

# Chapter 2

## Sequential Combining of Expert Information Using Mathematica

Patrizia Agati, Luisa Stracqualursi, and Paola Monari

### 2.1 Introduction

Knowledge-gaining and decision-making in real-world domains often require reasoning under uncertainty. In such contexts, combining information from several, possibly heterogeneous, sources ('experts,' such as numerical models, information systems, witnesses, stakeholders, consultants) can really enhance the accuracy and precision of the 'final' estimate of the unknown quantity (a risk, a probability, a future random event, ...).

Bayesian paradigm offers a coherent perspective from which to address the problem. It just suggests to regard experts' opinions/outputs as data from an experiment [7]: a likelihood function may be associated with them to revise the prior knowledge. A Joint Calibration Model (JCM) makes the procedure more easier to assess [1, 6]. In such a way, the information combining process just becomes a knowledge updating process.

An issue strictly related to information combining is how to perform an efficient process of sequential consulting. Investigators, indeed, often prefer to consult the experts in successive stages rather than simultaneously: so, they avoid wasting time (and money) by interviewing a number of experts that exceed what they need. At each stage, the investigator can select the 'best' expert to be consulted and choose whether to stop or continue the consulting.

The aim of this work is to rephrase Bayesian combining algorithm in a sequential context and use *Mathematica* to implement suitable selecting and stopping rules. The paper is organized as follows. Section 2.2 gives the notation and suggests a

---

P. Agati • L. Stracqualursi (✉) • P. Monari  
Department of Statistical Sciences "P. Fortunati", University of Bologna, Via Belle Arti 41,  
40126 Bologna, Italy  
e-mail: [patrizia.agati@unibo.it](mailto:patrizia.agati@unibo.it); [luisa.stracqualursi@unibo.it](mailto:luisa.stracqualursi@unibo.it); [paola.monari@unibo.it](mailto:paola.monari@unibo.it)

recursive algorithm for information sequential combining, while Sect. 2.3 proposes selecting and stopping criteria for the consulting process. Mathematica 4.1 [8] was used to develop the code implemented in the notebook `EXPS.nb` and showed in Sect. 2.4. Finally, Sect. 2.5 presents a case-study.

## 2.2 A Recursive Algorithm for Information Combining

Let's suppose that an investigator  $A$  is uncertain about the value of a random quantity  $\theta$  and decides to consult a number of 'experts' with the aim of gaining knowledge about it. According to [7], experts' answers can be viewed as data from an experiment: a likelihood function may be associated with them and used to revise a prior judgment via Bayes' theorem. In such a way, the information combining process just becomes an information updating process.

This general principle can be applied to the aggregation of any kind of information, ranging from the combination of point estimates to the combination of probability distributions, and Bayes' rule can be implemented:

- In a 'standard' form, to be used when the experts' answers are combined with the prior all at once,
- In a recursive form, to be used when the experts are consulted sequentially and each new answer gets to update the posterior output obtained at the previous stage.

In the following, we rephrase Morris' approach in a sequential context and show both the simultaneous form and the recursive form of Bayes' rule for combining information from different sources.

Let's suppose that  $A$ 's body of knowledge about  $\theta$  is represented as a (possibly uninformative) probability density function (in the following, pdf)  $h_0(\cdot)$  on the space of states  $\Theta \subseteq \mathfrak{R}$ . Due to efficiency reasons, he chooses to consult the experts sequentially: at each stage  $k$  ( $k = 1, 2, \dots, K$ ), he picks an expert  $Q_j$  ( $j = 1, 2, \dots, n$ ) from a pool of size  $n$  ( $n \geq K$ ). The selected expert  $Q_{j;k}^*$  (or, more briefly,  $Q_k$ ) provides subject  $A$  with a pdf  $g_k(\cdot)$  on the space of states. Using Bayes' theorem, the posterior pdf of the investigator  $A$  at stage  $k$  can be written as

$$h_k(\theta | g_1, \dots, g_i, \dots, g_k) \propto \ell(g_1, \dots, g_i, \dots, g_k | \theta) \cdot h_0(\theta) \quad (2.1)$$

where  $\ell(\cdot)$  denotes the likelihood function of  $\theta$  for the experimental data  $\{g_1, \dots, g_i, \dots, g_k\}$  and the constant of proportionality is  $[\int_{\Theta} \ell(g_1, \dots, g_i, \dots, g_k | \theta) \cdot h_0(\theta) d\theta]^{-1}$ .

