Confidence intervals, significance values, maximum likelihood estimates, etc. sharpened into Occam's razors

David R. Bickel

May 24, 2018

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology, and Immunology

Department of Mathematics and Statistics

University of Ottawa

451 Smyth Road

Ottawa, Ontario, K1H 8M5

+01 (613) 562-5800, ext. 8670

dbickel@uottawa.ca

Abstract

Confidence sets, p values, and maximum likelihood estimates may be adjusted to favor sampling distributions that are simple compared to others in the parametric family. The adjustments are derived from a prior likelihood function previously used to adjust posterior distributions.

Keywords: algorithmic probability; confidence distribution; differential entropy; fiducial inference; Kolmogorov complexity; probability matching prior distribution

1 Introduction

Occam's razor is the principle that simpler theoretical formulations are preferred to more complex ones when other things are equal (Baker, 2016). The principle has been invoked to support a wide variety of statistical methods. First, Bayesian (e.g., Wasserman, 2000; Ando, 2010), frequentist (e.g., Burnham and Anderson, 2002; Claeskens and Hjort, 2008), and information-theoretic methods of model selection such as that of minimum message length (Wallace, 2005; Dowe, 2011) have been interpreted as ways of choosing the parametric complexity of a model to optimize predictive accuracy, where parametric complexity increases with the number of free parameters. Second, simpler models are often used to enhance understanding even at the expense of predictive accuracy when a more accurate model would require so many parameters that interpretation would be difficult or impossible (Lindley, 2000, discussion by D. R. Cox). Third, the intuitive appeal of invariance and other probabilistic symmetries in statistics (e.g., Eaton, 1989; Helland, 2009) may be attributed to a different type of simplicity, that of mathematical elegance.

Another type of simplicity is that of a hypothesized distribution. Should simpler distributions of data have higher prior probabilities or prior probability densities, other things equal? If so, what is the effect on frequentist inference, especially in the form of p values, confidence intervals, and maximum likelihood estimates? This paper addresses the second question.

Assuming entropy as a measure of the complexity of sampling distributions should influence prior distributions as Bickel (2016) suggests, the impact on credible sets such as 95% credible intervals can be calculated from the posterior distribution. When the confidence sets such as 95% confidence intervals are approximately credible sets according to some prior distribution, the confidence sets may be adjusted for entropy by replacing them with the credible sets generated from that prior distribution after it has been adjusted to account for entropy.

Prior distributions generating to credible sets that match confidence sets in that way are called *matching prior distributions*. Why restrict frequentist inference to methods that have a Bayesian interpretation using a matching prior? With a "synchronic coherence" condition, automated decisions on the basis of confidence levels and p values leads to minimizing expected loss with respect to coherent fiducial distributions such as a class of confidence distributions (Bickel and Padilla, 2014). If those decisions also satisfy a "diachronic coherence" condition regarding self-consistency as data arrive in time, then the decisions minimize expected loss with respect to a posterior distribution corresponding to a matching prior. A related reason to consider matching priors is that frequentist methods compatible with them have desirable conditional inference properties (Datta and Sweeting, 2005, §2).

The statistical background for the proposed procedure is explained in Section 2. The method for adjusting a prior distribution for simplicity and then applying that to the adjustment of frequentist procedures using a matching prior is offered in Section 3. Section 4 proposes alternative procedures, covering ways to adjust statistical inference such as maximum likelihood estimation for the simplicity of distributions without requiring a matching prior distribution.

2 Background concepts

2.1 Approximate confidence distributions and probability matching priors

Consider sets Θ and Γ of real numbers or vectors, a parameter of interest $\theta \in \Theta$, and a nuisance parameter $\gamma \in \Gamma$. The observed sample x is modeled as a realization of X, a random vector of distribution $P_{\theta,\gamma}$, i.e., $X \sim P_{\theta,\gamma}$. If the components are independent and distributed as $P_{\theta,\gamma}^{(1)}$, then the distribution of X is $P_{\theta,\gamma}^{(1)}$'s n-product $P_{\theta,\gamma}^{(n)}$, written here as $P_{\theta,\gamma}$ for short. An *approximate confidence curve* is a function $(\theta, x) \mapsto p(\theta; x)$ such that, to some order of approximation $(\doteq)$,

