Statistical Science
2016, Vol. 31, No. 4, 490–494
DOI: 10.1214/16-STS559
Main article DOI: 10.1214/16-STS592

Model Uncertainty First, Not Afterwards

Ingrid Glad and Nils Lid Hjort

Abstract. Watson and Holmes propose ways of investigating robustness of statistical decisions by examining certain neighbourhoods around a posterior distribution. This may partly amount to ad hoc modelling of extra uncertainty. Instead of creating neighbourhoods around the posterior a posteriori, we argue that it might be more fruitful to model a layer of extra uncertainty first, in the model building process, and then allow the data to determine how big the resulting neighbourhoods ought to be. We develop and briefly illustrate a general strategy along such lines.

Key words and phrases: Envelopes, Kullback–Leibler distance, local neighbourhoods, model robustness.

1. INTRODUCTION

The Bayesian apparatus has a clear master recipe. With data y, sampled from a model with parameters θ, and with a loss function $L(a,\theta)$ for potential actions or decisions a, one computes the posterior expected loss

$$\psi(a) = \mathrm{E}\{L(a,\theta) \mid \text{data}\} = \int L(a,\theta)\pi_I(\theta)\,\mathrm{d}\theta \tag{1}$$

and chooses the decision $\widehat{a}$ which minimises this function. Here, $\pi_I(\theta)$ is the posterior distribution for the model parameters, building also on a prior.

It is an entirely sensible idea to investigate robustness of both the $\psi(a)$ function and of the recipe's suggested decision $\widehat{a} = \operatorname{argmin}(\psi)$ with respect to the different ingredients, from the prior and the model specification to indeed also the loss function employed. Watson and Holmes (WH) carry out such investigations by examining $\widehat{a}$ inside Kullback–Leibler (KL) type neighbourhoods around π_I. They do so with these neighbourhoods put up after the original analysis, without particular regard to what has been put into the prior and the data model, to what might have been wrong there, and without a clear recipe for how big these neighbourhoods perhaps ought to be.

We suggest it would be more coherent and potentially fruitful to admit such a layer of extra uncertainty as part of the prior and model building process, and then examine the consequences for $\psi(a)$ and $\widehat{a}$. This allows the data their natural voice in the matter, creating the right amount of extra uncertainty around the first attempt at summarising information via the $\pi_I(\theta)$, rather than constructing ad hoc "neighbourhoods around the posterior a posteriori". In particular, the WH approach remains centred at π_I, not able to pick up a real bias of misspecification; our methods, laid out below, handle this, via KL neighbourhoods in the model specification, rather than by introducing extra uncertainty after the full analysis.

Ingrid Glad and Nils Lid Hjort are Professors of Statistics, Department of Mathematics, University of Oslo, P.B.1053, Blindern, N-0316 Oslo, Norway (e-mail: glad@math.uio.no; nils@math.uio.no).

2. A NEIGHBOURHOOD ELABORATION OF THE MODEL

Suppose the initial model for observations has the form of some $f(y,\theta)$, with a parameter vector of dimension say p; this is the setup that along with a prior $\pi_0(\theta)$ leads to the posterior distribution $\pi_I(\theta)$ in (1). We now embed the start model in a larger model $f(y,\theta,\gamma)$, with $\gamma = (\gamma_1,\ldots,\gamma_q)$ being extra parameters reflecting different ways in which the start model might have been too simplistic. These could relate to missing interaction terms in a regression model, Gaussian components not quite being Gaussian, a not fully correct link function, elements of dependence where the start model claims independence, etc. The narrow model corresponds to a null value $\gamma = \gamma_0$ in the γ parameter region, assumed below to be an inner parameter.

Now consider a focus parameter $\mu = \mu(\theta, \gamma)$, a "primary interest" parameter with direct relevance for the loss function; we could, for example, have $L(a, \theta) = L_0(|\mu - a|)$ with an appropriate L_0 depending only on how well we estimate μ. In a Bayesian setting, we are then interested in both:

(i) the posterior $\pi(\mu_{\rm narr} \mid \text{data})$, where $\mu_{\rm narr} = \mu(\theta, \gamma_0)$; and
(ii) $\pi(\mu_{\rm wide} \mid \text{data})$, where $\mu_{\rm wide} = \mu(\theta, \gamma)$ is the real thing.