What makes the Bayesian approach rather difficult to apply is the assessment of the likelihood function. If the experts' answers are processed simultaneously, the function to be assessed is a joint probability distribution over the whole set of functions  $g_i(\cdot)$ , for  $i = 1, 2, \dots, k$ ; on the other hand, if the expert information are processed recursively, the likelihood is assessed at each stage as a conditional

probability distribution given the subset of functions  $g_i(\cdot)$  already acquired; in both the cases, the assessment requires to account for the different performances as well as the dependences between experts.

Some assumptions allow to express the likelihood function in a form easier to be modeled [7]:

- i) Each  $g_i(\cdot)$  is parameterized with a location parameter  $m_i$  and a shape parameter  $v_i$ . For example,  $g_i(\cdot)$  denotes the pdf of a Gaussian random variable  $N(m_i, v_i)$ . Then, Eq. (2.1) becomes

$$h_k(\theta | m^{(k)}, v^{(k)}) \propto \ell(m^{(k)} | v^{(k)}, \theta) \cdot \ell(v^{(k)} | \theta) \cdot h_0(\theta) \quad (2.2)$$

where:

- $m^{(k)}$  represents the event “the location parameters supplied by the first  $k$  experts will be  $m_1, \dots, m_i, \dots, m_k$ ”;
  - Analogously,  $v^{(k)}$  indicates the event “the scale parameters supplied by the first  $k$  experts will be  $v_1, \dots, v_i, \dots, v_k$ ”;
  - Both  $\ell(v^{(k)} | \theta)$  and  $\ell(m^{(k)} | v^{(k)}, \theta)$  are likelihood functions, to be viewed as functions of  $\theta$ . The former is the likelihood function of  $\theta$  for the data  $v^{(k)}$ : it is defined by the probabilities assigned by subject  $A$  to the event  $v^{(k)}$  for  $\theta$  varying. The latter, denoted in the following by  $\ell_k(\theta)$  for notational convenience, is the conditioned likelihood function of  $\theta$  for the data  $m^{(k)}$ , given the event  $v^{(k)}$ : it represents the joint probability—conditioned upon the event  $v^{(k)}$ —assigned by subject  $A$  to the event  $m^{(k)}$ , for  $\theta$  varying;
- ii) The probabilities that subject  $A$  assigns to the event  $v^{(k)}$  do not depend on  $\theta$ : in symbols,  $\ell(v^{(k)} | \theta) = \ell(v^{(k)}) = c$ , where  $c$  denotes a constant of proportionality.<sup>1</sup> Using this assumption, Eq. (2.2) takes the form:

$$h_k(\theta | m^{(k)}, v^{(k)}) \propto \ell(m^{(k)} | v^{(k)}, \theta) \cdot h_0(\theta) \quad (2.3)$$

where the constant of proportionality is  $[\int_{\Theta} \ell(m^{(k)} | v^{(k)}, \theta) \cdot h_0(\theta) d\theta]^{-1}$ ;

- iii) The conditional probabilities that subject  $A$  assigns to the event “the shape parameter given by the  $k$ th expert will be  $v_k$ ,” given  $m^{(k-1)}$  and  $v^{(k-1)}$ , do not depend on  $\theta$ : in symbols,  $\ell(v_k | m^{(k-1)}, v^{(k-1)}, \theta) = \ell(v_k | m^{(k-1)}, v^{(k-1)})$ . If such an assumption holds, then Eq. (2.3) can be written in a recursive form as

$$h_k(\theta | m^{(k)}, v^{(k)}) \propto \ell(m_k | m^{(k-1)}, v^{(k)}, \theta) \cdot h_{k-1}(\theta) \quad (2.4)$$

<sup>1</sup>Due to the reciprocity of the stochastic independence assumption ii) can be also expressed as *invariance to scale* about  $\theta$ , that is  $h(\theta | v^{(k)}) = h(\theta)$ : the event  $v^{(k)}$  alone gives no information regarding  $\theta$ .

where:

- $\ell(m_k | m^{(k-1)}, v^{(k)}, \theta)$  is the conditioned likelihood function of  $\theta$  for the only observation  $m_k$ , given the scale parameters of the first  $k$  experts ( $v_k$  included) and the location parameters of the first  $k - 1$  experts;
- The constant of proportionality is  $[\int_{\Theta} \ell(m_k | m^{(k-1)}, v^{(k)}, \theta) \cdot h_{k-1}(\theta) d\theta]^{-1}$ .