$$P_{\theta,\gamma}\left(p(\theta; X) < \alpha\right) \doteq \alpha \tag{1}$$

for all $\alpha \in]0,1[$, following the concept of confidence curves used in Birnbaum (1961) and Blaker (2000). Thus, $p(\theta_0; x)$ is an observed p value for testing the null hypothesis that $\theta = \theta_0$. It follows from equation (1) that the random set $\mathcal{C}(1-\alpha; X)$ defined by the function

$$\alpha, x \mapsto \mathcal{C}(1-\alpha; x) = \{\theta \in \Theta : p(\theta; x) \geq \alpha\} \tag{2}$$

is an approximate confidence set for θ in the sense that

$$P_{\theta,\gamma}\left(\theta \in \mathcal{C}(1-\alpha; X)\right) \doteq 1-\alpha \tag{3}$$

for all $\alpha \in]0,1[$. For an observed sample x, a probability distribution $P(\bullet; x)$ on a measurable space $(\Theta, \mathfrak{H})$ is an *approximate confidence distribution* if

$$P(\vartheta \in \mathcal{C}(1-\alpha; x); x) = P(\mathcal{C}(1-\alpha; x); x) \doteq 1 - \alpha \tag{4}$$

for all $\alpha \in]0,1[$, where $\vartheta \sim P(\bullet; x)$. Together, equations (3) and (4) say the probability that the random parameter ϑ is in an observed confidence set $\mathcal{C}(1-\alpha; x)$ is approximately equal to the probability that an approximate confidence set $\mathcal{C}(1-\alpha; X)$ covers the true value θ. An approximate confidence distribution is a special case of what Bickel and Padilla (2014) call a "confidence distribution" that is a Kolmogorov probability distribution as opposed to merely an incomplete probability distribution. Such confidence distributions are in turn special cases of "basic fiducial distributions" (Bickel and Padilla, 2014).

The order of approximation may be formalized in various ways. For example, if θ is a scalar and if equation (3) is understood to mean $p(\theta; X)$ weakly converges to $\mathrm{U}(0,1)$, then the approximate confidence distribution is isomorphic to the "asymptotic confidence distribution" of Singh et al. (2005). The definition of the order of approximation is left open herein in order to make the following connection to a wide variety of probability matching prior distributions.

Let $P^{\pi_0}(\bullet|X = x)$ denote the posterior distribution of θ according to applying Bayes's theorem to a prior density π_0, a function of θ and γ. Then π_0 is called a *probability matching prior distribution* if $P^{\pi_0}(\bullet|X = x)$ is an approximate confidence distribution. Classes of probability matching priors thus correspond to different definitions of the order of approximation ($\doteq$) such as the definitions found in Datta and Sweeting (2005) and Ghosh (2011).

2.2 Extended evidence values

The measure of evidence proposed in De Bragança Pereira and Stern (1999) is extended by generalizing its highest density regions to regions of the form

$$(\pi_0, \alpha, x) \mapsto \mathcal{C}^{\pi_0}(1-\alpha; x) = \{\theta \in \Theta : q^{\pi_0}(\theta; x) \geq \beta(\alpha)\} \tag{5}$$

for some real-valued functions $(\pi_0, \theta, x) \mapsto q^{\pi_0}(\theta; x)$ and $\alpha \mapsto \beta(\alpha)$ such that $\mathcal{C}^{\pi_0}(1-\alpha; x)$ is a (100%) $(1-\alpha)$ credible set in the sense that $P^{\pi_0}(\theta \in \mathcal{C}^{\pi_0}(1-\alpha; x) | X = x) = 1 - \alpha$ for all $\alpha \in]0,1[$. Define the *extended evidence value* at a fixed $\theta_0 \in \Theta$, with respect to a prior π_0, by

$$p^{\pi_0}(\theta_0; x) = 1 - \inf_{\alpha \in]0,1[: \theta_0 \in \mathcal{C}(1-\alpha; x)} P^{\pi_0}(\theta \in \mathcal{C}^{\pi_0}(1-\alpha; x) | X = x) = \sup\{\alpha \in]0,1[: \theta_0 \in \mathcal{C}^{\pi_0}(1-\alpha; x)\}. \tag{6}$$

