Tutorials in Quantitative Methods for Psychology
2013, Vol. 9(2), p. 61-71.

Improving maximum likelihood estimation using prior probabilities: A tutorial on maximum a posteriori estimation and an examination of the weibull distribution

Denis Cousineau *Université d'Ottawa*
Sebastien Helie *Purdue University*

This tutorial describes a parameter estimation technique that is little-known in social sciences, namely maximum a posteriori estimation. This technique can be used in conjunction with prior knowledge to improve maximum likelihood estimation of the best-fitting parameters of a data set. The estimates are based on the mode of the posterior distribution of a Bayesian analysis. The relationship between maximum a posteriori estimation, maximum likelihood estimation, and Bayesian estimation is discussed, and example simulations are presented using the Weibull distribution. We show that, for the Weibull distribution, the mode produces a less biased and more reliable point estimate of the parameters than the mean or the median of the posterior distribution. When Gaussian priors are used, it is recommended to underestimate the shape and scale parameters of the Weibull distribution to compensate for the inherent bias of the maximum likelihood and Bayesian methods which tend to overestimate these parameters. We conclude with a discussion of advantages and limitations of maximum a posteriori estimation.

[*]Parameter estimation techniques are used to estimate the parameters of a distribution model which maximizes the fit to a particular data set. The most commonly used technique in social sciences is maximum likelihood estimation (MLE, Edwards, 1992, Myung, 2000, but see Van Zandt, 2000, for alternatives). Software using this technique includes RTSYS (Heathcote, 1996), PASTIS (Cousineau and Larochelle, 1997), QMPE (Brown and Heathcote, 2003, Heathcote, Brown and Cousineau, 2004), DISFIT (Dolan, 2000), and Mathematica (version 8 and above, Wolfram Research inc., 2011); see Cousineau, Brown and Heathcote, 2004, for a review and a comparison). In MLE, the likelihood of a set of parameters given the data is computed and a search for the parameters that maximize the likelihood is performed. This technique is very general and can be applied to any population distribution (e.g., ex-Gaussian, lognormal, Weibull, etc.; see Luce, 1986, Appendix A, for a review of some distribution functions).

A more general technique used in mathematical statistics is Bayesian estimation (BE, Edwards, Lindman and Savage, 1963). This technique returns the posterior distribution of the parameters given the data. By taking the mean of the posterior distribution, a point estimate of the best-fitting parameters can be obtained. Alternatively, the median or the mode of the posterior distribution can be used instead of the mean. One important advantage of BE over MLE is the possibility to inject prior knowledge on the parameters. Interval priors can be used to restrict one parameter within some given bounds, e.g., a parameter that can only be

[*] Denis Cousineau, École de psychologie, Université d'Ottawa, 136, Jean-Jacques Lussier, Ottawa, K1N 6N5, CANADA, E-mail: denis.cousineau@uottawa.ca; Sebastien Helie, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907-2081 USA, E-mail: shelie@purdue.edu. This article contains equal contributions from both authors.

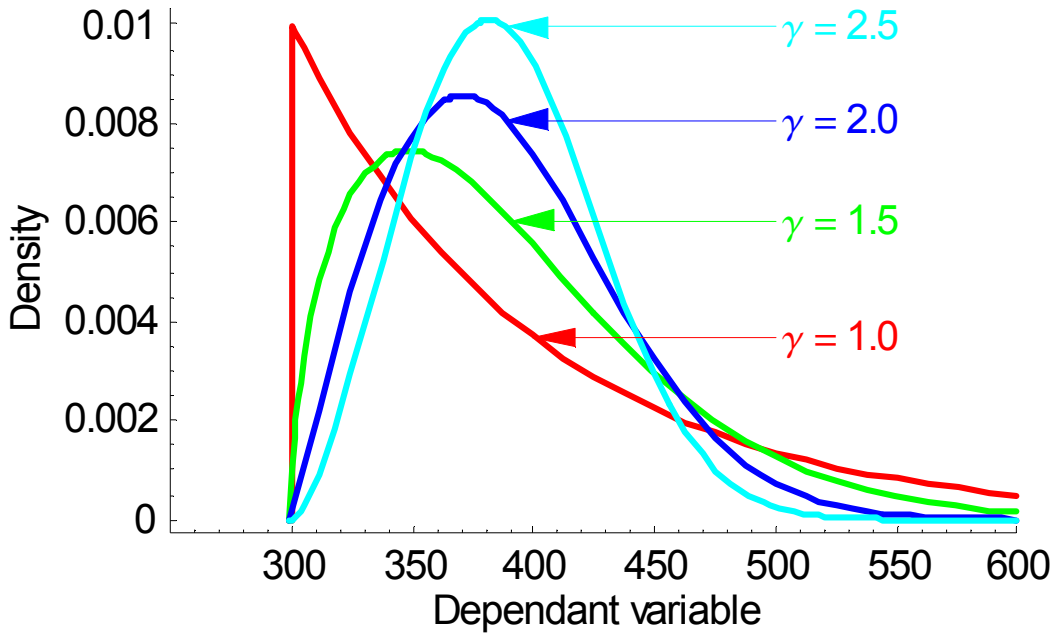

Figure 1. Four examples of Weibull distribution densities. In all cases, $\alpha = 300$ and $\beta = 100$, which correspond approximately to the shift and scale of well-trained participants in a simple task. When $\gamma \leq 1$, the distribution is J-shaped; when $\gamma = 3.6$ (not shown), the distribution is almost symmetrical with a Fisher skew of 0.

positive. Likewise, if a distribution can be positively or negatively skewed, but the data are known to always be positively skewed, the parameter(s) that determine the skew could be constrained by the use of an interval prior.

However, the use of priors is not limited to defining limits on the parameter domain: unbounded priors can also be used. For example, normal (Gaussian) priors are used when a parameter is believed to be normally distributed around a true population parameter. For instance, the γ parameter of the Weibull distribution applied to response time data is always found to be in the vicinity of 2 plus or minus 1 (Logan, 1992, Cousineau and Shiffrin, 2004, Huber and Cousineau, 2003, Rouder, Lu, Speckman, Sun, & Jiang, 2005). This prior knowledge could be entered in a BE analysis using a normal distribution with mean 2 and standard deviation 0.5.