For the purpose of assessing the likelihood  $\ell_k(\theta) = \ell(m^{(k)} | v^{(k)}, \theta)$  in Eq. (2.3), Morris introduced the notions of *performance indicator* and *performance function*.

The performance indicator  $\tau_i$  associated with  $g_i(\cdot)$  is defined as the cumulative distribution function  $G_i(\cdot | m_i, v_i)$  evaluated at the true value  $\theta_0$  of  $\theta$ :

$$\tau_i = \tau_i(m_i, v_i, \theta_0) = \int_{-\infty}^{\theta_0} g_i(\theta | m_i, v_i) d\theta = G_i(\theta_0 | m_i, v_i) \quad (2.5)$$

where  $0 \leq \tau_i \leq 1$ . For example, if the observed value is the 0.3-quantile of  $g_i(\cdot)$ , then  $\tau_i = 0.3$ .

The performance function, denoted by  $\varphi(\tau^{(k)} | v^{(k)}, \theta)$ , is defined as a conditional joint density on the event  $\tau^{(k)}$  “the performance indicators for the first  $k$  experts will be  $\tau_1, \dots, \tau_i, \dots, \tau_k$ ,” given  $v^{(k)}$  and  $\theta$ .

Given  $v^{(k)}$ , for any fixed value of  $\theta$ , a monotonic decreasing relationship exists between corresponding elements  $\tau_i$  and  $m_i$ . So, a change of variable allows to show that:

$$\ell(m^{(k)} | v^{(k)}, \theta) = C_k(\theta) \cdot \prod_{i=1}^k g_i(\theta | m_i, v_i) \quad (2.6)$$

where:

$$C_k(\theta) = \varphi[G(\theta)^{(k)} | v^{(k)}, \theta] = \varphi(\tau^{(k)} | v^{(k)}, \theta) \quad (2.7)$$

Equation (2.6) shows that the likelihood function can be obtained as the product of the pdfs from the experts, adjusted by a *joint calibration function*  $C_k(\cdot)$  that models the performance of the experts and their mutual dependence in assessing  $\theta$ :  $C_k(\cdot)$  is nothing but the performance function  $\varphi(\tau^{(k)} | v^{(k)}, \theta)$  viewed as a function of  $\theta$  (for fixed  $m^{(k)}$ ). It expresses the admissibility degrees assigned to each possible  $\theta$ -value regarded as the realization of the event  $\tau^{(k)}$ .

By substituting (2.6) into (2.3), the posterior pdf can be written as:

$$h_k(\theta | m^{(k)}, v^{(k)}) \propto C_k(\theta) \cdot \prod_{i=1}^k g_i(\theta | m_i, v_i) \cdot h_0(\theta) \quad (2.8)$$

which describes the structural form of what we call “Joint Calibration Model”(JCM).

It is worth noting that Eq. (2.6) can be also used to assess the likelihood function in Eq. (2.4), since the relation  $\ell(m_k | m^{(k-1)} v^{(k)}, \theta) = \ell_k(\theta) / \ell_{k-1}(\theta)$  holds.

Implementing JCM requires that function  $C_k(\theta)$  is properly specified. In other words, once the scale parameters are known to subject  $A$ , a conditional pdf  $\varphi(\tau^{(k)} | v^{(k)}, \theta)$  on the  $k$ -dimensional performance indicator variate  $\tau^{(k)}$  should be specified.

This task is less demanding if function  $\varphi(\tau^{(k)} | v^{(k)}, \theta)$  can be assumed to take the same value whatever the true value of  $\theta$  be (*equivariance to shift* assumption):

$$\varphi(\tau^{(k)} | v^{(k)}, \theta) = \varphi(\tau^{(k)} | v^{(k)}) \quad (2.9)$$

However, it still remains a frustratingly difficult task, especially in the absence of an adequate parametric modelling, which would allow to assess the entire function by means of a relatively small number of parameters.