If π_0 is a probability matching prior density function and $\mathcal{C}^{\pi_0}(1-\alpha; x) = \mathcal{C}(1-\alpha; x)$, then equation (2) implies that in this case the extended evidence value is a p value:

$$p^{\pi_0}(\theta_0; x) = \sup\{\alpha \in]0,1[: \theta_0 \in \mathcal{C}(1-\alpha; x)\} = \sup\{\alpha \in]0,1[: p(\theta_0; x) \geq \alpha\} = p(\theta_0; x). \tag{7}$$

3 Sharpened priors, sharpened p values, and sharpened confidence sets

A prior density function π_0 is considered *blunt* if its specification does not reflect the simplicity of a single-observation distribution $P_{\theta,\gamma}^{(1)}$ as a function of θ and γ. A prior density function π that is *sharpened* with respect to π_0 is defined by

$$\theta, \gamma \mapsto \pi(\theta, \gamma) \propto \pi_0(\theta, \gamma) e^{-H(\theta,\gamma)}, \tag{8}$$

where $H(\theta, \gamma)$ may be the Shannon entropy of $P_{\theta,\gamma}^{(1)}$ if θ and γ are discrete. Otherwise, letting ξ denote a measure that dominates $P_{\theta,\gamma}$ and letting $f_{\theta,\gamma} = dP_{\theta,\gamma}^{(1)}/d\xi$ denote the relevant probability density function,

$$H(\theta, \gamma) = -\int f_{\theta,\gamma}(x) \ln f_{\theta,\gamma}(x)\, d\xi(x),$$

commonly known as *differential entropy* when ξ is the Lebesgue measure. Bickel (2016) argued on the basis of Kolmogorov complexity for using sharpened priors in place of blunt priors.

The method may be generalized by replacing $\theta, \gamma \mapsto e^{-H(\theta,\gamma)}$ with $\theta, \gamma \mapsto e^{-\kappa H(\theta,\gamma)}$ for some $\kappa > 0$ or with some other function that monotonically increases with the simplicity of $P_{\theta,\gamma}^{(1)}$. The function $\theta, \gamma \mapsto e^{-\kappa H(\theta,\gamma)}$ with $\kappa = 1$ is used as a default in the rest of this paper, without loss of generality.

In analogy with the extended evidence value of equation (6), the *sharpened evidence value* at θ with respect to π is

$$p^{\pi}(\theta;x)=\sup\left\{\alpha\in\left]0,1\right[:\theta_0\in\mathcal{C}^{\pi}(1-\alpha;x)\right\},$$

where $(\pi,\alpha,x)\mapsto\mathcal{C}^{\pi}(1-\alpha;x)$ is the function defined by equation (5). Thus, $\mathcal{C}^{\pi}(1-\alpha;x)$ is the highest-$q^{\pi}(\bullet;x)$ $(100\%)(1-\alpha)$ credible set in the sense that $P^{\pi}(\theta\in\mathcal{C}^{\pi}(1-\alpha;x)|X=x)=1-\alpha$ for all $\alpha\in\left]0,1\right[$. For example, if $q^{\pi_0}(\theta;x)$ is the posterior probability density when π_0 is the prior, then $q^{\pi}(\theta;x)$ is the posterior probability density when π is the prior.

Now assuming that π_0 is a probability matching prior density function and that $\mathcal{C}^{\pi_0}(1-\alpha;x)=\mathcal{C}(1-\alpha;x)$, the prior density function π that is sharpened with respect to π_0 is a *sharpened matching prior distribution*, and $P^{\pi}(\bullet|X=x)$, the corresponding posterior distribution, is a *sharpened confidence distribution*, the simplicity-informed counterpart to $P^{\pi_0}(\bullet|X=x)$. Since equation (7) indicates that in this case the extended evidence value is a p value, the corresponding $p^{\pi}(\theta;x)$ may be considered a *sharpened p value* and $\mathcal{C}^{\pi}(1-\alpha;x)$ a *sharpened confidence set* of level $(100\%)(1-\alpha)$.