As will be outlined in the following section, MLE is a special case of BE in which (1) the estimate is based on the mode of the posterior distribution, and (2) all the parameter values are equally likely (i.e., there is no priors). In the following section, we show how priors from BE can be inserted back into MLE, a technique called *Maximum A Posteriori* (MAP) estimation (Neapolitan, 2004). The MAP estimates are much faster to compute than BE, as they do not require the estimation of integrals (which are generally not available in closed form) or the use of Markov chains. In this tutorial, we explain how MAP estimation is related to MLE and BE, and we use the Weibull distribution as an example model to illustrate its use. The first important result is that the mode is the most accurate central tendency statistic of the posterior distribution to infer best-fitting (point estimate) parameters in the context of the Weibull distribution. As such, the full flexibility of BE is not required in the Weibull case. This presentation is followed by an examination of the impact of priors on the estimates. The second important result is that normal priors, even when inaccurate, are useful to avoid outlier estimates and thus improve parameter estimation.

A primer on the Weibull distribution

The present tutorial uses the Weibull distribution as an example application of MAP estimation. However, MAP estimation is a general technique and any parametric model can be used with any type of priors. The Weibull distribution was selected here because it is simple yet convenient for describing a data set (generally, response times). Moreover, the Weibull distribution is described by three parameters, and each parameter quantifies a different aspect of the data, which is useful in conveying simple explanations. In addition, many psychological models predict a Weibull distribution (Cousineau, 2004, Miller and Ulrich, 2003, Tuerlinckx, 2004, Cousineau, Goodman and Shiffrin, 2003, Marley, 1989). Figure 1 shows four probability densities with various levels of asymmetry. The probability density function (pdf) of the Weibull distribution is given by:

$$f(x|\gamma,\beta,\alpha) = e^{-\frac{(x-\alpha)}{\beta}^{\gamma}}(x-\alpha)^{\gamma-1}\beta^{-\gamma}\gamma \quad (1)$$

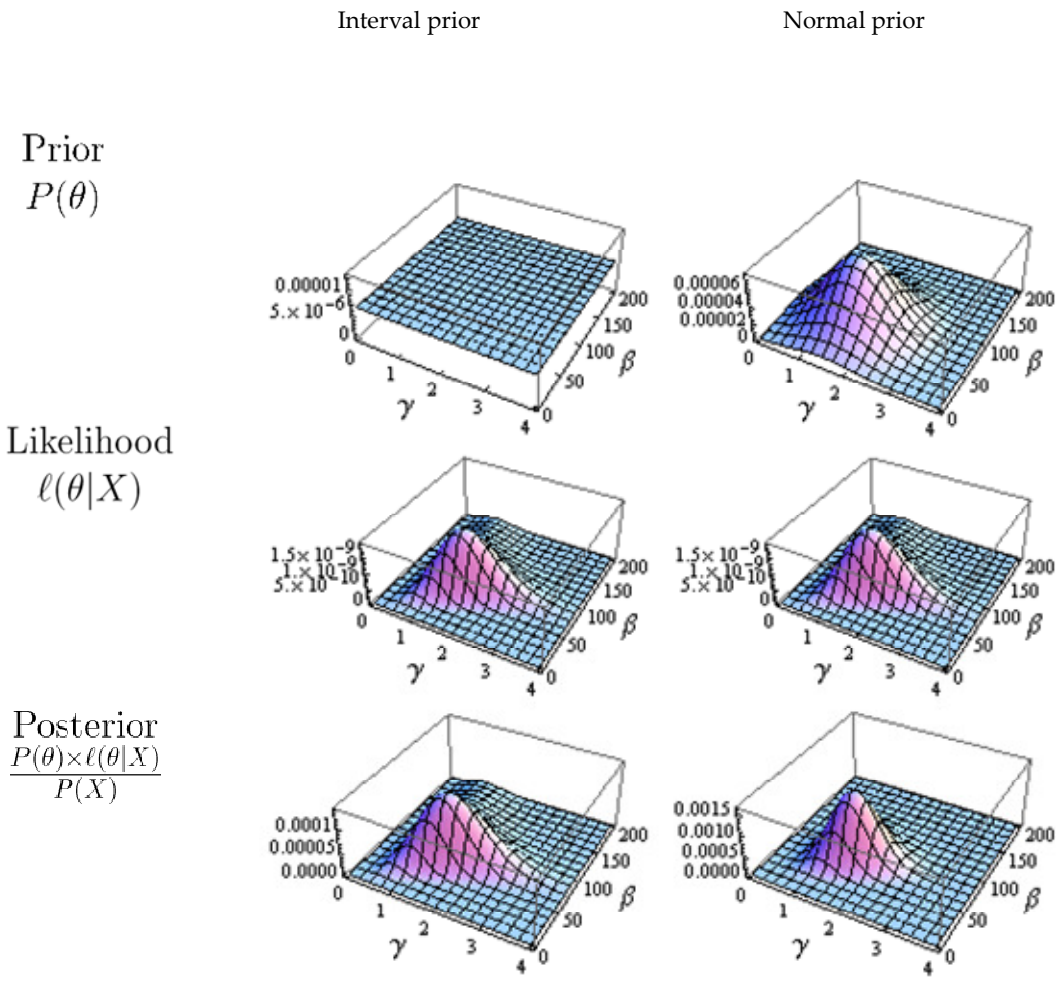

Figure 2. The steps to obtain a posterior distribution of parameters. The top row shows priors over all the parameters. The left panel shows an interval prior while the right panel shows a normal prior. The middle row shows the likelihood surface. The sample in both case is **X** = {303, 320, 370, 407} so that the likelihood surfaces shown in the middle row are identical. The bottom row shows the posterior distributions obtained by multiplying the prior and the likelihood pointwise, and then dividing by the probability of the data $P(\mathbf{X})$.

where x is one datum and $\theta = \{\gamma, \beta, \alpha\}$ is the set of parameters. The parameter α represents the shift of the distribution, the parameter β is related to the variance, and the parameter γ quantifies the degree of asymmetry (see Figure 1; Rouder et al., 2005, Cousineau and Shiffrin, 2004, Luce, 1986).