There exist several suitable choices about a parametric probabilistic model for the  $k$ -dimensional performance variate  $\tau^{(k)}$ . Some preliminary remarks are necessary in order to motivate our choice:

- According to definition (2.5), each element  $\tau_i$  is a (cumulate) probability;
- When modelling a joint pdf  $\varphi(\cdot | v^{(k)})$  on the variate  $\tau^{(k)}$ , it needs to take into account that “values [...] near 0 or 1 will ordinarily have smaller standard errors than those around 0.5. [...] A possibility is to suppose some transform of probability, like log-odds, has constant variance” [4];
- Log-odds lie in the range  $-\infty$  to  $+\infty$ : probabilities that are less, equal or greater than 0.5 correspond to negative, zero, or positive log-odds, respectively. Therefore, modelling the performance function in terms of log-odds, instead of probabilities, is advantageous also because the range of log-odds is coherent with a Gaussian distribution, which is attractive for its good analytic properties and the clear interpretation of its parameters.

For these reasons, a reasonable choice is to assume:

$$\tilde{\tau}^{(k)} \sim N_k(\tilde{t}^{(k)}, S) \quad (2.10)$$

where

- $\tilde{\tau}^{(k)}$  refers to the  $k$ -dimensional vector of log-odds  $[\tilde{\tau}_i]_{i=1, \dots, k}'$ , with  $\tilde{\tau}_i = \ln[\tau_i / (1 - \tau_i)] \in \Re$  for  $i = 1, 2, \dots, k$ ;
- $\tilde{t}^{(k)}$  and  $S$  denote the mean vector and the covariance matrix of the  $k$ -variate Gaussian distribution, respectively.

The analytical form of function  $\varphi(\tau^{(k)} | v^{(k)})$  can be obtained by using a change of variable from  $\tilde{\tau}^{(k)}$  to  $\tau^{(k)}$ . Denoting by  $\psi(\cdot | v^{(k)})$  the model in (2.10), the well-known change formula yields:

$$\varphi(\tau^{(k)} | v^{(k)}) = |J_{\tilde{\tau}^{(k)} \rightarrow \tau^{(k)}}| \cdot \psi_{\tau^{(k)}}(\tau^{(k)} | v^{(k)}) \quad (2.11)$$

As the Jacobian of the transformation  $\tilde{\tau}^{(k)} \rightarrow \tau^{(k)}$  is:

$$J_{\tilde{\tau}^{(k)} \rightarrow \tau^{(k)}} = \prod_{i=1}^k \frac{1}{\tau_i (1 - \tau_i)} \quad (2.12)$$

the resulting performance function of the variate  $\tau^{(k)}$  is:

$$\varphi\left(\tau^{(k)} \middle| v^{(k)}\right) = c \cdot \prod_{i=1}^k \frac{1}{\tau_i (1 - \tau_i)} \cdot \exp\left[-\frac{1}{2} \left(\tilde{\tau}^{(k)} - \tilde{t}^{(k)}\right)' \mathbf{S}^{-1} \left(\tilde{\tau}^{(k)} - \tilde{t}^{(k)}\right)\right] \quad (2.13)$$

where  $c$  denotes the normalization constant.

Finally, the calibration function  $C_k(\theta)$ , defined in (2.7), can be obtained as follows. Definition (2.5) implies that:

$$\tilde{\tau}^{(k)} = [\tilde{G}(\theta)]^{(k)} \quad (2.14)$$

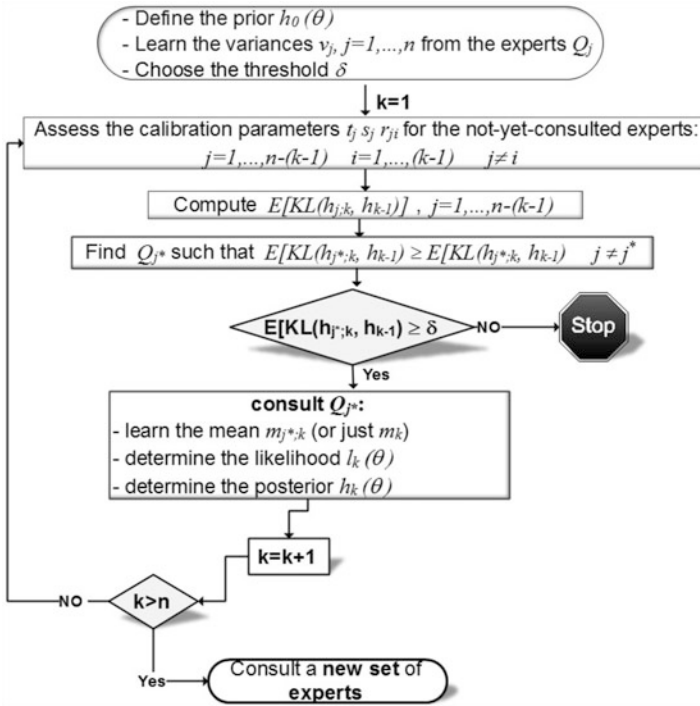
By substituting Eq. (2.14) in (2.13),  $C_k(\theta)$  takes the form:

$$\begin{aligned} C_k(\theta) &= \varphi\left([G(\theta)]^{(k)} \middle| v^{(k)}\right) \\ &= c \cdot \prod_{i=1}^k \frac{1}{G_i(\theta) \cdot [1 - G_i(\theta)]} \cdot \exp\left\{-\frac{1}{2} \left\{[\tilde{G}(\theta)]^{(k)} - \tilde{t}^{(k)}\right\}' \right. \\ &\quad \left. \mathbf{S}^{-1} \left\{[\tilde{G}(\theta)]^{(k)} - \tilde{t}^{(k)}\right\}\right\} \end{aligned} \quad (2.15)$$

It's worth noting that the calibration function, as expressed by (2.15), is univocally defined by two parameters only: the mean vector and the covariance matrix of the variate  $\tilde{\tau}^{(k)}$ .

## 2.3 Selecting and Stopping Rules

In designing and performing the sequential process, the purpose of expert consulting is reducing the uncertainty about the unknown quantity  $\theta$ . So, it is reasonable to found the selecting and stopping rules on some criterion of informativeness. More precisely, though no single number can convey the amount of information carried by a density function, a synthetic measure of the (*expected*) additional informative value of a not-yet-consulted expert  $Q_{j;k}$  is indispensable for selecting the one to be consulted at stage  $k$ , especially when the assessment of the calibration parameters, together with the shape parameters provided by the experts, leads to not-coinciding preference orderings. And, analogously, as likelihood functions and

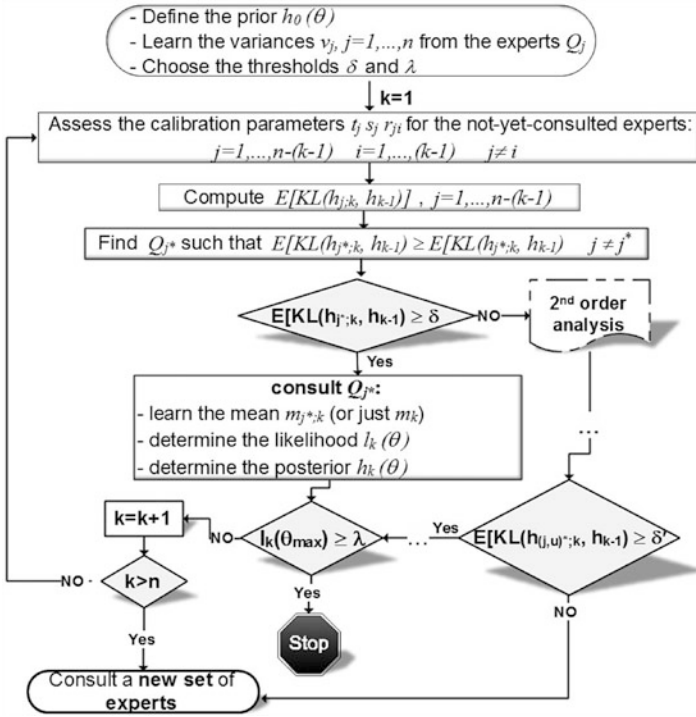


**Fig. 2.1** Flow-chart of the sequential procedure with stopping rule based on the expected KL-divergences between contiguous stages

posterior densities can display a wide variety of form, a synthetic measure of the knowledge level achieved about  $\theta$  is needed for picking out the ‘optimal’ stage  $k^*$  at which data acquiring can be stopped.

Let’s suppose the investigator  $A$  is performing the process of revising beliefs in light of new data according to the algorithm described in Sect. 2.2. The prior  $h_0(\theta)$  has already been specified; each expert  $Q_j$  in the pool of size  $n$  has revealed the variance  $v_j$ —assumed as uninformative about  $\theta$ : see assumption *ii*) in Sect. 2.2—of his own density  $g_j(\theta)$ , and  $A$  has already consulted  $k-1$  of them, so obtaining the locations of  $k-1$  expert densities:  $A$  is now at stage  $k$  of the process (Figs. 2.1 and 2.2), and must select one among the experts  $Q_{j;k}$  not yet consulted ( $j = 1, 2, \dots, n-k+1$ ).