Example 1. Let X denote a sample of n independent draws from $P^{(1)}_{\theta,\gamma}=\mathrm{N}(\theta,\gamma)$, the normal distribution of unknown mean θ, the parameter of interest, and unknown variance γ. The maximum likelihood estimates of θ and γ are denoted by $\widehat{\theta}$ and $\widehat{\gamma}$. According to the t test, the two-sided p value for testing the null hypothesis that $\theta=\theta_0$ is

$$p(\theta_0;x)=2\min\left(\Phi_{\theta_0,\widehat{\gamma}_{n-1}/n,n-1}\left(\widehat{\theta}/\sqrt{\widehat{\gamma}_{n-1}/n}\right),1-\Phi_{\theta_0,\widehat{\gamma}_{n-1}/n,n-1}\left(\widehat{\theta}/\sqrt{\widehat{\gamma}_{n-1}/n}\right)\right),\tag{9}$$

where $\widehat{\gamma}_{\nu}=\widehat{\gamma}n/\nu$, and $\Phi_{\mu,\sigma^2,\nu}$ is the cumulative distribution function (CDF) of the Student t distribution with location parameter μ, scale parameter σ, and ν degrees of freedom. The corresponding $(100\%)(1-\alpha)$ confidence interval is

$$\mathcal{C}(1-\alpha;x)=\{\theta_0\in\Theta:p(\theta_0;x)\geq\alpha\}=\left[\Phi_{\widehat{\theta},\widehat{\gamma}_{n-1}/n,n-1}\left(\frac{\alpha}{2}\right),\Phi_{\widehat{\theta},\widehat{\gamma}_{n-1}/n,n-1}\left(1-\frac{\alpha}{2}\right)\right].\tag{10}$$

In order to sharpen $p(\theta_0;x)$ to take into account the simplicity of $\mathrm{N}(\theta,\gamma)$ as θ and γ vary, a matching prior distribution resulting in a $(100\%)(1-\alpha)$ credible set equal to $\mathcal{C}(1-\alpha;x)$ is required. Box and Tiao (1992, §2.4.6) considered improper priors of probability density $\pi_0^{(\delta)}(\theta,\gamma)\propto\gamma^{-(\delta+1)/2}$ for $\delta\geq0$ and their posterior CDFs $\Phi_{\widehat{\theta},\widehat{\gamma}_{n+\delta-1}/n,n+\delta-1}$. The $\delta=0$ case, $\pi_0^{(0)}$, is a probability matching prior since the resulting posterior distribution of θ, also a confidence distribution of θ, has CDF $\Phi_{\widehat{\theta},\widehat{\gamma}_{n-1}/n,n-1}$, leading to $\mathcal{C}\left(1-\alpha;x\right)$ as the highest-density (100%) $(1-\alpha)$ credible set, that is, $\mathcal{C}\left(1-\alpha;x\right)=\mathcal{C}^{\pi_0^{(0)}}\left(1-\alpha;x\right)$. Using $\mathrm{N}\left(\theta,\gamma\right)$'s differential entropy, $H\left(\theta,\gamma\right)=\ln\gamma^{1/2}$ up to a constant (Michalowicz et al., 2013, p. 127), the corresponding sharpened matching prior density is

$$\pi^{(0)}\left(\theta,\gamma\right)\propto e^{-H(\theta,\gamma)}\pi_0^{(0)}\left(\theta,\gamma\right)\propto\gamma^{-1/2}\gamma^{-(0+1)/2}=\gamma^{-(1+1)/2}\propto\pi_0^{(1)}\left(\theta,\gamma\right),$$

which is $\pi_0^{(\delta)}\left(\theta,\gamma\right)$ with $\delta=1$. It follows that the sharpened confidence distribution is of CDF $\Phi_{\widehat{\theta},\widehat{\gamma}_n/n,n}$, that the sharpened p value is