The Weibull distribution does not meet the regularity conditions for MLE (see Kiefer, 2006, for a list of these conditions). In particular, the domain of one of its parameter (α) depends on the observed data (e.g., α has to be smaller than any of the data). As a result, it is not known whether the MLE technique (and more generally the mode of the posterior distribution) is the most efficient (least variable) approach to get point estimates (Rose and Smith, 2000). Rockette, Antle and Klimko (1974) showed that if the true shape parameter is below 1 (J-shaped distribution), MLE is an inconsistent technique because there may be more than one best-fitting set of parameters (i.e., more than one mode to the posterior distribution). Luckily, J-shaped distributions of response time data have never been observed (the main application of the Weibull distribution in psychology). Smith (1985) showed that for $1 \leq \gamma \leq 2$, the posterior distribution is unimodal but is not symmetrically distributed so that the mean, the mode and the median (among other central tendency statistics) are not equal. Smith also showed that for $\gamma > 2$, the posterior distribution tends towards a normal distribution (as a consequence, the mean, mode and median are equal) as the sample size is large (here, large is generally

believed to be above 100). Because the shape of response time data is generally estimated to be near 2, the last two scenarios are relevant and the subsequent simulations will explore them separately.

Bayesian estimates vs. Maximum likelihood estimates

The BE technique works by updating the prior probability of some parameters following the acquisition of new information (Bishop, 1995; Edwards, Lindman & Savage, 1963; Jeffreys, 1961; Hastie, Tibshirani & Friedman, 2001). The result is called a posterior probability distribution. In the present context, the new information is a sample **X** of size n. The posterior probability density of the parameters, noted $P(\theta \mid \mathbf{X})$, is given by applying Bayes' theorem:

$$P(\theta|\mathbf{X}) = l(\theta|\mathbf{X}) \times \frac{P(\theta)}{P(\mathbf{X})} \quad (2)$$

in which $l(\theta \mid \mathbf{X})$ is the likelihood of the data under the putative parameter set θ, and $P(\theta)$ models the prior knowledge available on the parameter θ (that is, $\{\gamma, \beta, \alpha\}$ in the case of the Weibull distribution). The priors can be of any type as long as they are a distribution model.

The term $P(\mathbf{X})$, called the probability of the observed data, is a normalizing constant ensuring that the posterior distribution has an area of 1. It is given by $\int_\theta l(\theta|\mathrm{X})P(\theta)d\theta$. The computation of this constant is sometimes cumbersome because it often involves solving multiple integrals that can only be estimated numerically (Bishop, 1995; Hastie et al., 2001). The complexity of its calculation might explain why the Bayesian approach is rarely used in psychology and may explain the preference of social scientists for the simpler MLE method.

Figure 2 illustrates the steps required to obtain a posterior distribution. The plots are a function of two parameters, γ and β. In the top left plot, an interval prior is shown where β is constrained within the interval [0.. 200] and γ is constrained within the interval [0..4]. Likewise, the parameter α (not seen) was constrained within the interval [200.. 400]. The prior assigns a probability of zero to the parameter values outside the intervals and a probability of $\frac{1}{200} \times \frac{1}{4} \times \frac{1}{200} = \frac{1}{160{,}000}$ inside the intervals. The top right plot represents a normal prior centered at $\gamma = 1.5$ and $\beta = 75$. The second row of each column shows the likelihood of a small data set of size $n = 4$ as a function of β and γ (in this plot, α was fixed at 300). If the data are all independently sampled and come from the same distribution (i. i. d. assumption), then the likelihood is given by:

$$l(\theta|\mathrm{X}) = \prod_i f(x_i|\theta) \quad (3)$$

where $f(x_i \mid \theta)$ is the pdf of the assumed distribution (e.g., Eq. 1) and x_i is the i^{th} item in the sample. In the middle row of Figure 2, the mode – the maximum of the function – is located at $\gamma = 1.42$ and $\beta = 79.3$ (at $\gamma = 1.45$, $\beta = 69.3$, and $\alpha = 289$ if α is free to vary). This is the maximum likelihood estimator.

By multiplying the two top surfaces and by dividing by $P(\mathbf{X})$ (a real number), the third row is obtained. This is the posterior distribution. Note that the distribution in the left column has two long tails, one in the direction of increasing γ and the other in the direction of increasing β. By using a normal prior (right column), the tails are almost non-existent. We will return to this when we examine the impact of priors in a later section.

The posterior distribution can be summarized by computing central tendency statistics. Often, the mean is computed. However, the median and the mode are also potentially useful statistics and we will argue that the mode is the most reliable for a Weibull distribution. To locate the mode of the posterior distribution, a search for the maximum of the function over the three parameters θ can be performed:

$$\max_{\theta\in\Theta} P(\theta|\mathrm{X}) = \max_{\theta\in\Theta} \frac{l(\theta|\mathrm{X}) \times P(\theta)}{P(\mathrm{X})}.$$

Because $P(\mathbf{X})$ is a constant independent of the parameters, it can be dropped, so that the mode is equally well localized by

$$\max_{\theta\in\Theta} l(\theta|\mathrm{X}) \times P(\theta) \quad (4)$$

where Θ is the domain of the parameters (i.e., $R^+ \times R^+ \times R$) for the parameters γ, β and α respectively). Eq. 4 is called the MAP estimator. In the case where there is no prior, i.e., when all the parameter values are equally likely, $P(\theta)$ becomes a constant and therefore can be dropped from the equation as well:

$$\max_{\theta\in\Theta} l(\theta|\mathrm{X}) \quad (5)$$

This last equation is exactly the MLE solution. It is a special case of BE if the mode is extracted from the posterior and if there is no prior.