We demonstrate below that both questions can be answered, in reasonable generality, in a local neighbourhood framework where $\gamma = \gamma_0 + \delta_0/\sqrt{n}$, in terms of the growing sample size n. The data generating mechanism is hence taken to be $f_{\rm true}(y) = f(y, \theta_0, \gamma_0 + \delta_0/\sqrt{n})$, for some (unknown) (θ_0, δ_0). The accompanying true value of the focus parameter is $\mu_{\rm true} = \mu(\theta_0, \gamma_0 + \delta_0/\sqrt{n})$. We take an interest in consequences for (i) and (ii), after having started with priors, say $\pi_0(\theta)$ for the θ part and $\pi_e(\delta)$ for the extra δ part. These questions and methods, leading to alternatives to the WH approach, may also be worked with in the frequentist framework of Schweder and Hjort (2016), where posterior distributions emerge without priors, but we here focus on the usual Bayesian approach. The local model framework also amounts to a KL neighbourhood setup; see (5) below. Our formalisation with $\mu = \mu(\theta, \gamma)$ and loss function built for that μ is a version of WH's Principles 1a and 1b.

Let $\widehat{\theta}_{\rm narr}$ be the maximum likelihood (ML) estimator of θ in the start model, having only θ on board, and let $(\widehat{\theta}, \widehat{\gamma})$ be the ML estimators in the $f(y, \theta, \gamma)$ model. These lead to ML estimators $\widehat{\mu}_{\rm narr} = \mu(\widehat{\theta}_{\rm narr}, \gamma_0)$ and $\widehat{\mu}_{\rm wide} = \mu(\widehat{\theta}, \widehat{\gamma})$ for the focus parameter, in the working model and the extended model, respectively. To explain what goes on, regarding the behaviour of both the ML estimators and with Bayes constructions, we need the Fisher information matrix $J_{\rm wide} = -\mathrm{E}\partial^2 \log f(Y, \theta_0, \gamma_0)/\partial\kappa\, \partial\kappa^{\rm t}$, writing $\kappa = (\theta, \gamma)$ for the full parameter vector of the extended model, but computed at the null model:

$$J_{\rm wide} = J(\theta_0, \gamma_0) = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix},$$

with inverse

$$J_{\rm wide}^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}.$$

The blocks indicated here of the $(p+q) \times (p+q)$ matrices have their appropriate sizes. Following Hjort and Claeskens (2003) and Claeskens and Hjort (2008), Chapters 5, 6, $D_n = \sqrt{n}(\widehat{\gamma} - \gamma_0) \to_d D \sim \mathrm{N}_q(\delta_0, Q)$, with $Q = J^{11}$, and

$$\begin{aligned} &\sqrt{n}(\widehat{\theta}_{\rm narr} - \theta_0) \to_d \mathrm{N}_p(J_{00}^{-1} J_{01}\delta_0, J_{00}^{-1}), \\ &\begin{pmatrix} \sqrt{n}(\widehat{\theta} - \theta_0) \\ \sqrt{n}(\widehat{\gamma} - \gamma_0) \end{pmatrix} \to_d \mathrm{N}_{p+q}\left(\begin{pmatrix} 0 \\ \delta_0 \end{pmatrix}, J_{\rm wide}^{-1} \right). \end{aligned} \tag{2}$$

Using the perhaps too simple start model amounts to smaller variability but a certain modelling bias, and vice versa with the extended model. This is captured in the following results, valid in the frequentist framework with a fixed $\delta_0/\sqrt{n}$ distance from the working model. Define $\omega = J_{10}J_{00}^{-1}\frac{\partial\mu}{\partial\theta} - \frac{\partial\mu}{\partial\gamma}$ and $\tau_0^2 = (\frac{\partial\mu}{\partial\theta})^{\rm t} J_{00}^{-1}\frac{\partial\mu}{\partial\theta}$, with partial derivatives evaluated at the null point. Then

$$\begin{aligned} &\sqrt{n}(\widehat{\mu}_{\rm narr} - \mu_{\rm true}) \to_d \mathrm{N}(\omega^{\rm t}\delta_0, \tau_0^2), \\ &\sqrt{n}(\widehat{\mu}_{\rm wide} - \mu_{\rm true}) \to_d \mathrm{N}(0, \tau_0^2 + \omega^{\rm t} Q \omega). \end{aligned} \tag{3}$$