$$p^{\pi^{(0)}}\left(\theta_0;x\right)=2\min\left(\Phi_{\theta_0,\widehat{\gamma}_n/n,n}\left(\widehat{\theta}/\sqrt{\widehat{\gamma}/n}\right),1-\Phi_{\theta_0,\widehat{\gamma}_n/n,n}\left(\widehat{\theta}/\sqrt{\widehat{\gamma}/n}\right)\right),\tag{11}$$

and that the sharpened (100%) $(1-\alpha)$ confidence set is

$$\mathcal{C}^{\pi^{(0)}}\left(1-\alpha;x\right)=\left\{\theta_0\in\Theta:p^{\pi^{(0)}}\left(\theta_0;x\right)\geq\alpha\right\}=\left[\Phi_{\widehat{\theta},\widehat{\gamma}_n/n,n}\left(\frac{\alpha}{2}\right),\Phi_{\widehat{\theta},\widehat{\gamma}_n/n,n}\left(1-\frac{\alpha}{2}\right)\right].\tag{12}$$

Figures 1 and 2 indicate that while the sharpened confidence distributions and sharpened confidence intervals differ markedly from their blunt counterparts for $n=2$ observations, they become closer by $n=4$ observations but still with substantial differences.

The requirement of $n\geq2$ imposed by equations (9) and (10) prevents equations (11) and (12) from degenerating due to the fact that $\widehat{\gamma}_n=\widehat{\gamma}=0$ when $n=1$. That is a safeguard that the frequentist aspect of the proposed method puts in place, a safeguard missing in this case from a purely Bayesian approach to improper priors. ▲

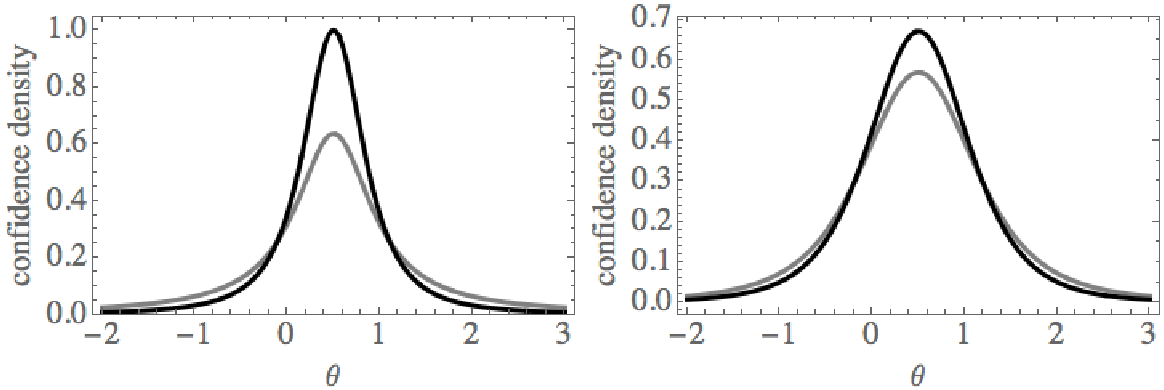

Figure 1: Probability density function of the normal mean for the blunt confidence distribution (gray) and the sharpened confidence distribution (black). The samples are $x = (0, 1)$ on the left and $x = (-1, 0, 1, 2)$ on the right.