More interesting is Eq. 4, the MAP estimator, which is a search for the mode of the likelihood weighted by the priors. If we consider a search for the logarithm of $l(\theta \mid \mathbf{X}) \times P(\theta)$, we get

$$\begin{aligned} &\max_{\theta\in\Theta} \log\left(l(\theta|\mathrm{X}) \times P(\theta)\right) \\ &= \max_{\theta\in\Theta} \left(\log l(\theta|\mathrm{X}) + \log P(\theta)\right) \end{aligned} \quad (6)$$

Because the log of a probability is a negative number, the quantity $\log P(\theta)$ can be interpreted as a penalty term. In particular, when the estimated parameters fall outside an interval prior, the penalty becomes log(0), that is, $-\infty$. Since

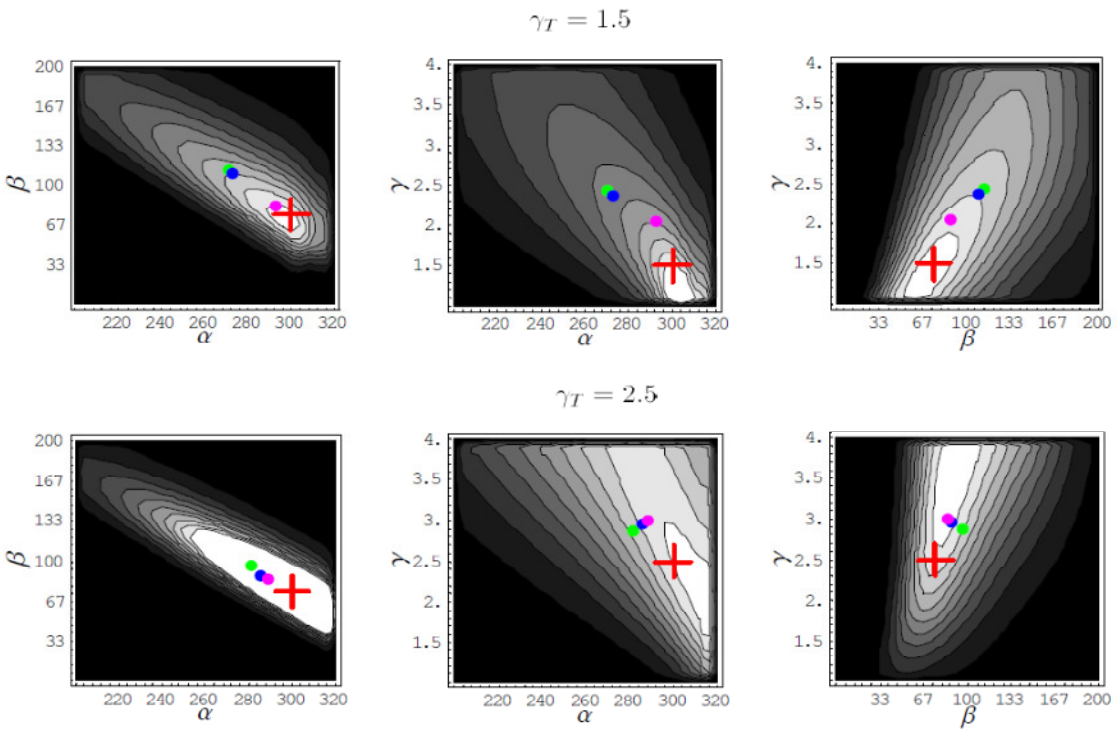

Figure 3. The expected posterior distribution estimated using 1,000 samples of size 8. The top row shows a true shape parameter $\gamma_T = 1.5$ and bottom shows a true shape parameter of $\gamma_T = 2.5$. Because the parameters are located in a three-dimensional space, the three projections on a plane are illustrated. The green dot shows the estimation obtained by using the mean of the expected posterior distribution, the blue dot shows the estimation obtained by using the median of the expected posterior distribution, and the magenta dot shows the estimation obtained by using the mode of the expected posterior distribution. The cross indicates the position of the true parameters $\{\gamma_T, \beta = 75, \alpha = 300\}$.

there cannot be worse penalty, the search returns inside the prior interval. However, whenever the estimated parameters fall near the center of mass of the prior, the penalty term plays only a marginal role in the estimation process. Hence, parameter values far away from the center of mass would be pushed back toward the center of mass (an effect termed *shrinkage* in Rouder et al., 2005) to increase the MAP estimation. The push (given by the penalty term) is stronger whenever the parameter value is unlikely according to the prior.

Unlike in BE, using the MAP estimator (Eq. 4) does not require the computation of the normalizing constant $P(\mathbf{X})$. One consequence is that the computation times are decreased by a factor of 1000 when using the MAP technique. However, unlike BE, only a point estimate is possible, and it can only be the mode. This is because computing the mean (or the median) requires integrating the posterior, a slower and more complex process achieved with numerical integration techniques (e. g. the Gauss-Kronrod algorithm or the Gibbs sampling algorithm). Nonetheless, MAP estimation is useful when the additional flexibility of BE is not strictly necessary. In the following section, we compare parameter point estimates of BE obtained using the mean, mode, and median in the context of the Weibull distribution.

Mean, Mode or Median?

The usefulness of the mode of the posterior distribution in recovering the true parameter values is assessed in this section in two ways. First, we examine how the mode compares with the mean and median to recover the true parameters of the expected posterior distribution. Second, simulations are run to estimate the best-fitting parameters describing individual samples of data using the mean, median, and mode of the posterior distributions. These estimates are then used to compute the bias and the efficiency of the point estimations obtained using the three central tendency statistics.

The expected posterior distribution

We propose using the *Expected Posterior Distribution* in order to have a mean to visualize the *ideal* posterior distribution. For large n, this notion is not required as the posterior distribution will be smooth. However, for small

sample sizes, the posterior distribution depends on the specifics of the sample and its appearance changes considerably from one sample to the other. To remedy this problem, we generated a large number of posterior distributions based on many different samples of the same size and then averaged the densities. Formally, the expected posterior distribution is given by a mixture:

$$\overline{P}(\theta|\mathrm{X}) = \frac{1}{m}\sum_{j=1}^{m} P(\theta|_j\mathrm{X})$$

where ${}_j\mathbf{X}$ is the j^{th} sample (j = 1..m). In the limit, we would get:

$$\lim_{m\to\infty}\frac{1}{m}\sum_{j=1}^{m} P(\theta|_j\mathrm{X}) = \int_X P(\theta|\mathrm{X})d\mathrm{X}$$

that is, the ultimate posterior based on all the possible samples from the sample space. However, we restricted ourselves to m = 1,000 in the simulations. We used a very small sample (n = 8) because we wanted the differences between the mean, the median and the mode to be easy to detect.

