parallel race model can mimic the predictions of a serial model; see Van Zandt & Ratcliff, 1995). In this respect, median RT does not fare better than mean RT (Miller, 1988; Ratcliff, 1993). One solution is to consider RT means and variances simultaneously (Cousineau & Larochelle, in press). Although this provides greater constraint, some model mimicking can still occur (Townsend & Colonius, 2001). Higher order moments (e.g., skew and kurtosis) are of little help because their sample estimates are unreliable for the sample sizes typically available in empirical research. As a result,

the importance of considering the whole RT distribution for testing formal models is now generally acknowledged.

Nonparametric approaches to the description of RT distributions are possible. For example, estimating the cumulative distribution function (CDF) is easily achieved with the cumulative frequencies of observed RTs. However, estimating the probability density function (PDF) and the hazard function is more difficult (see Silverman, 1986, on the former, and Bloxom, 1984, on the latter). This is a problem because some models are most easily tested with nonparametric approaches (e.g., tests of the hazard function, Burbeck & Luce, 1982, and the crossing points of two PDFs, Ashby, Tein, & Balakrishnan, 1993).

A parametric approach to RT distribution is achieved by introducing an important piece of information: a density or cumulative density function of the distribution. As we will discuss in the next section, fitting a distribution is rather easy, and there are many software packages that can automate this procedure.<sup>1</sup> In addition, the estimation method used, maximization of the likelihood function, is well understood and is not dependent upon the use of approximate heuristics (as opposed to nonparametric PDF and hazard function estimates; see Silverman, 1986). Also, once the distribution has been fitted, all associated functions (CDF, PDF, hazard, and log-survivor functions)

---

This research was supported in part by the Fonds pour la formation de Chercheurs et l'Aide à la Recherche and the Conseil de Recherches en Sciences Naturelles et en Génie du Canada. We thank C. Dolan, G. Francis, G. Giguère, and S. Hélie for their comments on an earlier version of this article. Correspondence should be addressed to D. Cousineau, Département de psychologie, Université de Montréal, C. P. 6128, succ. Centre-ville, Montréal, PQ H3C 3J7, Canada (e-mail: [denis.cousineau@umontreal.ca](mailto:denis.cousineau@umontreal.ca)).

are completely determined. Finally, the fitting process consists simply of finding estimated values for a few parameters (generally three for RT distributions). Thus, the whole RT distribution is summarized with a very compact representation.

All these benefits come at a cost, however. An incorrect distribution function, even one fitting the data reasonably well, may give a wrong indication of what kind of psychological model has produced the data. For this reason, many authors prefer to use distribution functions as an atheoretical tool or a descriptive model (Heathcote, Popiel, & Mewhort, 1991; Ratcliff, 1979). In addition, if the true RT distribution is in fact different from the fitted distribution in some fundamental way, the parameters may not capture the regularities that exist across different distributions (Schwarz, 2001). For that reason, it is desirable for experimenters using the parametric approach to fit more than one distribution function.

In this article, we review software that can perform distribution fitting. All the software packages reviewed can fit many distinct distribution functions. The most commonly used distributions in cognitive psychology are the ex-Gaussian (Hockley, 1984), the Gumbel (Gumbel, 1958; Yellot, 1977), the lognormal (Ulrich & Miller, 1993), the Wald (Burbeck & Luce, 1982), and the Weibull (Cousineau, Goodman, & Shiffrin, 2002) distributions (see Heathcote, Brown, & Cousineau, 2004, and Luce, 1986, Appendix A, for details).<sup>2</sup> The software reviewed can all fit these distributions, although some can fit others, as well. They are PASTIS (Cousineau & Larochelle, 1997), QMPE (previously called QMLE; Brown & Heathcote, 2003), DISFIT (Dolan, van der Maas, & Molenaar, 2002), and MATHEMATICA (Wolfram, 1996). This review will be carried out in Section 2. Because the methods and the specific details of a fitting procedure are numerous, we provide in Section 1 some information to readers interested in programming their own fitting procedure. Although many readers will prefer to rely on existing software, these details are useful to know, because they can differ from one package to another.

## 1. DISTRIBUTION FITTING METHODS

### Estimation Methods

On one hand, there is a data set  $\mathbf{T} = \{t_i\}$ ,  $i = 1 \dots n$ , a sample containing  $n$  response times (RT). On the other hand, there is a distribution to fit that depends on a parameter set  $\theta$ . The distribution is given by its probability density function  $f$ . The objective of the fitting procedure is to find the estimated parameters  $\hat{\theta}$  so that the theoretical distribution will be most similar to the distribution of the data set.

Many methods can be used to fit a distribution. Van Zandt (2000) reviewed sum of square error (*SSE*) methods and the maximum likelihood (ML) method (presented next). Using simulations, she found the standard ML method to be the best and *SSE* based on CDF almost as good. The criteria used were (1) bias—repeated over mul-

iple samples, the average  $\hat{\theta}$  should be exactly the true parameter set  $\theta$  of the population where the samples were taken; and (2) efficiency—when repeated over different samples of data, the estimates  $\hat{\theta}$  should have smaller variance than when estimated with other methods.

We will concentrate on the ML method. This method searches for a set of parameters  $\hat{\theta}$  that maximizes the likelihood of observing such data (Hays, 1973). The likelihood of the data set  $\mathbf{T}$  is the joint probability of the sample for a given model and set of parameters,  $\Pr(\{RT\} = \mathbf{T})$ . When the observations are independent and measured with infinite precision, the probability of the sample becomes

$$\Pr(\{RT\} = \mathbf{T}) = \Pr(RT_1 = t_1 \cap RT_2 = t_2 \cap RT_3 = t_3 \dots \cap RT_n = t_n) \\ = \prod_{i=1}^n \Pr(RT_i = t_i).$$

For a given parameter set  $\hat{\theta}$  and a PDF  $f$ , and assuming the sample to contain independent deviates from this distribution, we may write

$$L(\hat{\theta}, \mathbf{T}) = \prod_{i=1}^n f(t_i | \hat{\theta}). \quad (1)$$

The function  $L$ , called the likelihood function, is a measure of how likely the data set is, given  $\hat{\theta}$ . The larger it is, the more likely  $\hat{\theta}$  is the true  $\theta$ .<sup>3</sup>

Occasionally, Equation 1 can be solved analytically. For example, we know that the sample mean is a maximum likelihood estimator of the population mean  $\mu$  if the distribution is normal (Gaussian). Maximum likelihood estimators, whether analytic or obtained from numerical optimization, have a desirable property: They are asymptotically ( $n \rightarrow \infty$ ) the most efficient; that is, they make maximum use of the information contained in the sample, resulting in the least variable estimation method (Van Zandt, 2000). Of course, it is not clear whether such asymptotic property holds for small samples as well (Heathcote, Brown, & Cousineau, 2004).

One limitation of Equation 1 is that it assumes infinite precision measurements. In practice, this is never achieved. In some RT experiments, precision is within 1 msec ( $\pm 0.5$  msec), sometimes more (e.g., for responses collected through a mouse,  $\pm 6$  msec, see Beringer, 1992). For data measured with precision  $2\epsilon$ , the exact probability of observing the response time  $t_i$  is given by

$$\Pr(RT_i = t_i \pm \epsilon | \hat{\theta}) = \int_{t_i - \epsilon}^{t_i + \epsilon} f(t | \hat{\theta}) dt.$$

We can introduce a rectangular approximation by using

$$\Pr(RT_i = t_i \pm \epsilon | \hat{\theta}) \approx 2\epsilon \times f(t_i | \hat{\theta})$$

so that Equation 1 becomes

$$L(\hat{\theta}, \mathbf{T}) \approx (2\epsilon)^n \prod_{i=1}^n f(t_i | \hat{\theta}) \\ \propto \prod_{i=1}^n f(t_i | \hat{\theta}). \quad (2)$$

Because the  $(2\epsilon)^n$  term is generally ignored, numerical values of  $L(\hat{\theta}, \mathbf{T})$  cannot be compared between appara-

tuses that do not have the same accuracy. However, it has no influence on the maximum likelihood estimates  $\hat{\theta}$  obtained because the term  $(2\varepsilon)^n$  is independent of  $\theta$  and may be ignored. The method based on the continuous approximation to the likelihood function of raw data (Equation 2), denoted  $L(\hat{\theta}, \mathbf{T})$ , will be called CML.