Note that different focus parameters μ give rise to different ω, so some types of model misspecifications might cause little or no damage to some types of inferences or decisions, whereas other aspects missed by the working model might lead to misleading inference. The degree to which misspecification of the start model is crucial for the later inference hinges on the sizes of $|\omega^{\rm t}\delta_0|$ and $(\omega^{\rm t} Q \omega)^{1/2}$, depending in particular on the focus parameter, or, in yet other words, the loss function. We shall now see that results parallelling the frequentist findings (2)–(3) may be reached for Bayes solutions, of crucial relevance for questions (i) and (ii) above, depending however also on the precise prior $\pi_e(\delta)$ used for the $\gamma_0 + \delta/\sqrt{n}$ part.

First, consider $S_n = \sqrt{n}(\theta - \widehat{\theta}_{\rm narr})$ and its posterior distribution. Starting from

$$\begin{aligned} &\pi(s \mid \text{data}) \\ &\quad \propto \pi_0(\widehat{\theta}_{\rm narr} + s/\sqrt{n}) \\ &\qquad \cdot \exp\{\ell_n(\widehat{\theta}_{\rm narr} + s/\sqrt{n}, \gamma_0) - \ell_n(\widehat{\theta}_{\rm narr}, \gamma_0)\}, \end{aligned}$$

with $\ell_n(\theta, \gamma)$ the likelihood, one learns upon Taylor expansion and some further analysis that $\pi(s \mid \text{data}) \to_d \text{const.}\exp(-\frac{1}{2}s^{\rm t}J_{00}s)$, which is the $\mathrm{N}_p(0, J_{00}^{-1})$ density. With the delta method type of arguments, this leads to $\sqrt{n}(\mu_{\rm narr} - \widehat{\mu}_{\rm narr}) \mid \text{data} \to_d \mathrm{N}(0, \tau_0^2)$. In view of (3), this means first-order approximation agreement for frequentist and Bayesian analyses for μ via the narrow vehicle model. Confidence and credibility intervals are equal, to the first order,

they have sensible widths, but they are biased, thanks to $\omega^{\mathrm{t}}\delta_0$.

Second, consider $T_n = \sqrt{n}(\theta - \widehat{\theta})$ and the joint posterior for (T_n, δ). We find

$$\begin{aligned}&\pi(t,\delta \mid \text{data})\\ &\to_d \text{const.}\,\pi_e(\delta)\\ &\quad\cdot\exp\Big\{-\frac{1}{2}\begin{pmatrix} t\\ \delta - D\end{pmatrix}^{\mathrm{t}} J_{\text{wide}}\begin{pmatrix} t\\ \delta - D\end{pmatrix}\Big\}.\end{aligned}\tag{4}$$

This means that the part of the prior relating to θ is being washed out by the data, with $\theta \mid \text{data} \sim \mathrm{N}_p(\widehat{\theta}, J^{00}/n)$; this aspect of (4) corresponds to a Bernshteĭn–von Mises theorem for the θ part. The $\pi_e(\delta)$ part is not being washed out; however, in the limit, $\pi_e(\delta \mid \text{data}) \propto \pi_e(\delta)\exp\{-\frac{1}{2}(\delta - D)^{\mathrm{t}}Q^{-1}(\delta - D)\}$, where $D \mid \delta \sim \mathrm{N}_q(\delta, Q)$. It follows that

$$\begin{aligned}&\sqrt{n}(\mu_{\text{wide}} - \widehat{\mu}_{\text{wide}}) \mid \text{data}\\ &\quad\to_d \Big(\frac{\partial\mu}{\partial\theta}\Big)^{\mathrm{t}} T + \Big(\frac{\partial\mu}{\partial\gamma}\Big)^{\mathrm{t}}(\delta - D),\end{aligned}$$

with (T, δ) having the joint limiting distribution indicated in (4). If in particular a flat prior is used for $\pi_e(\delta)$, then this results in Bayesian inference matching frequentist inference to the first order, as is seen from (3).