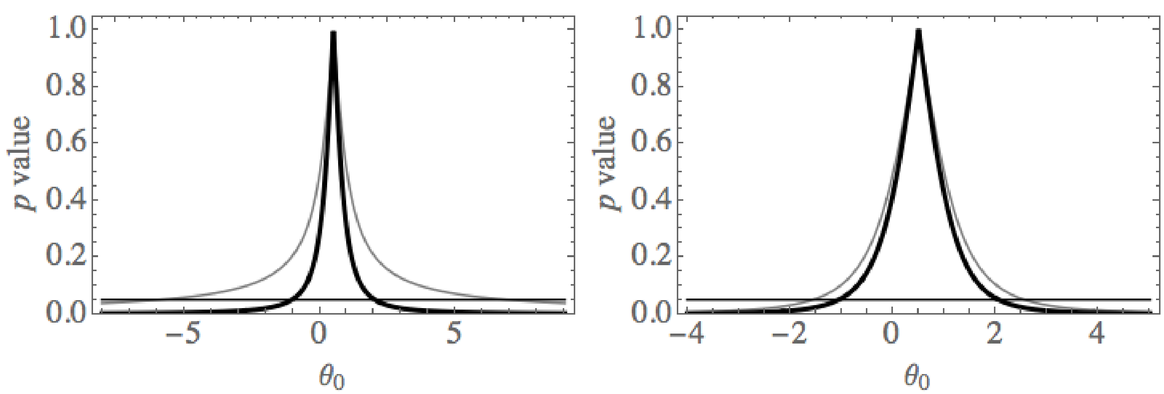

Figure 2: Observed p values for testing whether the normal mean is θ_0 for the blunt p value function (gray) and the sharpened p value function (black). Intersections with the horizontal line indicate the 95% confidence intervals. The samples are $x = (0, 1)$ on the left and $x = (-1, 0, 1, 2)$ on the right.

4 Sharpened statistical inference with prior likelihood instead of prior probability

4.1 Sharpened likelihood function

Equation (8) may be interpreted in terms of Bayes's theorem as updating the prior density function π_0 to a posterior density π according to an observation about simplicity that induces the likelihood function

$$(\theta, \gamma) \mapsto L(\theta, \gamma) = e^{-H(\theta,\gamma)},$$

defined up to multiplication by a positive constant. Since that observation is conditionally independent of the sample X, L qualifies as a prior likelihood function (Bickel, 2016).

That interpretation suggests dispensing with π_0, replacing π with methods of inference based on likelihood functions without prior distributions. Those procedures are sharpened to account for simplicity by replacing each *blunt likelihood function*, $(\theta, \gamma) \mapsto f_{\theta,\gamma}(x)$, with the *sharpened likelihood function*, $(\theta, \gamma) \mapsto L(\theta, \gamma) f_{\theta,\gamma}(x)$. Some special cases follow.

4.2 Sharpened likelihood asymptotics

Schweder and Hjort (2002) suggested accounting for pre-sample information by multiplying a likelihood function on which a confidence distribution is based by a prior likelihood function that encodes more subjective considerations. Likewise, first-order and higher-order asymptotic methods of deriving confidence sets and p values from the likelihood function (e.g., Severini, 2000; Brazzale et al., 2007; Butler, 2007) may take simplicity into account by using the sharpened likelihood function instead of the blunt likelihood function in quantities such as the score function, the Wald statistic, and Fisher information.

Example 2. For inference about θ as the normal mean in Example 1, a pseudo-likelihood function may be used to eliminate the nuisance parameter γ, the variance. For example, the profile likelihood function is

$$\theta \mapsto L_0(\theta; x) = \sup_{\gamma > 0} f_{\theta,\gamma}(x),$$

and the likelihood ratio test (LRT) of the hypothesis that $\theta = \theta_0$ yields a p value equal to the probability that