An alternative estimation method, quantile maximum probability (QMP), was introduced by Heathcote, Brown, and Mewhort (2000). First, a set of raw data  $\mathbf{T}$  is grouped into  $m$  categories described by the location of the  $m + 1$  bounds,  $\mathbf{Q}$ , called quantiles ( $\hat{q}_j, j = 0 \dots m \leq n$ ). The quantiles are calculated from the order statistics of  $\mathbf{T}$  (see Heathcote et al., 2000, for details), with the number of observations falling into the interval  $[\hat{q}_{j-1}, q_j]$  denoted  $n_j$ . In general, the  $\hat{q}_j$  are chosen so that there are approximately the same number of observations in each quantile, but this need not always be true. Furthermore, the bounds  $\hat{q}_0$  and  $\hat{q}_m$  are set at the limit of the fitted distribution, sometimes  $-\infty$  and  $+\infty$ . Estimates are obtained by choosing parameters that maximize the product of the probability in each category:

$$\prod_{j=1}^m \left( \int_{\hat{q}_{j-1}}^{\hat{q}_j} f(t|\hat{\theta}) dt \right)^{n_j}. \quad (3)$$

One may intuitively expect QML to be inefficient, since transforming the raw RT into quantiles involves a reduction of the information available. On the other hand, if  $m$  is close to  $n$ , the loss of information can be quite small and may also provide a benefit for finite samples. This is because there may be outliers in the data, and creating quantiles simply replaces the absolute value of an outlying observation with an additional count in  $n_j$  or  $n_m$ . Also, estimates are robust to the addition of a small amount of error to a given RT, because Equation 3 will not change at all, as long as the RT does not move across a quantile bound.

Heathcote et al. (2000) showed that QML (Equation 3) is superior to CML (Equation 2) in terms of bias and efficiency when tested on simulated data generated by the ex-Gaussian distribution; Heathcote et al. (2004) showed QML to be equal to or better than CML on other distributions. In order to do so, they extended a software package called QMPE to include the lognormal, Gumbel, Wald, and Weibull distributions with both CML and QMP estimation. Before proceeding to comparisons across software, we will discuss some issues related to the implementation of a maximum likelihood fitting technique because the software reviewed in Section 2 differ on these implementation details.

### Implementing a Maximum Likelihood Fitting Procedure

In order to implement an ML procedure, three ingredients are required: (1) a distribution function to be fitted, (2) an optimization routine, and (3) starting values for  $\hat{\theta}$ .

To be a reasonable candidate for characterizing RT, a distribution function must be able to accommodate pos-

itively skewed data. The most commonly used distributions are briefly described in Heathcote et al. (2004), and their equations are presented in Table 1. The choice of a distribution provides the PDF equation  $f$  that is inserted into the function to be minimized (Equation 2 or 3), which in this context is usually called the *objective function*. Due to numerical considerations, the logarithm of the objective function is usually employed because the use of the summation avoids numerical underflows. For likelihood, for example, maximizing  $L(\hat{\theta}, \mathbf{T})$  over the range 0 to 1 is then replaced by minimizing  $-\ln L(\hat{\theta}, \mathbf{T})$ , the value ranging from  $+\infty$  [ $-\ln(0)$ , unlikely] to zero [ $\ln(1)$ , absolutely certain]. The same rationale applies if quantiles  $\mathbf{Q}$  are used instead of raw data  $\mathbf{T}$ . Most computer software packages report the minimized value of  $-\ln L(\hat{\theta}, \mathbf{T})$ .

The second ingredient is an optimization procedure to minimize the objective function. Various algorithms exist, the oldest of which was introduced by Newton. All these methods are iterative, starting with a tentative  $\hat{\theta}_0$  and updating it through various iterations until an optimal value,  $\hat{\theta}_p$ , is found.<sup>4</sup>

Appendix A summarizes the most commonly used minimization algorithms. They are generally distinguished (Box, Davies, & Swann, 1969) by whether they use analytic derivatives of the objective function (gradient methods) to guide search, or whether they use numerical approximations to the derivative (direct search methods). In general, gradient methods can find a minimum with a smaller number of iterations. However, it may take more time to perform the iterations if the gradients are not available in closed-form equations (as is the case for the ex-Gaussian distribution).

The last ingredient in obtaining a solution consists of finding reasonable starting values  $\hat{\theta}_0$ . If the surface of the objective function is quadratic, it has only one minimum, and thus, all starting points will lead the minimization routine to the same optimal solution  $\hat{\theta}_p$ . In practice, however, there may be many local minima. The best way to avoid false convergence in a local minimum is to start the routine at various locations or to start as close as possible to the optimal solution. To achieve this, heuristic estimates can be developed, often based on the first few moments of the data in order to automate the starting point selection. These heuristics are not always accurate, due to sampling variance in the moment estimates.

## 2. TESTING THE SOFTWARE PACKAGES

We compare different software packages aimed at fitting distributions. These packages, briefly described in Appendix B, differ in the minimization routines used and in the heuristics used for starting values. All of these packages allow the user to alter the starting value parameters. Table 2 reviews some features of the software, and Appendix C shows examples of commands for a typical fitting session with each.

Table 1  
Overview of the Distribution Functions

Distribution	Probability Density Function	Moments
ex-Gaussian $\theta = \{\mu, \sigma, \tau\}$	$F(RT) = \Phi\left(\frac{RT - \mu}{\sigma}\right) - \text{Exp}\left(\frac{\sigma^2}{2\tau^2} - \frac{RT - \mu}{\tau}\right) \Phi\left(\frac{RT - \mu - \sigma}{\sigma}\right)$ $f(RT) = \frac{1}{\tau} e^{-\frac{\mu - RT + \sigma^2}{2\tau^2}} \Phi\left(\frac{RT - \mu - \sigma}{\sigma}\right) + \frac{1}{\sigma} e^{-\frac{RT - \mu}{\sigma}}$	$E(RT) = \mu + \tau$ $\text{Var}(RT) = \sigma^2 + \tau^2$ $\text{Skew}(RT) = \frac{2\tau^3}{(\sigma^2 + \tau^2)^{3/2}}$
Gumbel $\theta = \{\mu, \sigma\}$	$F(RT) = e^{-e^{-\frac{RT - \mu}{\sigma}}}$ $f(RT) = \frac{1}{\sigma} e^{-e^{-\frac{RT - \mu}{\sigma}}} - \frac{RT - \mu}{\sigma} e^{-e^{-\frac{RT - \mu}{\sigma}}}$	$E(RT) = \psi(1)\sigma + \mu$ $\text{Var}(RT) = \frac{\sigma^2 \pi^2}{6}$ $\text{Skew}(RT) = \frac{-6\sqrt{6}}{\pi^3} \psi(2,1) \approx 1.13955$
Lognormal $\theta = \{\kappa, \mu, \sigma\}$	$F(RT) = \Phi\left[\frac{\log(RT - \kappa) - \mu}{\sqrt{2}\sigma}\right]$ $f(RT) = \frac{e^{-[\log(RT - \kappa) - \mu]^2 / (2\sigma^2)}}{\sqrt{2\pi\sigma(RT - \kappa)}}$	$E(RT) = e^{\frac{\sigma^2}{2} + \mu}$ $\text{Var}(RT) = e^{\sigma^2 + 2\mu} (e^{\sigma^2} - 1)$ $\text{Skew}(RT) = \sqrt{e^{\sigma^2} - 1} (e^{\sigma^2} + 2)$
Wald $\theta = \{\kappa, \mu, \lambda\}$	$F(RT) = \Phi\left[\frac{\sqrt{\lambda}}{\sqrt{RT - \kappa}} \left(\frac{RT - \kappa}{\mu} - 1\right)\right] - e^{-\mu} \Phi\left[-\sqrt{\frac{\lambda}{RT - \kappa}} \left(\frac{RT - \kappa}{\mu} + 1\right)\right]$ $f(RT) = \sqrt{\frac{\lambda}{2\pi(RT - \kappa)^3}} \text{Exp}\left[-\frac{\lambda(RT - \mu - \kappa)^2}{2\mu^2(RT - \kappa)}\right]$	$E(RT) = \kappa + \mu$ $\text{Var}(RT) = \mu^2 / \lambda$ $\text{Skew}(RT) = 3\sqrt{\frac{\mu}{\lambda}}$
Weibull $\theta = \{\alpha, \beta, \gamma\}$	$F(RT) = 1 - e^{-\left(\frac{RT - \alpha}{\beta}\right)^\gamma}$ $f(RT) = \beta^{-\gamma} \gamma (RT - \alpha)^{\gamma-1} e^{-\left(\frac{RT - \alpha}{\beta}\right)^\gamma}$	$E(RT) = \beta \Gamma(1 + 1/\gamma) + \alpha$ $\text{Var}(RT) = \beta^2 \left[ \Gamma(1 + 2/\gamma) - \Gamma(1 + 1/\gamma)^2 \right]$ $\text{Skew}(RT) = \frac{2\Gamma^3(1 + 1/\gamma) - 3\Gamma(1 + 1/\gamma)\Gamma(1 + 2/\gamma) + \Gamma(1 + 3/\gamma)}{\left[ \Gamma(1 + 2/\gamma) - \Gamma(1 + 1/\gamma)^2 \right]^{3/2}}$