Informative priors may be used, however, reflecting the assumption that the start model should not be very wrong. A natural prior on these extra parameters is $\delta \sim \mathrm{N}_q(0, \nu Q)$. Then the posterior is approximately a $\mathrm{N}_q(\rho D, \rho Q)$, from the above, with $\rho = \nu/(\nu + 1)$. We can infer the size of ν, and hence ρ and for later degrees of robustness, from data. We may specifically use the natural statistic $Z_n = n(\widehat{\gamma} - \gamma_0)^{\mathrm{t}}\widehat{Q}^{-1}(\widehat{\gamma} - \gamma_0)$, via the ML for γ and an estimate for Q inferred from that of J. It has the property that $Z_n \mid \delta \to_d D^{\mathrm{t}}Q^{-1}D \sim \chi^2_q(\delta^{\mathrm{t}}Q^{-1}\delta)$, and its unconditional limit mean is $q + q\nu$. This leads to the natural estimator $\widehat{\rho} = \text{clip}(1 - q/Z_n)$, where $\text{clip}(x)$ truncates x to the unit interval. The corresponding empirical Bayes scheme can then be followed by simulating from $\mu(\widehat{\theta} + t/\sqrt{n}, \gamma_0 + \delta/\sqrt{n})$, with (t, δ) drawn from the relevant (4) distribution.

WH construct KL neighbourhoods around the posterior from the start model. Our approach can be seen as constructing neighbourhoods around the model itself, via extra extension parameters γ, and then allowing the data to tell us how far these are from their null values. It turns out that these neighbourhoods also correspond to the KL metric. Writing for simplicity $f_0(y) = f(y, \theta_0, \gamma_0)$ and $f_\delta(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n})$, Taylor expansion and some analysis lead to both

$$\begin{aligned}\mathrm{KL}(f_0, f_\delta) &\doteq \frac{1}{2}(1/n)\delta^{\mathrm{t}}J_{11}\delta \quad\text{and}\\ \mathrm{KL}(f_\delta, f_0) &\doteq \frac{1}{2}(1/n)\delta^{\mathrm{t}}J_{11}\delta,\end{aligned}\tag{5}$$

implying in particular that the KL and the reverse KL neighbourhoods agree, to this order of approximation. Note that KL distances are "quadratic" and are easier to interpret on the square-root scale; densities $O(1/\sqrt{n})$ apart have KL distances $O(1/n)$.

3. AN ILLUSTRATION: ALMOST FLAT REGRESSION

Our methods and findings briefly exposited above generalise suitably to regression settings, partly following the methods of Claeskens and Hjort (2008), Chapters 6, 7. For an illustration, consider a simple regression setup where $y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i$ for $i = 1, \ldots, n$, for ε_i being i.i.d. $\mathrm{N}(0, \sigma^2)$, with $\bar{x} = n^{-1}\sum_{i=1}^n x_i$, and for simplicity of presentation take $\sigma = 1$ known. We take an interest in $\mu = \mathrm{E}(Y \mid x_0) = \beta_0 + \beta_1(x_0 - \bar{x})$. We take the narrow starting model to correspond to $\beta_1 = 0$ and the wider extension to have $\beta_1 = \delta/\sqrt{n}$, fitting with our general apparatus above. The ML estimators for β_0 and β_1 in the wider model are the familiar $\bar{y}$ and $(1/M_n)n^{-1}\sum_{i=1}^n(x_i - \bar{x})y_i$, where $M_n = n^{-1}\sum_{i=1}^n(x_i - \bar{x})^2$. In this case, the ML for β_0 in the narrow model is the same as in the wide model. We hence have $\widehat{\mu}_{\text{narr}} = \bar{y}$ and $\widehat{\mu}_{\text{wide}} = \bar{y} + (x_0 - \bar{x})\widehat{\beta}_1$. The Fisher information matrix is $J_n = \text{diag}(1, M_n)$, and $Q_n = 1/M_n$. We also need $D_n = \sqrt{n}\widehat{\beta}_1$, which has the $\mathrm{N}(\delta_0, Q_n)$ distribution. With a prior $\delta \sim \mathrm{N}(0, \nu Q_n)$, we have $\delta \mid D_n \sim \mathrm{N}(\rho D_n, \rho Q_n)$, with $\rho = \nu/(\nu + 1)$. The empirical Bayes estimate for this shrinkage parameter is $\text{clip}(1 - 1/Z_n)$, with $Z_n = nM_n\widehat{\beta}_1^2$.