To build the expected posterior, we generated a thousand samples (m) each of size 8 (n). For each sample, we generated the posterior distribution using the interval prior of the top left plot seen in Figure 2. These posterior probabilities were evaluated and averaged at the intersections of a grid subdividing the 3D parameter space in 30 × 30 × 30 = 27,000 points. In one set of simulations, the true shape parameter was γ_T = 1.5; in the other, γ_T = 2.5. The former case corresponds to situations where the posterior is not normally distributed and the latter, to situations where the posterior should be normally distributed when the sample size is large. However, the difference between the mean, the median and the mode is smaller in the second case but still not zero, suggesting that for a small sample size (n = 8), the posterior distribution is not yet normal.

Figure 3 shows the projections on the three planes of the three dimensional expected posterior distribution. In both cases the distribution is elongated, suggesting the presence of parameter correlations (Bates and Watts, 1988). We also see that the distribution is truncated at γ = 4 by the interval prior. As noted by Rouder et al. (2005), the asymmetry changes very slowly when γ >> 3.602 so that a Weibull distribution with γ equals to, e.g., 10, does not differ much from a Weibull distribution with γ equals to 10,000. Rouder et al. suggested limiting the allowable γ < 5.

We also obtained point estimates of the parameters of the expected posterior distribution using the mean, median and mode. None of the estimates was equal to the true parameter values. For all three statistics (i.e., mean, median, mode), the estimated shift α was underestimated whereas the other two parameters were overestimated. More specifically, the bias was much smaller when using the mode than when using the mean or median when γ_T = 1.5. However, the differences in bias between the central tendency statistics was smaller when γ_T = 2.5. Note that the posterior estimates are constrained to be below γ = 4 by the interval prior. Had the shape parameter be allowed a larger range, the estimation bias when using the mean and the median would have increased correspondingly.

To assess the magnitude of the biases, as well as the general precision of the estimates, we now turn to simulations in which the parameters are estimated for each sample individually.

Bias and efficiency of the estimates

Three series of simulations were run to examine the quality of the parameter estimates obtained by using the mean, the median and the mode of the posterior distribution. As in the previous section, we used a small sample size (n = 8). Three shape parameters were explored: γ_T = 1.5 and γ_T = 2.5 as previously, but also an intermediate case, γ_T = 2.0. In each simulation, a random sample of size 8 was generated with true parameters {γ_T, β, α}. The parameters β and α are scaling parameters and were fixed at 100 and 300 respectively. The resulting sample was then analyzed using BE and point estimates of the distribution parameters were obtained using the mean, the median and the mode of the posterior distribution. This procedure was repeated a thousand times. The prior used in BE was an interval prior {1 ≤ γ ≤ 4}, {0 ≤ β ≤ 200} and {0 ≤ α ≤ 400}. Additionally, because α cannot exceed the smallest item in the sample, the upper bound was the smallest of 400 and the smallest of the sample.

The mode was located using the Simplex (Nelder and Mead, 1965). The mean of the posterior distribution was obtained by extracting the marginal distribution of a single parameter (integrating out the other two) and then computing the mean value of that univariate distribution. For instance, the mean of the shift parameter was obtained by getting its marginal distribution $f(\alpha|\mathrm{X}) = \int_0^{200}\int_1^4 P(\theta|\mathrm{X})d\gamma d\beta$ and then the mean was obtained with $\hat{\alpha}_{Mn} = \int_0^{Min(Min(\mathrm{X}),400)} \alpha f(\alpha|\mathrm{X})d\alpha$. The limits of the integral correspond to the intervals of the priors. The median of a single parameter was obtained by first getting its marginal distribution as above, then the cumulative density function (cdf, e.g., $F(\alpha|\mathrm{X}) = \int_0^{\alpha} f(\alpha|\mathrm{X})d\alpha$), and finally searching for the value at which the cdf equals 1/2. (e.g., $\hat{\alpha}_{Md}$ such that $F(\hat{\alpha}_{Md}|\mathrm{X}) = \frac{1}{2}$).

We summarized the results by first computing the mean of the estimated parameters obtained using the mean, mode,

Table 1. Parameter estimates for three central tendency measures as a function of the true shape parameter.

		Mean parameter estimates			Reliability		
True scale	Statistics	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\alpha}$	Bias	Eff	RMSE
$\gamma_T = 1.5$	Mode	1.63	106.5	299.9	6.5	34.0	34.6
	Median	2.05	133.9	283.8	37.6	29.9	48.0
	Mean	2.09	133.3	275.7	41.2	21.8	46.6
$\gamma_T = 2.0$	Mode	2.57	110.2	290.9	13.7	35.0	37.6
	Median	2.60	126.5	281.6	32.3	21.1	38.5
	Mean	2.59	127.9	276.9	36.2	18.4	40.6
$\gamma_T = 2.5$	Mode	3.21	115.9	285.0	21.9	32.4	39.1
	Median	2.96	119.6	287.0	23.5	21.1	31.6
	Mean	2.90	122.6	281.4	29.3	18.2	34.5

Note: the true parameters are {γ_T, 100, 300}

and median. The results are shown in the three leftmost columns of Table 1. As seen, the parameters γ and β were always overestimated and α was always underestimated (although negligibly so in one condition). For the mode, the bias was small but tended to increase with increasing γ_T. For the mean and median, the bias was very large but diminished as γ_T increases. Remember that the parameter γ was bounded from above at 4, favoring fewer overestimation and therefore smaller biases.