Note— $E(RT)$  is the expected (mean) value of the distribution.  $\text{Var}(RT)$ , its variance, and  $\text{Skew}(RT)$  is the Fisher measure of skewness, given by

$$\frac{\frac{1}{n} \sum (x_i - \bar{x})^3}{\text{Var}(x)^{3/2}}$$

$\Phi$  is the standard normal integral, given by

$$\Phi(x) = \int_{-\infty}^x e^{-y^2/2} dy$$

and implemented on many computer languages with the expression:

$$\Phi(x) = \frac{1}{2} [1 + \text{Erf}(x)].$$

$\psi(1) = 0.577216$  is the Euler constant, and  $\psi(2,1) = -2.40411$  is the digamma function.  $\Gamma(x)$  is the gamma function; for  $x \in \text{integer}$ ,  $\Gamma(x + 1) = x!$

**Table 2**  
**Relevant Features of the Software Packages Tested**

Software	Optimization Procedure	Warning on Unsuccessful Fit	File Format Flexible	Starting Point Suggested	Standard Error Provided	Fitting Other Distributions Requires:
PASTIS	Quasi-Newton (Chandler, 1965)	no	yes	yes	no	programming in C
QMPE	Conjugate gradient (Press et al., 1986)	yes	no, 2 columns	yes	yes	programming in Fortran
DISFIT	Sequential quadratic (Gill et al., 1986)	yes	no, 1 sub. per file	yes	yes	programming in Fortran*
MATHEMATICA	Conjugate gradient	yes	yes	no†	no†	algebraic formula

Note—\*In addition, a special library NPSOL must be obtained from their authors (Gill, Murray, Sanders, & Wright, 1986). †These features can be added.

### Simulation Methods

The simulations repeatedly sample random deviates from one of the five distributions, with known parameters, and then estimate those parameters with each software package. The parameter estimates are then compared against the known values for both accuracy (bias) and variability (efficiency). For each of the five source distributions (ex-Gaussian, Gumbel, lognormal, Wald, and Weibull), we sampled  $n = 250$  random deviates. The parameter values appear in Table 3, along with the associated (theoretical) mean, standard deviation, and skew. They were chosen so that (1) the means and standard deviations are all approximately 1,000 and 100, respectively; (2) the overall distribution shapes are positively skewed (the Gumbel distribution has a constant skew). We repeated the sample-and-fit process 1,000 times, making sure that the same samples were fitted by each software package.

The random samples were generated for each of the five distributions by variously transforming random uniform deviates.<sup>5</sup> The source code of a program that generated these random values is available on the archive site of the Psychonomic Society.

For all QMP calculations, the number of quantiles used was 32. This decision probably put the QMP method at a relative disadvantage because the small number of quantiles was unnecessarily restrictive.

### Simulation Results

The programs reviewed were quite robust and never crashed; MATHEMATICA could not find a solution for only one simulated set of Wald deviates. Some analyses did not finish before they reached the maximum number of iterations allowed. However, because only QMPE and

MATHEMATICA return this information, we did not remove these solutions from further analyses.

**Parameter space.** Before turning to the computation of bias and efficiency, we take a look at parameter space. In Figures 1–5, we plotted the estimated parameters as points in the appropriate parameter spaces (two-dimensional for the Gumbel distribution and three-dimensional for the other). Each point represents one of 100 different samples. The central cross shows the position of the true parameter set used to generate each sample. The purpose of these graphs is to see to what extent parameter dependencies are present and, most importantly, whether some software packages are less sensitive to them than others.

Figure 1 shows the parameters  $\hat{\mu}$  and  $\hat{\sigma}$  estimated from the Gumbel random deviates. As seen, the estimates are all spread out around the true parameters with no systematic deviations, indicating no important bias. Furthermore, all software shows the same dispersion.

Figure 2 shows the parameters  $\hat{\mu}$ ,  $\hat{\sigma}$ , and  $\hat{\tau}$  estimated from ex-Gaussian distributed random deviates. One thing to note is that the cloud is not uniformly spread in all directions but tends to form an ellipse. This is easier to see with the projections on the sides of the plot box. This ellipse more or less goes through the main diagonal of the box, which illustrates that the parameter estimates are not independent. For example, a moderately small estimate for  $\hat{\mu}$  can be compensated by a moderately large value of  $\hat{\tau}$  and a moderately small value of  $\hat{\sigma}$ . QMPE and DISFIT return information about this fact in the form of estimated parameter correlations, but the other software packages do not.

The results for the Weibull distribution were similar, as seen in Figure 3, except that the ellipse is oriented along a different diagonal of the cube. This indicates that

**Table 3**  
**Parameter Values Used to Generate Random Values as a Function of the Distribution**

Distribution	Parameter $\theta$	Parameter Value	Mean	Standard Deviation	Skew
ex-Gaussian	$\{\mu, \sigma, \tau\}$	910.56, 44.721, 89.443	1,000	100	1.43
Gumbel	$\{\mu, \sigma\}$	955, 74	998	94.9	1.13955
Lognormal	$\{\kappa, \mu, \sigma\}$	745, 5.45, 0.36	993	92.4	1.16
Wald	$\{\kappa, \mu, \lambda\}$	725, 275, 2,000	1,000	102	1.11
Weibull	$\{\alpha, \beta, \gamma\}$	800, 220, 2.0	995	102	0.631

a moderately large estimate for  $\hat{\alpha}$  can be compensated by a moderately small estimate for  $\hat{\beta}$  and  $\hat{\gamma}$ . Except for one outlier obtained by QMPE, the efficiencies are roughly comparable (that outlier generated an error exit code and so could have been either censored or remedied by manually setting the starting points).

As can be seen from comparison across the panels of Figures 2 and 3, all software packages returned an ellipse of about the same shape and orientation. In all cases, the centers of gravity of the clouds are near the central cross, suggesting only a small bias, and the overall volume of the clouds suggest equal efficiency for all the software packages. Further investigation of biases and efficiency will be performed later.

Figure 4 shows the—more complicated—results using lognormally distributed random deviates: The points form

a crescent. As a consequence, many of the estimates are close to the true parameter values. However, the center of gravity, because of the curvature, will not be on the cross, resulting in mean bias. In addition, all four software packages are subject to this pathology (although to a lesser extent for MATHEMATICA), suggesting that it is due to the distribution function, not the optimization capabilities of the software. Finally, QMPE has a few outliers near the bottom of panel a. On these occasions, a singular Hessian matrix error was also returned by QMPE.

This pathology is not unique to the lognormal distribution. Wald-distributed random deviates also produced estimates that form a crescent when the parameter space is plotted, as seen in Figure 5. It had the same volume and orientation, regardless of the software package used. Such nonlinear pathology cannot be detected by the es-