Figure 1 relates to a simulated dataset with $n = 100$, with $(\beta_0, \beta_1) = (2.00, 3.50/\sqrt{n})$, and the x_i taking values $1/n, 2/n, \ldots, n/n$, and with interest in $\mu = \beta_0 + \beta_1(x_0 - \bar{x})$ at the next position $x_0 = 1 + 1/n$. The true value is 2.177, marked in the figure. The left-hand curve corresponds to WH's π_I, the posterior density for μ, computed based on the initial (and slightly wrong) model, missing the target due to the model bias. The right-hand curve corresponds to Bayesian analysis in the wider model, and also to a flat prior on δ in the $\beta_1 = \delta/\sqrt{n}$ setup. The middle curve is the empirical Bayes compromise, emerging from using the

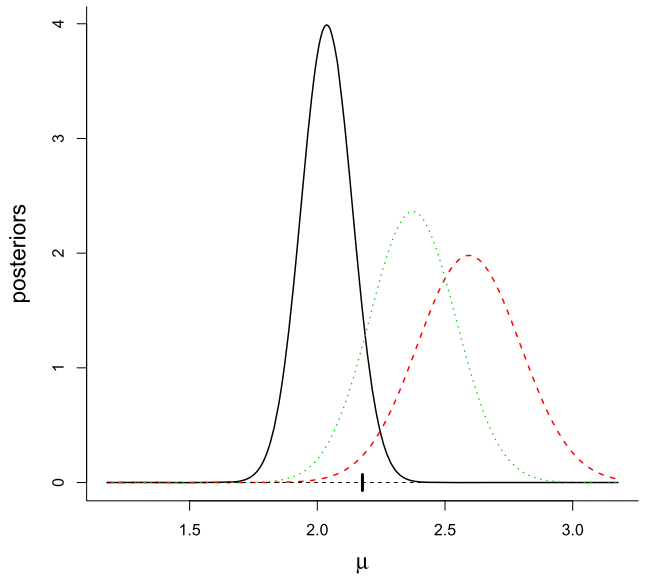

FIG. 1. *Posterior distributions for* $\mu = \mathrm{E}(Y \mid x_0)$ *in the simple regression illustration, with* $x_0 = 1.01$, *and true value* 2.177 *marked on the* μ *axis. Left curve*: $\pi_I(\mu)$, *from narrow model*; *right curve*: $\pi(\mu \mid \text{data})$ *in wide model*; *middle curve*: *the empirical Bayes compromise*.

$\delta \sim \mathrm{N}(0, \nu Q_n)$ prior and then estimating ν from data. It would be of interest to see in suitable detail how the WH approach would pan out in such a setting, given a relevant loss function, for example, of the type $L_0(|\mu - a|)$, via KL neighbourhood tilting of the $\pi_I(\mu)$ distribution.

4. KL NEIGHBOURHOODS WITH DIRICHLET PROCESSES DO NOT WORK

Given the authors' approach (though we found difficulties with this, conceptually and operationally, discussed above), it is at the outset also sensible to follow such ideas nonparametrically. The authors do so employing Dirichlet processes (Section 4.3), in effect attempting to examine posterior loss inside KL Dirichlet process neighbourhoods centred at π_I. It turns out that this is problematic, however. First, examining robustness within a random neighbourhood, say the set where $d(P, \pi_I) \le c$ (direct) or $d(\pi_I, P) \le c$ (reversed), clashes with WH's coherence principle, as they here seem to rely on a single realisation of a Dirichlet process $P \sim \mathrm{DP}(\alpha, \pi_I)$; even letting $m \to \infty$ in their favoured way of sampling from a DP, with an infinite bag of samples [cf. their equation (4)], corresponds to a single realisation; see the discussion in Hjort (2003), Section 2. It would perhaps make better sense to define such neighbourhoods via the means of these random distances. We note, incidentally, that WH's equation (7), giving a correct formula for the expected absolute deviation around the mean for a Beta distribution, seems to be taken as indication that the expected L_1 distance between the random Dirichlet process distribution function F and its mean F_0 ought to be of size $O(1/\alpha)$ (cf. WH's Figure 5). The real expected L_1 distance is however considerably bigger, and indeed of size $O(1/\sqrt{\alpha})$ as the concentration index α grows. This is seen from the Brownian motion limit of $\sqrt{\alpha}(F - F_0)$.