Next, we computed (a) the three-dimensional bias, i.e., the distance between the mean of the estimated parameters and the true parameters in a 3D space and (b) the three-dimensional efficiency of the estimates, i.e., the variability in the distance between the individual estimates and the mean estimate in a 3D space. Formally:

$$Bias = \left\| \overline{\theta} - \theta_T \right\|$$

$$Eff = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left\| \theta_i - \overline{\theta} \right\|^2}$$

where $\overline{\theta}$ is the mean estimated parameter vector, θ_T is the true parameter vector, and θ_i is the estimated parameter vector from the i^{th} sample (i = 1 .. m, with m = 1000). The symbol $\left\| \bullet \right\|$ denotes the Euclidian distance. From these definitions, the global error of prediction, i.e., the Root Mean Squared Error (RMSE) is equal to:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left\| \theta_i - \theta_T \right\|^2} = \sqrt{Bias^2 + Eff^2}$$

These results are shown in the three rightmost columns of Table 1. First, we note that efficiency is smaller (i.e. the estimates are more stable) when the mean posterior distribution is computed. This was a predictable result since, for any distribution, the standard deviation about the mean is smaller than the standard deviation about any other value (Cramèr, 1947). Further, as γ_T increased, the advantage in efficiency of the mean over the mode became more important. This advantage in efficiency is however accompanied by an important bias (the bias is nearly five times as important for small γ_T compared to the modal estimates' bias). The RMSE reflects this fact: for small γ_T, the advantage of the modal estimates is important. For large γ_T, the differences tend to reverse. However, the empirical data generally do not show large γ_T (the largest mean γ_T in Logan, 1992, was 2.264). So, for the kind of RT data typically obtained in social sciences, the mean estimates are the most biased and this is caused by the long tail of the posterior distribution in the particular case of the Weibull distribution. Using an interval prior presumably reduced the impact of this long tail. We examine in the next section the impact of more aggressive priors. Finally, the bias and efficiency of estimates obtained using the median tended to be somewhere between those obtained with the mean and mode, but the median's middle of the road scores sometimes resulted in smaller RMSE.

Impact of priors, right or wrong

In this section, we examine the influence of normal priors on the estimates. A good set of priors (e.g., that model

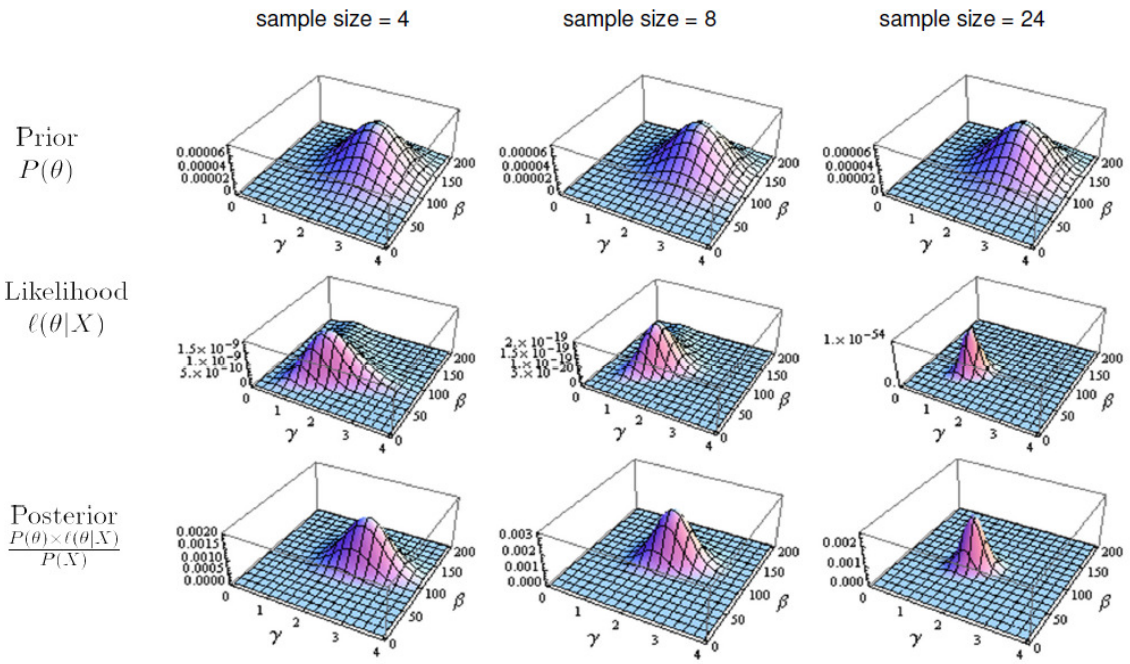

Figure 4. The steps to obtain a posterior distribution using the same format as in Figure 2. Here, the prior is a normal distribution centered on {γ = 2.5, β = 125} with variances of {0.5, 1250} and no covariation. The prior on the parameter α was an interval [200..400]. The posterior distribution is more similar to the prior for small samples than for large samples.

the parameters of the population accurately) should increase the precision of the estimates (smaller bias and hence, smaller RMSE, as shown next). Another advantage of a normal prior with a reasonably small variance is that it will "truncate" the long tails of the likelihood function (as was seen in Figure 2, second row). This in turn reduces the probability of outlier estimates so that the efficiency will be improved (smaller variance of the estimates). However, one risk of using a normal prior is that it could determine the outcome entirely. Figure 4 illustrates this last danger (in a format similar to that of Figure 2). The top row shows a normal prior centered at {γ = 2.5, β = 125} (the parameter α, not shown, is modeled by an interval prior on [200..400]). As in Figure 2, the middle row shows the likelihood function for three samples of increasing sizes. These samples were taken from a Weibull-distributed population with true parameters {γ = 1.5, β = 75, α = 300}. In all three cases, the likelihood function is reasonably well centered above the true parameters of the population. The bottom row shows the posterior distributions. As seen, the smaller the sample size, the more similar the posterior is to the prior.