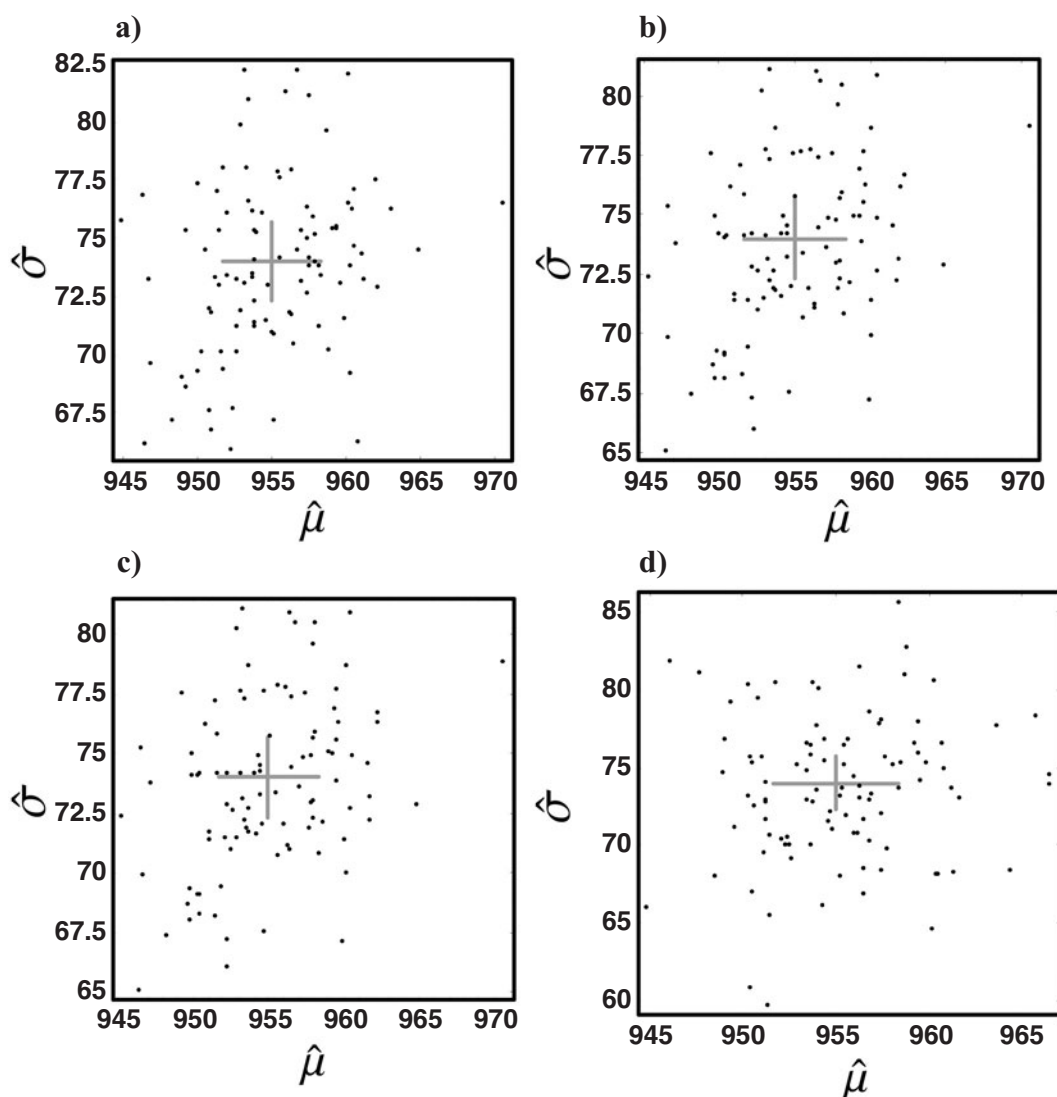


Figure 1. Two-dimensional plots of the estimated parameters  $\hat{\theta}_i$  when the samples were generated with a Gumbel distribution. The central cross shows the position of the true parameter  $\theta$ . Only 100 samples are shown. Plot in panel a was obtained using QMPE, (b) using PASTIS, (c) using MATHEMATICA, and (d) using DISFIT.

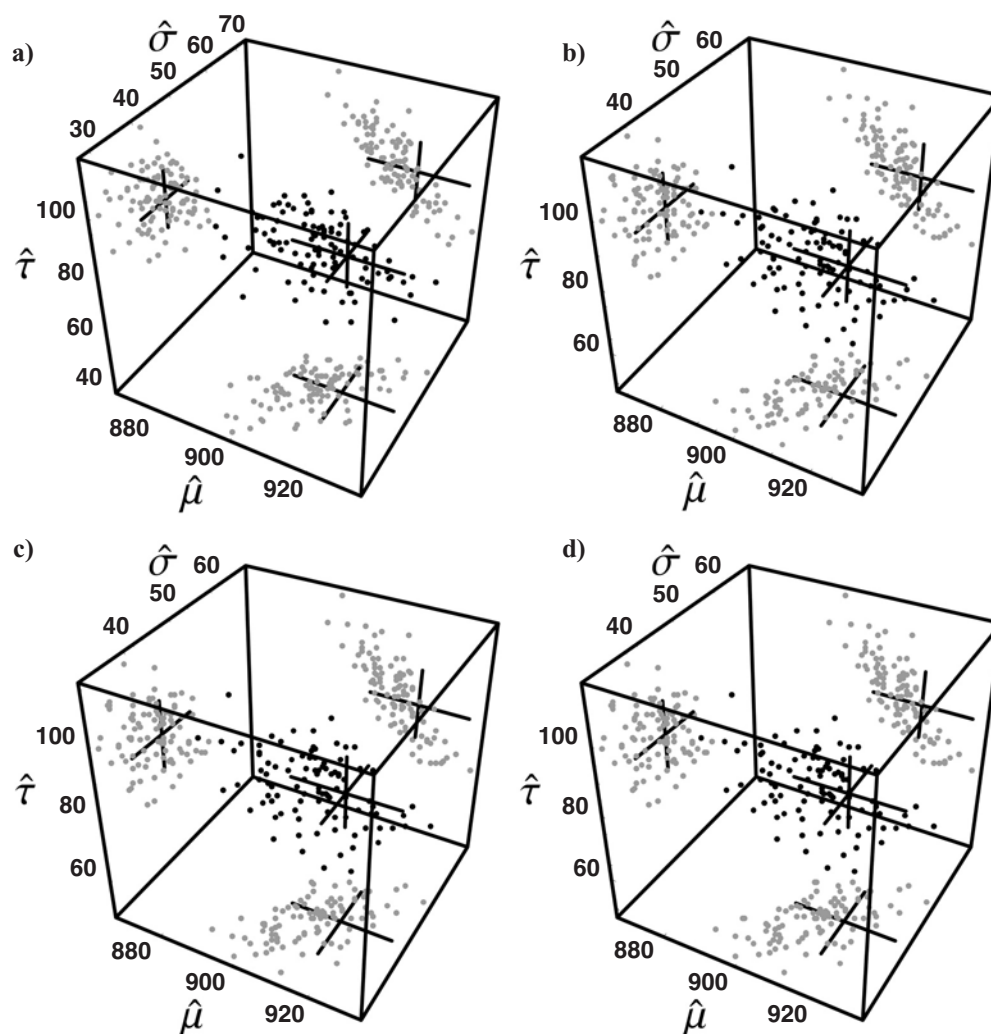


Figure 2. Three-dimensional plots of the estimated parameters  $\hat{\theta}_i$  when the samples were generated with an ex-Gaussian distribution. The central cross shows the position of the true parameter  $\theta$ . The gray points represent the 2-D projections on the  $\mu$ - $\sigma$ ,  $\mu$ - $\tau$ , and  $\sigma$ - $\tau$  planes. Only 100 samples are shown. Plot in panel a was obtained using QMPE, (b) using PASTIS, (c) using MATHEMATICA, and (d) using DISFIT.

estimated parameter correlations; only visual inspection of the parameter space shows it.

**Package capabilities.** In the following, we proceed to an examination of bias and efficiency across packages. However, we will not consider single-parameter biases but rather concentrate on the bias shown by the whole estimated parameter set, relative to the true parameter set. To achieve this, bias was computed as the distance between the center of gravity of all the estimated parameters  $\hat{\theta}_i$ ,  $i = 1 \dots 1,000$ , and the true  $\theta$ . Thus, bias =  $\|E(\hat{\theta}) - \theta\| = E(\|\hat{\theta}_i - \theta\|)$  where  $E(\hat{\theta})$  denotes the average position of all the estimates, and  $\|\cdot\|$  denotes the Euclidean distance (the norm). Efficiency was computed as the standard deviation in the distances between each point  $\hat{\theta}_i$  and  $\theta$ ,  $SD(\|\hat{\theta}_i - \theta\|)$ .

Figure 6 shows the results expressed as a percentage relative to  $\|\theta\|$ . Note that the scales for each panel differ. The two most biased distributions are the lognormal and

the Wald (bottom row), reaching an average of 1% and 25% biases. These are exactly the distributions showing nonlinear dependencies between parameter estimates, as seen in Figures 4 and 5. The other three distribution functions (Weibull, Gumbel, and ex-Gaussian) have much smaller biases, less than 1% in all cases. The bias is even smaller than 0.1% for the Gumbel. In this last case, since the parameter space has only two dimensions, there is less potential for bias. DISFIT turned out to be very apt (low bias, high efficiency), fitting Weibull deviates, whereas MATHEMATICA outperformed the other packages for lognormal deviates.

Overall, the QMP estimates produced by QMPE are as good as those produced by the CML methods obtained from the other software packages (DISFIT being worst for lognormal deviates). This is surprising, considering the major information reduction imposed on the data: They were reduced from 250 raw data points to only 32 quan-