There are yet further technical issues with these KL neighbourhoods around π_I, as we shall now explain. For simplicity of presentation, take π_I to be the uniform distribution on the unit interval; the problems we point to with the Dirichlet process approach to KL neighbourhoods persist, and in the same manner, for other choices of the centre distribution π_I.

For the direct neighbourhood, let $P \sim \mathrm{DP}(\alpha, \pi_I)$, and consider the KL distance from P_m to π_I, where P_m is the inherited Dirichlet distribution on a fine partition of m intervals of length $1/m$. This is $\mathrm{KL}(P_m, \pi_I) = \sum_{i=1}^m p_i \log(p_i/q_i)$, with $q_i = 1/m$ and $(p_1, \ldots, p_m) \sim \mathrm{Dir}(\alpha/m, \ldots, \alpha/m)$. Writing $p_i = G_i/G$, with the $G_i \sim \mathrm{Gam}(\alpha/m, 1)$ independent and with sum $G \sim \mathrm{Gam}(\alpha, 1)$, one finds

$$\begin{aligned}\mathrm{KL}(P_m, \pi_I) &= \sum_{i=1}^m \frac{G_i}{G} \log \frac{G_i}{G} + \log m \\ &= -\frac{V_m}{G} - \log G + \log m,\end{aligned}$$

with $V_m = -\sum_{i=1}^m G_i \log G_i$. Here, V_m tends to a certain complicated distribution with mean $0.5772\,\alpha$ and variance $0.8237\,\alpha$; the main point is, however, that the real KL distance from the Dirichlet process to its centre approaches infinity. Consider also what WH term the reverse KL neighbourhood, involving $\mathrm{KL}(\pi_I, P_m) = \sum_{i=1}^m q_i \log(q_i/p_i)$. With the same representation as above, one finds

$$\begin{aligned}\mathrm{KL}(\pi_I, P_m) &= \sum_{i=1}^m q_i \log q_i - \sum_{i=1}^m q_i (\log G_i - \log G) \\ &= \log(G/m) + W_m,\end{aligned}$$

with $W_m = -\sum_{i=1}^m q_i \log G_i$. Via $\mathrm{E} \log G_i = \psi(\alpha/m) = \psi(1 + \alpha/m) - m/\alpha$, where $\psi = \Gamma'/\Gamma$ is the digamma function, and some further analysis one finds that $W_m \doteq m/\alpha$. Hence, $\mathrm{KL}(\pi_I, P_m) \doteq m/\alpha$ and tends to infinity in the limit from fine partition to a genuine Dirichlet process; $(1/m)\mathrm{KL}(\pi_I, P_m) \to 1/\alpha$.

So the direct and the reverse KL distances involved for this fine partition version P_m of $P \sim \mathrm{DP}(\alpha, \pi_I)$ are of size $\log m$ and m/α, both tending to infinity, indicating that KL neighbourhoods don't work in the intended fashion.

REFERENCES

CLAESKENS, G. and HJORT, N. L. (2008). *Model Selection and Model Averaging*. Cambridge Univ. Press, Cambridge. MR2431297

HJORT, N. L. (2003). Topics in non-parametric Bayesian statistics. In *Highly Structured Stochastic Systems* (P. J. Green, N. L. Hjort and S. Richardson, eds.). *Oxford Statist. Sci. Ser.* **27** 455–487. Oxford Univ. Press, Oxford. With discussion. MR2082419

HJORT, N. L. and CLAESKENS, G. (2003). Frequentist model average estimators (with discussion). *J. Amer. Statist. Assoc.* **98** 879–899. MR2041481

SCHWEDER, T. and HJORT, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge Univ. Press, Cambridge.