This influence of the prior on the estimate is best understood with Eq. 6. In that equation, log $P(\theta)$ is a penalty term that increases with the distance between the point estimates and the mode of the prior distribution. However, the magnitude of the penalty term is independent of sample size. The log of the likelihood, on the other hand, becomes more peaked with increasing sample sizes so that in the limit ($n \rightarrow \infty$), the area surrounding the peak is zero. As such, the prior no longer has any influence on the estimates. In practical applications, however, it is not known if there is a critical sample size beyond which the prior has no influence anymore. The effects of biased priors and sample size are explored next.

Simulation method

We ran simulations to explore the effect of sample size and biased priors. We used three priors: two normal priors centered on θ_{Low} = {γ = 1.5, β = 75, α = 300} and θ_{High} = {γ = 2.5, β = 125, α = 300}, and the interval prior used in the previous section. In half of the simulations, the true parameters used to sample random deviates were θ_{Low}; in the other half, θ_{High} was used. The sample size was also varied from a very small sample size (n = 1) to a very large sample size (n = 128). The sample sizes used were 1, 8, 16, 32, 64, 80, 96 and 128. In all the conditions of true parameters by sample size (2 × 8 conditions), we generated 1000 (m) samples from which estimates were obtained using the three priors (for a total of 48 estimates). In all simulations, we estimated the parameters from the mode of the posterior distribution using MAP estimation. The results were then summarized using the (three-dimensional) bias and efficiency as in the previous section.

Hypothesis on sample size

Figure 4 suggests that, with small sample sizes, the priors determine the estimates. Hence, when the prior is not based on the true parameters, the three-dimensional bias should equal the distance between θ_{High} and θ_{Low}, that is $\|\theta_{High} - \theta_{Low}\|$ which is equal in this case to 50.01. In addition, the efficiency of the estimates should equal the total variation found in the priors. Because the priors were modeled with a variance on γ of 0.5 and a variance on β of 1,250, the total variation of the priors should be 49.5. These predictions should occur for very small sample sizes. The first goal of the simulations is to verify this hypothesis.

Hypothesis on biased priors

Figure 4 also suggests that biased priors can be harmful with small sample sizes. However, biased priors may be helpful with known biased estimators. For instance, we have shown earlier that the estimates obtained from BE (or its special case, MAP estimation) are biased upward for γ and β. Keeping this bias in mind, it may be possible to use priors for γ and β that are smaller than the true γ and β to compensate for this bias. The second goal of these simulations is to test this possibility.

Simulation results

The results of the simulations are shown in Figure 5. The most striking aspect of the plots is the leveling of the curves for sample sizes 80 and above. Both bias and efficiency became nearly flat with no influence of the priors injected in the process. Hence, answering the first question above, there is a critical sample size past which the priors stop affecting the MAP estimate. This critical sample size is n = 80. This is rather small and would suggest that all the asymptotic properties of the MLE apply past n = 80 (remember that, following Smith, 1985, the Weibull distribution has the usual MLE properties only when the true population shape parameter γ exceeds 2).

Interestingly, Figure 5 also suggests an interaction involving the location of the prior. First, correct priors (centered at the true parameter values) improve bias markedly. The bias is the smallest in our simulations for a sample size of 8. Second, incorrect priors do not necessarily produce bad estimates: when the prior accentuate the tendency of MLE to overestimate γ and β, the results are very biased. For a sample size of 1, the bias is 50.0, equal to the distance between the true parameters and the parameters of the prior, suggesting in this case that the prior uniquely determines the estimates. However, when the prior is opposing the tendency of BE and MLE to overestimate γ and β, bias is bad for very small samples (n = 1 or n = 8) but much reduced for medium sample sizes (n = 32 or n = 64). Figure 5 suggests that at n = 64, the pull of the prior is equal (in the opposite direction) to the push of the bias inherent to BE and MLE so that they nearly cancel each other out. The optimal sample size in our simulations (n = 64) is however certainly dependent on the distance between the true parameters and the parameters of the priors. Had the distance been larger (or smaller), the minimum of the function would have occurred at a smaller (respectively larger) sample size. Hence, for all practical application, when devising priors, the modeler should be very conservative and underestimate the parameters γ and β in proportion to the uncertainty on their true values and in proportion to the smallness of the available sample.

Regarding efficiency, any normal prior is preferable to an interval prior. Such priors diminish the influence of long tails, resulting in less variable (i.e., more efficient) estimates. Also, efficiency was not affected much by the quality of the priors (among normal priors). As such, RMSE mainly reflected the bias term in the estimates.

Discussion

In this tutorial, we reviewed a seldom-used technique for parameter estimation in social sciences, namely MAP estimation. MAP estimation is an extension of the regular MLE technique which can be used in conjunction with priors of any types. Likewise, it is a special case of the Bayesian estimation technique from which only the modal value of the posterior distribution can be obtained. The results of the example simulations included in this tutorial suggest that the full capabilities of BE technique is unnecessary for the case of the Weibull distribution. Modal estimates were accurate and priors could be injected into the MLE technique directly. This conclusion is useful because BE is difficult to implement and slow to operate: It requires the numerical estimation of a large number of nested integrals (none available in closed form when the Weibull distribution is assumed) or the use of Markov Chain Monte Carlo techniques. Such calculations are slow (estimating the mean of the posterior distribution is approximately 1,000 times slower than estimating the mode using MLE or MAP) and become a real concern in models involving convolved stages.