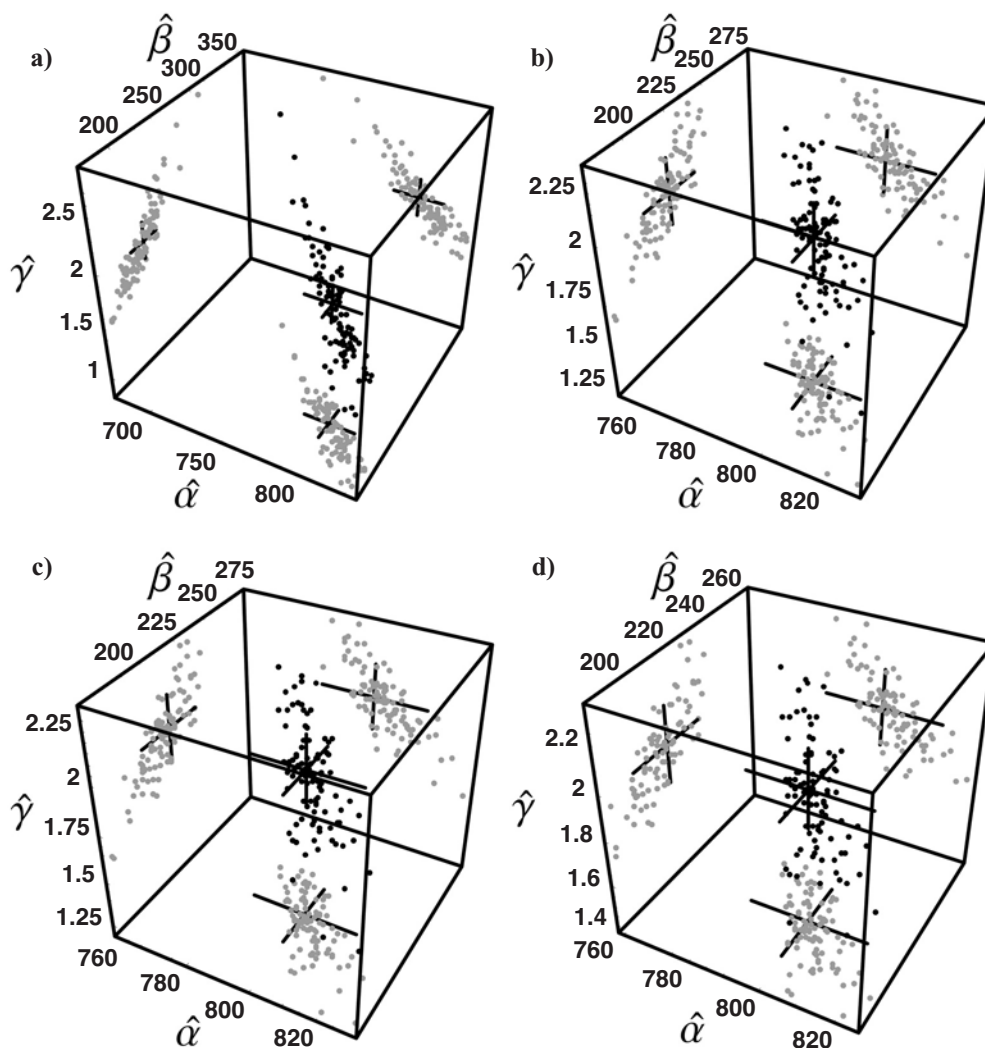


Figure 3. Three-dimensional plots of the estimated parameters  $\hat{\theta}_i$  when the samples were generated with a Weibull distribution. The central cross shows the position of the true parameter  $\theta$ . The gray points represent the 2-D projections on the  $\alpha$ - $\beta$ ,  $\alpha$ - $\gamma$ , and  $\beta$ - $\gamma$  planes. Only 100 samples are shown. Panel a was obtained using QMPE, (b) using PASTIS, (c) using MATHEMATICA, and (d) using DISFIT.

tiles, an almost eight-fold compression. QMPE efficiency, indicated by the error bars in Figure 6, is slightly worse for the Weibull deviates, but results almost entirely from a few outliers (one is visible in Figure 3). Outliers generated by QMPE were often accompanied by an error exit code related to the singularity of the Hessian matrix. Therefore, a very strict selection of the successful fits would have increased considerably the efficiency of the QMPE method, to the detriment of having a little less than 5% of the data set either rejected or requiring refit. When manually fitting a data set, the user should consider changing the starting points or changing the criteria for ending a search.

### Conclusions

Overall, the four software packages lead to very similar bias and efficiency measures, confirming that they all

work properly and that the different platforms and algorithms used make little difference, at least with simulated data. This was true even though different optimization routines and different starting value heuristics underlie each package. The single most important factor on the quality of the estimates was the presence of nonlinear relationships between the parameter estimates. This has implications for comparing groups of subjects. For example, the Wald estimates are so inefficient that they are likely to differ more within group than between groups. If the purpose is to see differences, the ex-Gaussian and Weibull distributions are preferable as atheoretical summaries of shape. Strategies (such as reparameterization, Bates & Watts, 1988) can reduce such nonlinearities. However, the required transformations are difficult to find, sometimes relying on a trial-and-error process.



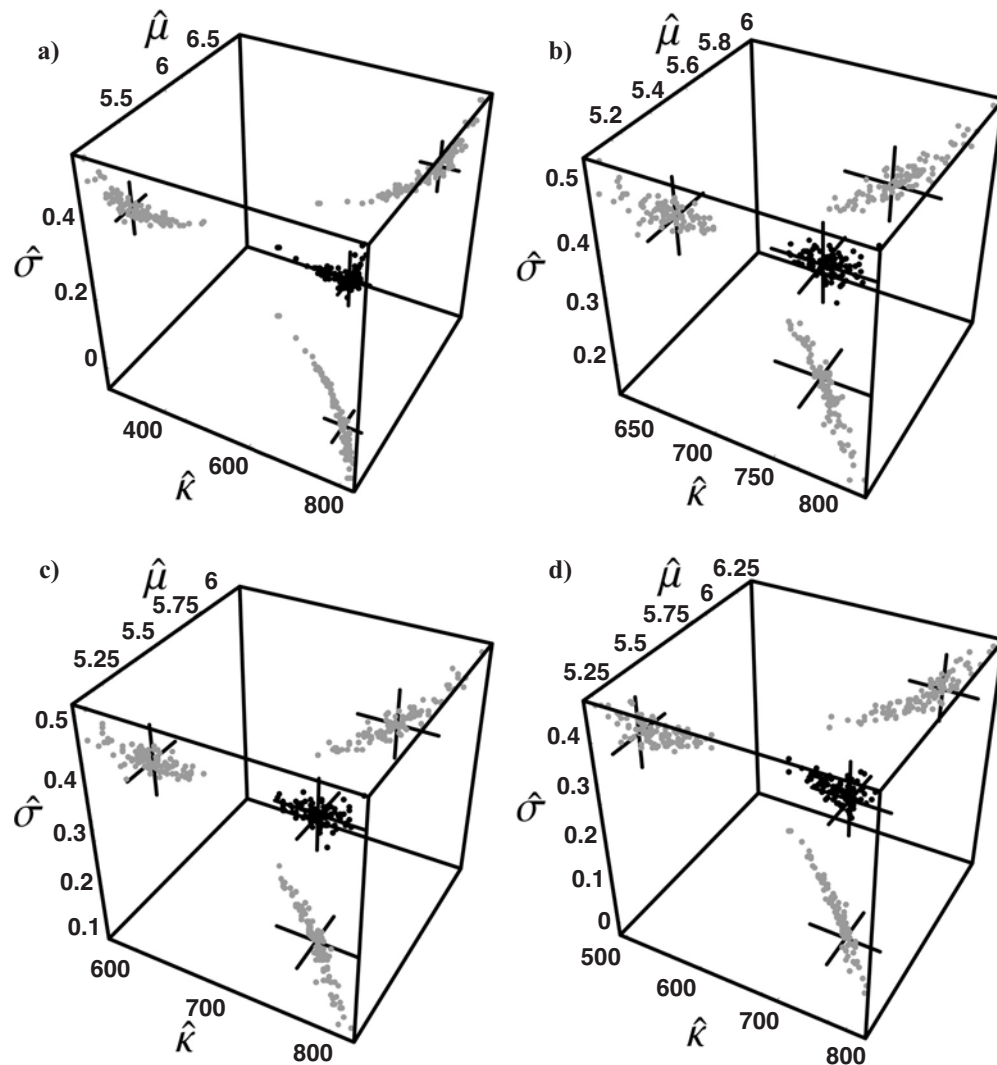


Figure 4. Three-dimensional plots of the estimated parameters  $\hat{\theta}_i$  when the samples were generated with a lognormal distribution. The central cross shows the position of the true parameter  $\theta$ . The gray points represent the 2-D projections on the  $\kappa$ - $\mu$ ,  $\kappa$ - $\sigma$ , and  $\mu$ - $\sigma$  planes. Only 100 samples are shown. Panel a was obtained using QMPE, (b) using PASTIS, (c) using MATHEMATICA, and (d) using DISFIT.

### 3. GENERAL CONCLUSIONS From Descriptive Models of RT to RT Models

Parametric estimation of RT distribution provides a compact description of RT data. In addition, once the distribution is fitted, it is easy to calculate the PDF, CDF hazard, and more. A main point of this article was to show that there are good quality software packages to perform fits and that these packages are reliable and easy to use.

A more theoretical question is to decide which distribution function to fit. As seen in this paper, five candidates can readily be explored. Although there is no consensus at this time, two points should guide one's choice.

The first point concerns the informative utility of the parameters across samples. For example, if a single change

in the experimental procedure results in changes in all the parameters of the distribution, the representation is not compact across conditions. Thus, in choosing a distribution function as a descriptive model, the researcher should be mostly interested in how concisely the parameters capture the experimental manipulation. This should be sought even if it sacrifices some quality of the fit.