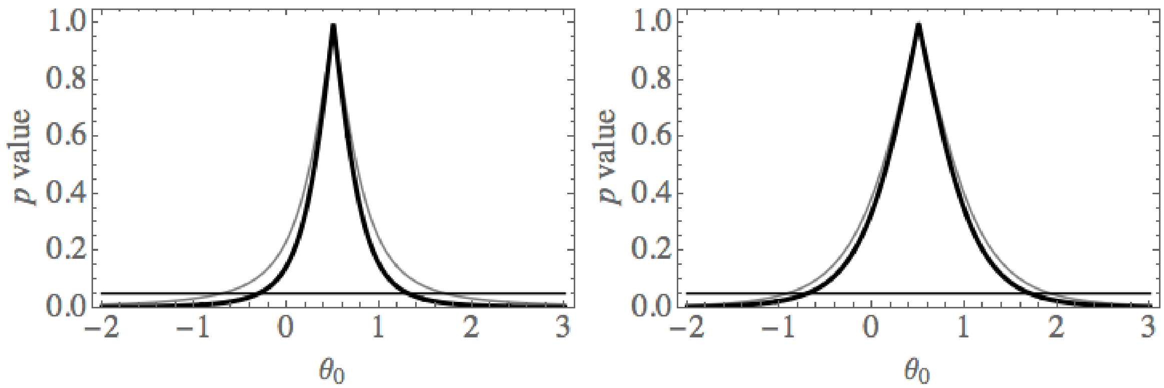

Figure 3: Profile-likelihood p values for testing whether the normal mean is θ_0 based on the blunt likelihood function (gray) and on the sharpened likelihood function (black). Intersections with the horizontal line indicate the 95% confidence intervals. The samples are $x = (0, 1)$ on the left and $x = (-1, 0, 1, 2)$ on the right.

a χ^2 variate with 1 degree of freedom is greater than the observed LRT statistic, $-2 \ln \left(L_0 (\theta_0; x) / L_0 \left(\widehat{\theta}; x \right) \right)$. The p value based on the sharpened likelihood replaces the profile likelihood function with

$$\theta \mapsto L(\theta; x) = \sup_{\gamma > 0} e^{-H(\theta, \gamma)} f_{\theta,\gamma}(x) \propto \sup_{\gamma > 0} \gamma^{-1/2} f_{\theta,\gamma}(x)$$

and replaces the LRT statistic with $-2 \ln \left(L(\theta_0; x) / L \left(\widehat{\theta}; x \right) \right)$. While sharpening the likelihood function has a notable effect on inference (Figure 3), it is less pronounced than that under the matching prior approach (Figure 2). ▲

4.3 Sharpened estimation, model selection, and model averaging

Certain methods of model selection and model averaging, such as the Bayesian information criterion (BIC) (Carlin and Louis, 2009, p. 53), the Akaike information criterion (AIC) (Burnham and Anderson, 2002), and the minimum description length (MDL) principle (Rissanen, 2007; Grünwald, 2007), are based on maximizing a product of the likelihood function and other factors, especially those reflecting the parametric complexity of each model, over the free parameters. (Methods assessing the evidence for a composite hypothesis may (Bickel, 2013) or may not (Zhang and Zhang, 2013) include other factors in the product.) Such methods of model selection, model averaging, and evidence measurement may incorporate information about the simplicity of distributions by multiplying the product by $e^{-H(\theta,\gamma)}$ before the maximization step.

Example 3. Many methods of model selection, including BIC, AIC, and MDL, reduce to maximum likelihood estimation (MLE) when each model consists of a single distribution. MLE, however, fails to incorporate the simplicity of each distribution. By contrast, *maximum sharpened likelihood estimation* results in the estimates

$$\arg \sup_{\theta\in\Theta,\gamma\in\Gamma} L(\theta,\gamma) f_{\theta,\gamma}(x) = \arg \sup_{\theta\in\Theta,\gamma\in\Gamma} e^{-H(\theta,\gamma)} f_{\theta,\gamma}(x).$$

Alternatively, if $\theta \mapsto L_0(\theta; x)$ is a pseudo-likelihood function such as a marginal, conditional, estimated, or integrated likelihood function that is free of γ, then *maximum sharpened pseudo-likelihood estimation* results in the estimate

$$\arg \sup_{\theta\in\Theta} L(\theta; x),$$

where $\theta \mapsto L(\theta; x)$ is the *sharpened pseudo-likelihood function*, the same transform of $(\theta,\gamma) \mapsto e^{-H(\theta,\gamma)} f_{\theta,\gamma}(x)$ that $L_0(\theta; x)$ is of $(\theta,\gamma) \mapsto f_{\theta,\gamma}(x)$. Example 2 illustrates a special case of $\theta \mapsto L(\theta; x)$. ▲

Acknowledgments

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada and by the Faculty of Medicine of the University of Ottawa.

References

Ando, T., 2010. Bayesian Model Selection and Statistical Modeling. Statistics: A Series of Textbooks and Monographs. Taylor & Francis.