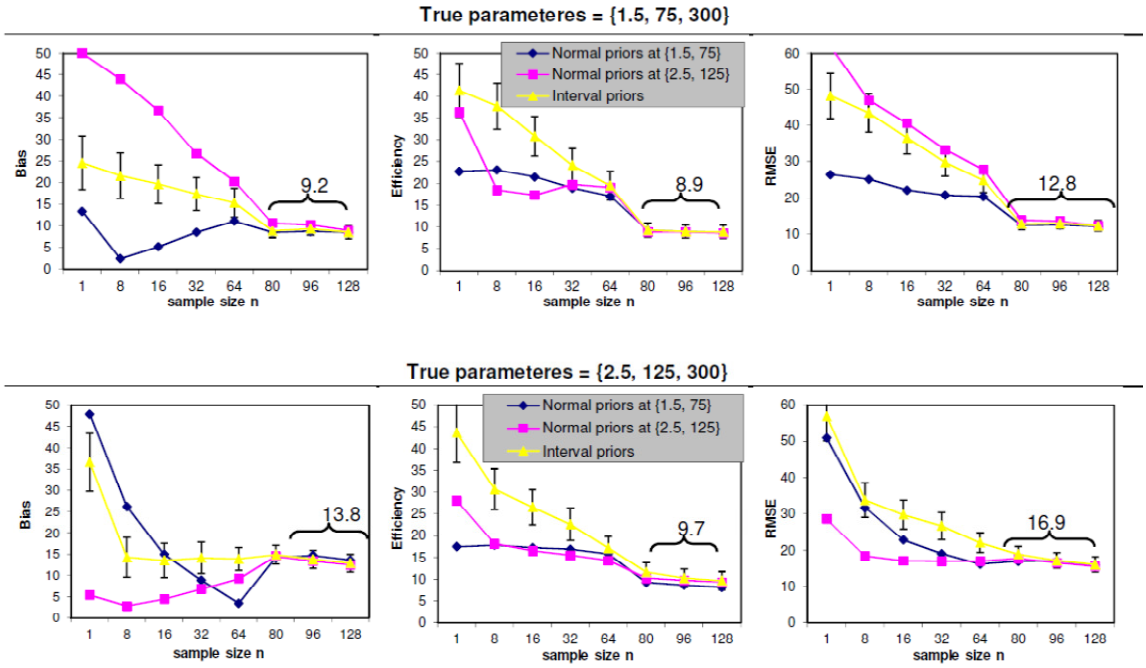

Figure 5. Bias, Efficiency and RMSE of the estimates as a function of the true parameters (rows) and the sample sizes. In each condition, the statistics were calculated with two different normal priors and an interval prior. The numbers above the curled braces are the means of the 9 conditions underneath the brace. Only the priors on the parameters γ and β were varied; the prior on α was an interval on [200, 400]. If the priors perfectly determined the estimates, then Bias should be 50.01 and Efficiency should be 49.5 for biased priors. Error bars are the standard error of the RMSE.

However, by using MAP estimation instead of BE, it is no longer possible to compute the posterior distribution of the parameters, only a point estimation is returned. This lost of information follows from the maximization procedure which summarizes the entire distribution using its mode. Because the mode of the posterior distribution is used, the amount of information lost is a negative function of the sample size. It is well-known in Bayesian statistics that the precision of a measurement is the reciprocal of the posterior distribution's variance, and that this measure often diminishes as new data is made available (Edwards et al., 1963; Jeffreys, 1961). While there is no general formula to describe the diminution of the posterior distribution's variance following Bayesian updating, it is usually very fast so that the posterior mode can adequately summarize the posterior distribution as soon as a fairly small amount of data is available. This result was confirmed by the included simulations which showed that modal estimates are less biased than the mean (or the median) of the posterior distribution. Overall, errors of estimation (measured by RMSE) were favorable to modes when γ_T was smaller than 2 and comparable when γ_T was larger than 2.

A limitation of MAP estimation that was not discussed previously concerns the construction of confidence intervals. Because BE provides a full distribution, the posterior can be used to infer confidence intervals and standard errors. Such quantities cannot be derived directly for MLE or MAP. However, it is possible to estimate standard errors using the Hessian matrix of the MAP estimators, as with regular MLE (see Dolan and Molenaar, 1991, Rose and Smith, 2001). It is our hope that this tutorial will increase the use of MAP estimation in social sciences, and future work should be devoted to testing the efficacy of the posterior mode in estimating parameters for other common parametric models in psychology.

References

Bates, D. M. & Watts, D. G. (1988). *Nonlinear regression analysis and its application*. New York: J. Wiley and son.

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. New York: Oxford University Press.

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.

Cousineau, D. (2004). Merging race models and adaptive networks: A parallel race network. *Psychonomic Bulletin & Review*, *11*, 807-825.

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.

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

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

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Dolan, C. (2000). *DISFIT version 1.0: Provisional program documentation* (Technical report series No. not numbered). University of Amsterdam: Department of psychology.

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 and Statistical Psychology*, *44*, 359-368.

Edwards, W., Lindman, H., Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.

Heathcote, A. (1996). *RTSYS:* A computer program for analysing response time data. *Behavior Research Methods, Instruments, & Computers*, *28*, 427-445.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

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.

Huber, D. E., & Cousineau, D. (2003, June). A race model of perceptual forced choice reaction time. *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, Boston.

Jeffreys, H. (1961). *Theory of Probability. Third Edition.* Glasgow: Oxford University Press.

Kiefer, N. M. (2005). *Maximum likelihood estimation (MLE),* Internet resource found at http://instruct1.cit.cornell.edu/courses/econ620/reviewm5.pdf, last consulted 6/october/2006.

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.

Marley, A. A. J. (1989). A random utility family that include many of the "classical" models and has closed form choice probabilities and choice reaction times. *British Journal of Mathematical and Statistical Psychology*, *42*, 13-36.

Miller, J., Ulrich, R. (2003). Simple reaction time and statistical facilitation: A parallel grains model. *Cognitive Psychology*, *46*, 101-151.

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

Neapolitan, R. E. (2004). *Learning Bayesian Networks*. Upper Saddle River, NJ: Prentice Hall.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*, 308-313.

Rockette, H., Antle, C., & Klimko, L. A. (1974). Maximum likelihood estimation with the Weibull model. *Journal of the American Statistical Association*, *69*, 246-249.

Rose, C., & Smith, M. D. (2000). Symbolic maximum likelihood estimation with Mathematica. *The Statistician*, *49*, 229-240.

Rose, C., Smith, M., D. (2001). *Mathematical Statistics with Mathematica*. New York: Springer-Verlag.

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.

Smith, R. L. (1985). Maximum likelihood estimation in a classe of nonregular cases. *Biometrika*, *72*, 67-90.

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

Wolfram Research inc. (2011) . Wolfram Mathematica 8 [Computer software].

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

Manuscript received 31 March 2013.