Differences in  $L(\hat{\theta}, T)$  across distributions cannot be compared since distribution functions can have different capabilities for fitting random data. For example, a distribution with more free parameters has more liberty to fit the data and will likely have a smaller  $-\ln L(\hat{\theta}, T)$ . One solution is to penalize for extra parameters, as in the AIC test (Bozdogan, 1987). However, even with an equal number of parameters, some functions may be able to accommodate more data sets, a property often termed *geometric*

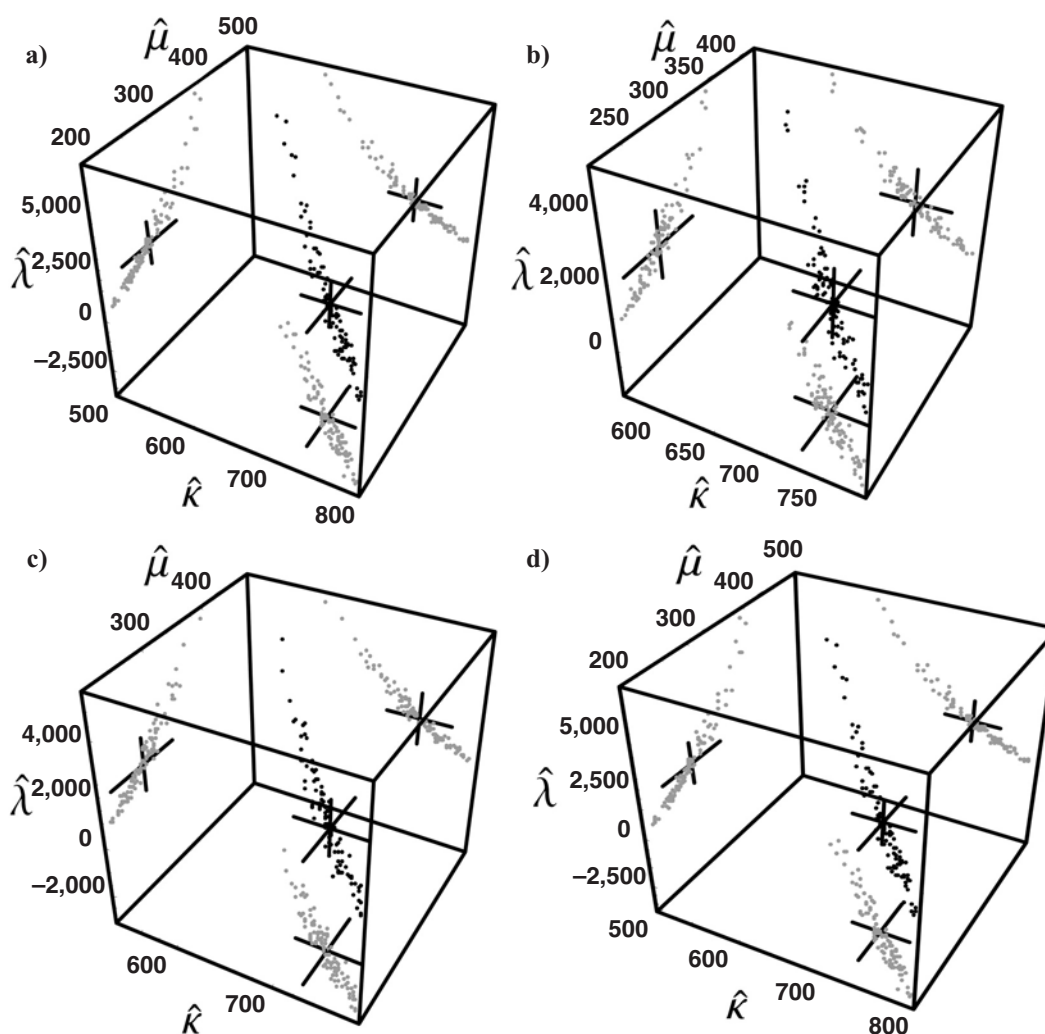


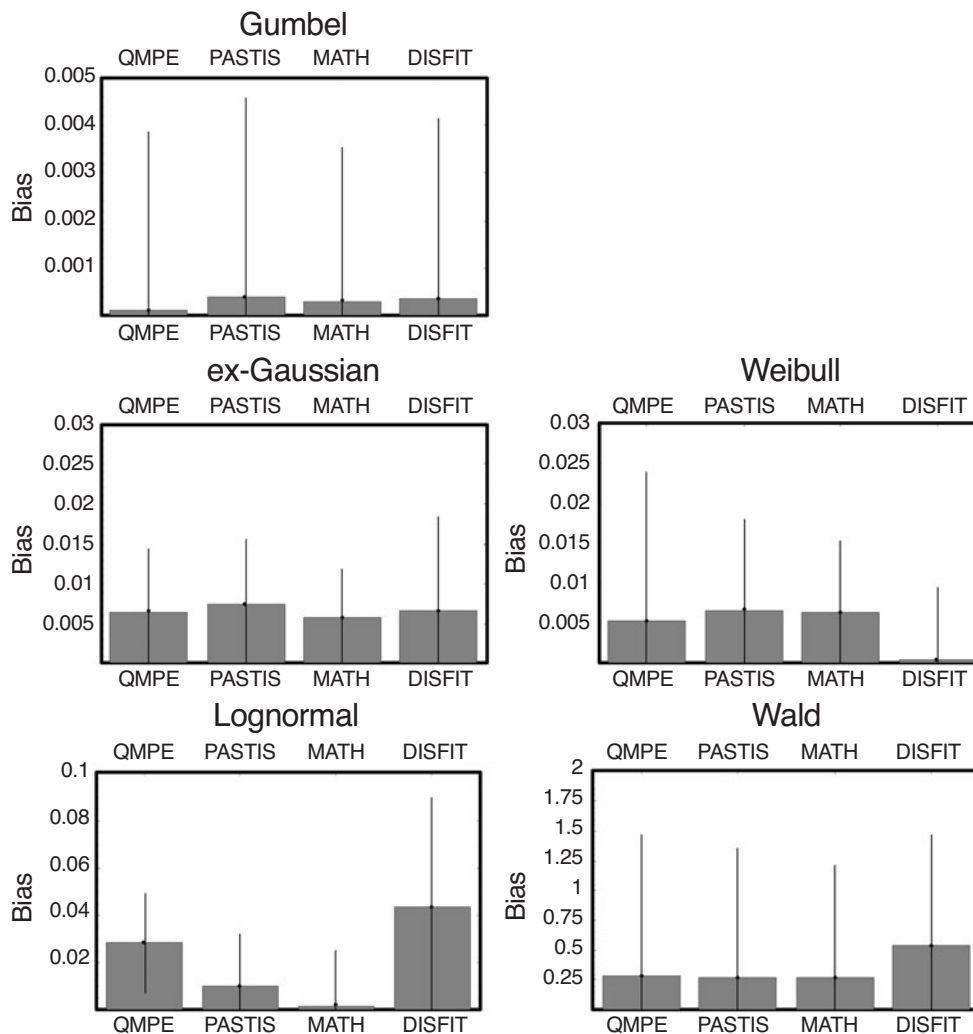
Figure 5. Three-dimensional plots of the estimated parameters  $\hat{\theta}_i$  when the samples were generated with a Wald distribution. The central cross shows the position of the true parameter  $\theta$ . The gray points represent the 2-D projections on the  $\hat{\kappa}$ - $\hat{\mu}$ ,  $\hat{\kappa}$ - $\hat{\lambda}$ , and  $\hat{\mu}$ - $\hat{\lambda}$  planes. Only 100 samples are shown. Panel a) was obtained using QMPE, (b) using PASTIS, (c) using MATHEMATICA, and (d) using DISFIT.

*complexity*. There are methods to adjust the penalty term to compensate for complexity, but these can be computationally difficult (Myung, 2000; Grünwald, 2000).

The second important point in the choice of a distribution to fit is related to psychological models of cognition. Whereas a researcher might simply be interested in a descriptive model of RT for convenient communication of the results, a more ambitious approach is to have a model based on psychological mechanisms that can predict not only RT but also the shape and scale of the whole RT distribution. Two cases are then possible: First, the model can be analytically solved to yield an algebraic formula for the RT distribution (see Cousineau, in press). It can either be one of the distributions reviewed here or a yet-unknown distribution function. In this case, the re-

searcher can fit this distribution and ensure that the parameters are acting according to a priori predictions (Schwarz, 2001). Second, in the case in which the model cannot be solved analytically, the researcher can simulate the model and choose a descriptive model to fit the simulated RT. By doing the same to the observed RT distribution, the researcher can check that the descriptions are convergent. This is the approach used in Ratcliff (1979), where the ex-Gaussian was the intermediary between empirical and simulated data.