Baker, A., 2016. Simplicity. In: Zalta, E. N. (Ed.), The Stanford Encyclopedia of Philosophy, winter 2016 Edition. Metaphysics Research Lab, Stanford University.
URL https://plato.stanford.edu/archives/win2016/entries/simplicity/

Bickel, D. R., 2013. Minimax-optimal strength of statistical evidence for a composite alternative hypothesis. International Statistical Review 81, 188–206.

Bickel, D. R., 2016. Computable priors sharpened into Occam's razors, Working Paper, University of Ottawa, hal-01423673.
URL https://hal.archives-ouvertes.fr/hal-01423673

Bickel, D. R., Padilla, M., 2014. A prior-free framework of coherent inference and its derivation of simple shrinkage estimators. Journal of Statistical Planning and Inference 145, 204–221.

Birnbaum, A., 1961. Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. Journal of the American Statistical Association 56, pp. 246–249.

Blaker, H., 2000. Confidence curves and improved exact confidence intervals for discrete distributions. Canadian Journal of Statistics 28 (4), 783–798.

Box, G. E. P., Tiao, G. C., 1992. Bayesian Inference in Statistical Analysis. Wiley-Interscience.

Brazzale, A. R., Davison, A. C., Reid, N., 2007. Applied Asymptotics: Case Studies in Small-sample Statistics. Cambridge University Press, Cambridge.

Burnham, K. P., Anderson, D. R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, New York.

Butler, R. W., 2007. Saddlepoint Approximations with Applications, 1st Edition. Cambridge University Press.

Carlin, B. P., Louis, T. A., 2009. Bayesian Methods for Data Analysis, Third Edition. Chapman & Hall/CRC, New York.

Claeskens, G., Hjort, N. L., 2008. Model Selection and Model Averaging. Cambridge University Press, Cambridge.

Datta, G. S., Sweeting, T. J., 2005. Probability matching priors. In: Dey, D., Rao, C. (Eds.), Bayesian Thinking. Vol. 25 of Handbook of Statistics. Elsevier, pp. 91 – 114.

De Bragança Pereira, C. A., Stern, J. M., 1999. Evidence and credibility: Full Bayesian significance test for precise hypotheses. Entropy 1 (4), 99–110.

Dowe, D. L., 2011. MML, Hybrid Bayesian Network Graphical Models, Statistical Consistency, Invariance and Uniqueness. In: Bandyopadhyay, P. S., Forster, M. R. (Eds.), Philosophy of Statistics. North Holland, Amsterdam, pp. 901–982.

Eaton, M. L., 1989. Group Invariance Applications in Statistics. Vol. 1. Institute of Mathematical Statistics, Hayward, California.

Ghosh, M., 2011. Objective priors: An introduction for frequentists. Statistical Science 26 (2), 187–202.

Grünwald, P. D., 2007. The Minimum Description Length Principle. MIT Press, London.

Helland, I. S., 2009. Steps Towards a Unified Basis for Scientific Models and Methods. World Scientific Publishing Company, Singapore.

Lindley, D. V., 2000. The philosophy of statistics. Journal of the Royal Statistical Society: Series D (The Statistician) 49 (3), 293–337.

Michalowicz, J. V., Nichols, J. M., Bucholtz, F., 2013. Handbook of Differential Entropy. CRC Press, New York.

Rissanen, J., 2007. Information and Complexity in Statistical Modeling. Springer, New York.

Schweder, T., Hjort, N. L., 2002. Confidence and likelihood. Scandinavian Journal of Statistics 29, 309–332.

Severini, T., 2000. Likelihood Methods in Statistics. Oxford University Press, Oxford.

Singh, K., Xie, M., Strawderman, W. E., 2005. Combining information from independent sources through confidence distributions. Annals of Statistics 33, 159–183.

Wallace, C. S., 2005. Statistical And Inductive Inference By Minimum Message Length. Springer, New York.

Wasserman, L., 2000. Bayesian model selection and model averaging. Journal of Mathematical Psychology 44 (1), 92–107.

Zhang, Z., Zhang, B., 2013. A likelihood paradigm for clinical trials (with discussion). Journal of Statistical Theory and Practice 7, 157–203.