The best solution is to fill the gap between a model and RT data with more than just the predicted means. However, it is possible that the observed RTs are contaminated by other factors, such as fatigue or fast guess. We thus have to keep in mind the possibility of fitting



**Figure 6.** Bias and efficiency for the four software packages reviewed. Histograms indicate the bias—that is, the average distance between the estimated parameter  $\hat{\theta}$  and the true parameter  $\theta$ . Error bars are the efficiency—that is, the standard deviation in the distance between the estimated parameter  $\hat{\theta}$  and the true parameter  $\theta$ . Shorter bars indicate better efficiency. The top row comes from the two-parameter distribution (the Gumbel); the middle row comes from distributions that had linear dependencies between parameter estimates (ex-Gaussian and Weibull); the bottom row comes from distributions with nonlinear dependencies between parameter estimates (Lognormal and Wald).

mixtures of distributions (Cousineau & Shiffrin, 2004; Dolan, van der Maas, & Molenaar, 2002) or that the parameters of the distributions are changing over time. In this context, QMP estimation is likely to be more robust to the effects of outliers and measurement noise than standard CML estimation.

#### REFERENCES

- AHRENS, J. H., & DIETER, U. (1982). Computer generation of Poisson deviates from modified normal distributions. *ACM Transactions on Mathematical Software*, *8*, 163-179.
- ASHBY, F. G., TEIN, J.-Y., & BALAKRISHNAN, J. D. (1993). Response time distributions in memory scanning. *Journal of Mathematical Psychology*, *37*, 526-555.
- BATES, D. M., & WATTS, D. G. (1988). *Nonlinear regression analysis and its application*. New York: Wiley.
- BERINGER, J. (1992). Timing accuracy of mouse response registration on the IBM microcomputer family. *Behavior Research Methods, Instruments, & Computers*, *24*, 486-490.
- BLOXOM, B. (1984). Estimating response time hazard functions: An exposition and extension. *Journal of Mathematical Psychology*, *28*, 401-420.
- BOX, M. J., DAVIES, D., & SWANN, W. H. (1969). *Non-linear optimization techniques*. Edinburgh: Oliver & Boyd.
- BOZDOGAN, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345-370.
- BROWN, S., & HEATHCOTE, A. (2003). QMLE: Fast, robust, and efficient estimation of distribution functions based on quantiles. *Behavior Research Methods, Instruments, & Computers*, *35*, 485-492.

- BURBECK, S. L., & LUCE, R. D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Perception & Psychophysics*, **32**, 117-133.
- 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.
- COUSINEAU, D. (in press). Merging race models and adaptive networks: A parallel race network. *Psychonomic Bulletin & Review*.
- COUSINEAU, D., & SHIFFRIN, R. M. (2002). Extending statistics of extremes to distributions varying in position and scale and the implications for race models. *Journal of Mathematical Psychology*, **46**, 431-454.
- 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. (in press). Visual-memory search: An integrative perspective. *Psychological Research*.
- COUSINEAU, D., & SHIFFRIN, R. M. (2004). Termination of a visual search with large display size effect. *Spatial Vision*, **17**, 327-352.
- DAGPUNAR, J. (1988). *Principles of random variate generation*. Oxford: Oxford University Press.
- DAWSON, M. R. W. (1988). Fitting the ex-Gaussian equation to reaction time distributions. *Behavior Research Methods, Instruments, & Computers*, **20**, 54-57.
- DOLAN, C. V., & MOLENAAR, P. C. M. (1991). A comparison of four methods of calculating standard errors of maximum-likelihood estimates in the analysis of covariance structure. *British Journal of Mathematical & Statistical Psychology*, **44**, 359-368.
- DOLAN, C. V., VAN DER MAAS, H. L. J., & MOLENAAR, P. C. M. (2002). A framework for ML estimation of parameters (mixtures of) common reaction time distributions given optional truncation or censoring. *Behavior Research Methods, Instruments, & Computers*, **34**, 304-323.
- FLETCHER, R. (1980). *Practical methods of optimization*. New York: Wiley.
- GILL, P. E., MURRAY, W., SANDERS, M. A., & WRIGHT, M. H. (1986). *User's guide for NPSOL (Version 4.0): A FORTRAN package for nonlinear programming* (Tech. Rep. No. SOL 86-2). Stanford University: Department of Operation Research.
- GILL, P. E., MURRAY, W., & WRIGHT, M. H. (1981). *Practical optimization*. London: Academic Press.
- GRÜNWARD, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, **44**, 133-152.
- GUMBEL, E. J. (1958). *The statistics of extremes*. New York: Columbia University Press.
- HAYS, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart & Winston.
- HEATHCOTE, A. (2004). Fitting Wald and ex-Wald distributions to response time data: An example using functions for the S-PLUS package. *Behavior Research Methods, Instruments, & Computers*, **36**, 678-694.
- 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.
- HEATHCOTE, A., BROWN, S., & MEWHORT, D. J. K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin & Review*, **9**, 394-401.
- HEATHCOTE, A., POPIEL, S. J., & MEWHORT, D. J. K. (1991). Analysis of response-time distributions: An example using the Stroop task. *Psychological Bulletin*, **109**, 340-347.
- HOCKLEY, W. E. (1984). Analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 598-615.
- KEMP, C. D. (1986). A modal method for generating binomial variables. *Communications on Statistics—Theoretical Methods*, **15**, 805-813.
- LUCE, R. D. (1986). *Response times, their role in inferring elementary mental organization*. New York: Oxford University Press.
- MARSAGLIA, G., & TSANG, W. W. (2000). A simple method for generating gamma variables. *ACM Transactions on Mathematical Software*, **26**, 363-372.
- MILLER, J. (1988). A warning about median reaction time. *Journal of Experimental Psychology: Human Perception & Performance*, **14**, 539-543.
- MYUNG, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, **44**, 190-204.
- NELDER, J. A., & MEAD, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308-313.
- PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A., & VETTERLING, W. T. (1986). *Numerical recipes: The art of scientific computing*. New York: Cambridge University Press.
- RATCLIFF, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, **86**, 446-461.
- RATCLIFF, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, **114**, 510-532.
- SCHWARZ, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, & Computers*, **33**, 457-469.
- SILVERMAN, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.
- TOWNSEND, J. T., & ASHBY, F. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University Press.
- TOWNSEND, J. T., & COLONIUS, H. (2001, July). *Variability of MAX and MIN statistics: A theory of the quantile spread as a function of sample size*. 34th Annual Meeting of the Society for Mathematical Psychology, Providence, RI.
- ULRICH, R., & MILLER, J. (1993). Information processing models generating lognormally distributed reaction times. *Journal of Mathematical Psychology*, **37**, 513-525.
- ULRICH, R., & MILLER, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, **123**, 34-80.
- VAN ZANDT, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, **7**, 424-465.
- VAN ZANDT, T., & RATCLIFF, R. (1995). Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures. *Psychonomic Bulletin & Review*, **2**, 20-54.
- WOLFRAM, S. (1996). *The Mathematica Book (3rd Ed.)*. New York: Cambridge University Press.
- YELLOTT, J. I., JR. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, **15**, 109-144.

## NOTES

1. In the following, we will use the word *fitting* to mean the process of parametric estimation.
2. PASTIS, QMPE, DISFIT, and MATHEMATICA use the same parameterization for the ex-Gaussian, the Gumbel, the lognormal, and the Weibull. For the Wald distribution, QMPE differs in using  $a = \sqrt{\lambda}$ , and  $m = a/\mu$ . The Wald and the ex-Gaussian are not built into MATHEMATICA.
3. To illustrate, consider a distribution that predicts only one possible value, say  $\alpha$  with probability 1. If 100 values are sampled, all identical and equal to say 10, the likelihood that  $\alpha$  is 10,  $L(10)$  is  $1 \times 1 \dots \times 1 = 1$ . On the other hand, for any  $\alpha$  not 10,  $L(\alpha) = 0 \times 0 \dots \times 0 = 0$ . Thus, the parameter  $\alpha = 10$  maximizes the likelihood of observing such data. In general, distribution functions do not predict a single value, so the parameters that maximize the likelihood will not reach a value of 1 (and may even not be close to 1) but again, values closest to 1 indicate the most likely parameter estimates.
4. Three kinds of stopping criteria are used in optimization software. One is related to the decrease in the objective function. If the difference in the objective function from one iteration to the next is smaller than a fixed proportion (say  $10^{-9}$ ), the algorithm considers itself close enough to the minimum and exits. The second stopping criterion concerns the

change in the solution  $\hat{\theta}$ . If the Euclidian distance between two successive solutions, measured in the  $n$ -dimensional space, is smaller than a fixed proportion (say  $10^{-4}$ ), the algorithm considers itself close enough to be minimum and exits. The last stopping criterion is based on the maximum number of iterations (say 150), after which the algorithm considers that the minimum cannot be found and exits.

5. The algorithm supplied with the Sun Forte Professional Workshop Fortran compiler (Version 6, Update 2) generated the uniform deviates. This algorithm uses a pair of linear congruential generators, with a cycle length in excess of  $2^{50}$ . Because only approximately  $2^{24}$  samples were used for the present study, the long cycle time ensures that samples were independent. Exponential, Weibull, Gumbel, and  $\chi^2$  deviates were all generated from uniform deviates by transforming them with the respective inverse cumulative distribution functions. Normal deviates were produced from uniform deviates (Press, Flannery, Teukolsky, & Vetterling, 1986). The sum of normal and exponential deviates produced ex-Gaussian deviates. Exponentially transforming normal deviates and adding a shift parameter produced lognormal deviates. Finally, Wald deviates were produced from  $\chi^2$  deviates followed by the addition of a shift parameter.

## ARCHIVED MATERIALS

The following materials associated with this article may be accessed through the Psychonomic Society's Norms, Stimuli, and Data archive, <http://www.psychonomic.org/archive/>.

To access these files or links, search the archive for this article using the journal (*Behavior Research Methods, Instruments, & Computers*), the first author's name (Cousineau) and the publication year (2004).

FILE: Cousineau-BRMIC-2004.zip

DESCRIPTION: The compressed archive file contains three files: `randmod.f90` and `random.f90` are the two parts of a Fortran 90 program that generates sets of random numbers corresponding to samples from the following distributions: ex-Gaussian, Gumbel, lognormal, Wald, and Weibull. The code is adapted from the work of Dagpunar (1988), Marsaglia and Tsang (2000), Ahrens and Dieter (1982), and Kemp (1986).

`readme.txt` is a text file explaining the purpose of the program and how to compile it on most stations.

AUTHOR'S E-MAIL ADDRESS: [denis.cousineau@umontreal.ca](mailto:denis.cousineau@umontreal.ca).

AUTHOR'S WEB SITE: <http://mapageweb.umontreal.ca/cousined>.

## APPENDIX A Optimization Procedures

**Simplex.** This derivative-free routine constructs a  $d$ -dimensional polygon having  $d + 1$  vertices (a simplex, where  $d$  is the number of parameters to be minimized) in the parameter space. For example, in a three-dimensional space, the simplex is a pyramid. The value of the objective function is computed at each corner, and the simplex reorients itself around the corner with the lowest value, effectively approximating the gradient, and either contracting or expanding. This procedure repeats until the simplex is reduced to a very small volume around a minimum. For more, see Nelder and Mead (1965) and Van Zandt (2000).

**Steepest descent.** This gradient method searches the space by looking at the slope. If the derivatives are available in algebraic form, it is easy to locate the direction of steepest descent. If they are not available, the routine probes a few points in the parameter space surrounding the current  $\hat{\theta}_i$  to find the ideal direction by numerically approximating the derivatives. For more details, see Box, Davies, and Swann (1969).

The two above methods are not quadratically convergent. If the objective function near the minimum is well approximated by a quadratic function (as is often the case), quadratically convergent methods are guaranteed to find the minimum to a certain precision with a fixed number of iterations (Fletcher, 1980). The following four methods have been proven to be quadratically convergent:

**Newton method.** This method is the most efficient, given that its assumptions are true. It is based on the second-order derivatives of the objective function, called the *Hessian matrix*,  $H$ . Since analytic Hessians can be difficult to determine (either numerically or algebraically), this method is rarely used. Hessians are also useful to compute approximate standard errors (*SE*) and parameter intercorrelations (see Bates & Watts, 1988, and Dolan & Molenaar, 1991).

**Quasi-Newton and conjugate gradient methods.** These two methods generate an approximate Hessian  $\hat{H}$  using first-order derivatives. They differ in terms of memory requirements. The first method updates large matrices to build  $\hat{H}$ , whereas the second updates eigenvectors (Chandler, 1965). For the small number of parameters typically minimized in RT distribution estimation, these differences are mostly immaterial.

**Sequential quadratic method.** This method uses Lagrangians instead of Hessians (Gill, Murray, & Wright, 1981). It is mostly useful when very restrictive constraints (in addition to non-negative parameters) are imposed on the parameters, which is rarely the case with distribution fitting.

**APPENDIX B****Packages That Perform Likelihood Fitting**

---

**PASTIS.** A very simple software package which returns  $-\ln L(\hat{\theta}, T)$  and the best-fitting parameters  $\hat{\theta}_p$ . It has very few limits on the format of the input file (Cousineau & Larochelle, 1997; available at <http://www.mpageweb.umontreal.ca/cousined/papers/02-pastis>).

**QMPE.** In addition to fitting with both CML and QML, QMPE can compute quantiles and vincentiles (Dawson, 1988; Ratcliff, 1979). It also returns estimated standard errors of the parameters, calculated with the second derivative Hessian matrix. File format, however, is restricted to two columns: subject number and one RT per line (Brown & Heathcote, 2003; Heathcote, Brown, & Cousineau, 2004, available at <http://www.newcastle.edu.au/school/behav-sci/ncl/>).

**DISFIT.** A versatile fitting package, aimed at entirely different objectives than QMPE. While it can minimize  $L(\hat{\theta}, T)$ , it can also fit mixtures of two or more distributions, and fit data that have been truncated or censored (see Ulrich & Miller, 1994). It also returns statistics of goodness of fit ( $\chi^2$  and Kolmogorov-Smirnov) and estimated standard errors of the parameters. However, like QMPE, the file format is not flexible: one subject per file, one column per file (Dolan et al., 2002, available at [users.fmg.uva.nl/cdolan](http://users.fmg.uva.nl/cdolan)).

**MATHEMATICA/FindMinimum.** The most versatile software package of all, an all-purpose framework for data and function manipulation. It can fit any distribution that can be expressed by an algebraic formula, and, appropriately programmed, can compute starting values and quantiles as well as perform minimizations on truncated and censored data. However, this flexibility comes with large costs: (1) the user must be trained on how to use MATHEMATICA since it is a very rich environment; and (2) the analyses are much slower because the commands are interpreted, not compiled. MATHEMATICA, in contrast to all the above, is a commercial product (see [www.wolfram.com](http://www.wolfram.com) for details). Other commercial software packages that allow the same flexibility are readily available; S-Plus, SAS, and Matlab are perhaps the most widely used of these. For example, Heathcote (2004) provides S-Plus routines to fit the Wald.

---

**APPENDIX C**  
**Examples of Commands Used for Each of the Four Software Packages**

---

**Listing 1: Typical PASTIS call for a file with 2 columns, the second being the dependent (RT) variable**

```
pastis -r file.dat -w file.out -c 2 -d 2 -a WEIBULL
```

**Listing 2: Typical QMPE script.** Extra information at the end of each line consists of comments.

```
file.dat    input file
file        output stem
1           measurement precision, i.e. 2 epsilon
1           mode: 0: silent, 1: one output/cell, 2: trace mode
1.OE-09     proportional objective function change tolerance
1.OE-04     proportional L (inf) norm tolerance
150         maximum number of iterations allowed
2           distribution to fit 1: exgaussian, 2: weibull, ...
2           type of analysis, 1: raw fitting, 2: quantile fitting
1           input type, 1: raw data, 2: quantiles, 3: vincentiles
32          how many quantiles to compute since raw is provided
```

**Listing 3: Typical DISFIT script.** Lines beginning with ! are comments

```
title "fitting file.dat using a Weibull distribution"
!
! subject number and use of starting values heuristics
ns=1 st=yes bs=0
!
! input file and number of observation (missing is -999)
df=file.dat nc=1 no=250 id="One_subject" mi=-999 cp=0
!
! truncation information
tt=no th=no xl=no xh=no
!
! description of the distribution to fit, with starting
! values (not used since st=yes) and range of each parameters.
di=wei
a=[ 1 200.0 50.0 2000.0] c=[ 2 2.0 0.5 3.6] k=[ 3 750.0 400.0 1200.0]
```

**Listing 4: Typical MATHEMATICA script.** Lines enclosed between (\* and \*) are comments.

```
(* Load distribution-related package *)
<< Statistics`ContinuousDistributions`

(*read the input file into "data"*)
data = ReadList[ "file.dat", Real];

(* define the log likelihood function *)
LogLikelihood[ data_, a_, b_, c_] :=
  -Plus @@ Log[ PDF[ WeibullDistribution[ c,b] ,data-a] /;a<Min[ data]

(* error: return a large value if the parameter a is too large *)
LogLikelihood[ data_, a_, b_, c_] :=
  $MaxMachineNumber /;a>=Min[ data]

(* perform a search over the parameter space *)
(* in Mathematica, a direct search requires two starting values *)
FindMinimum[ LogLikelihood[ data, a, b, c] ,
  { a, { 200, Min[ data] } },
  { b, { 40, 60 } },
  { c, { 0.5, 3 } }
]
```

---