For each  $Q_{j;k}$ , the investigator  $A$  assesses—conditionally on  $v_j$ , on the basis of the information at his disposal (including all the expert locations  $m_i$  revealed up to stage  $k-1$ )—the parameters of the  $k$ -stage calibration function  $C_{j;k}(\theta)$ : that is,  $t_j$ ,  $s_{jj}$  and the covariances  $s_{ji}$  (or the linear correlations  $r_{ji}$ ) between  $Q_{j;k}$  and each expert  $Q_i$  already consulted,  $i = 1, 2, \dots, k-1$ . At this point of the procedure, no  $Q_{j;k}$  has revealed the location value  $m_j$  of his own  $g_j(\theta)$ : the several ‘answers’  $m_j$  which each expert can virtually give are not all equally informative,



**Fig. 2.2** Flow-chart of the sequential procedure with stopping rule based on the observed curvature  $I_k(\theta)$  of the log-likelihood valued at  $\theta := \theta_{\max}$

so the (informative) value of each expert at stage  $k$ —to be measured with regard to  $A$ 's current knowledge<sup>2</sup> of  $\theta$  expressed by the posterior density  $h_{k-1}(\theta)$  of the previous stage—is an *expected* value, computed by averaging a suitable measure of relevant information about  $\theta$  in  $Q_{j;k}$ 's answer over the space  $M_j$  of the virtually possible  $m_j$  values.

By reasoning in a mere *knowledge* context, which is an *inductive* context, where an expert opinion is more relevant the more it is able to modify the posterior distribution on the unknown quantity, a suitable measure of  $Q_{j;k}$ 's informative value can be the *expected Kullback–Leibler divergence* of the density  $h_{j;k}(\theta)$  with respect to the posterior  $h_{k-1}(\theta)$  obtained at the previous stage,

$$E[KL(h_{j;k}, h_{k-1})] := \int_{M_j} f(m_{j;k} | v_{j;k}, m^{(k-1)}, v^{(k-1)}) \cdot KL(h_{j;k}, h_{k-1}) dm_j \quad (2.16)$$

<sup>2</sup>In fact, all the other elements being equal, the more  $A$  is uncertain about  $\theta$ , the more an answer  $m_j$  is worthy.



where the KL-divergence [3],

$$KL(h_{j;k}, h_{k-1}) := \int_{\Theta} h_{j;k}(\theta) \cdot \ln[h_{j;k}(\theta) / h_{k-1}(\theta)] d\theta \quad (2.17)$$

measures indirectly the information provided by an answer  $m_{j;k}$  in terms of the changes it yields on the density  $h_{k-1}(\theta)$ . The conditional density  $f(\cdot)$  in (2.16) is equal to the reciprocal of the constant of proportionality of Eq. (2.4), regarded as a function of  $m_{j;k}$  and normalized; when assumptions i), ii), and iii) hold, it can be determined as

$$f(m_{j;k} | v_{j;k}, m^{(k-1)}, v^{(k-1)}) = f(m^{(j;k)} | v^{(j;k)}) / f(m^{(k-1)} | v^{(k-1)}) \quad (2.18)$$

where the density  $f(m^{(j;k)} | v^{(j;k)})$ , and analogously  $f(m^{(k-1)} | v^{(k-1)})$ , is equal, up to the normalization term, to the reciprocal  $\int_{\Theta} \ell(m^{(k)} | v^{(k)}, \theta) \cdot h_0(\theta) d\theta$  of the constant of proportionality of Eq. (2.2), regarded as a function of  $m^{(k)}$ .

The expert  $Q_{j;k}^*$  characterized by the greatest expected KL-divergence is, at stage  $k$ , the most informative: but is he an expert worth consulting? The answer is yes, if the information he provides is, on average, *enough* different from what  $A$  already knows about  $\theta$ , i.e. if the expected divergence of  $h_{j^*;k}(\theta)$  with respect to  $h_{k-1}(\theta)$  is not less than a preset threshold  $\delta$  ( $0 \leq \delta < \infty$ ). About the choice of the threshold  $\delta$ , a very useful tool is the scheme proposed by McCulloch [5], who suggested to connect any value  $\delta$  of a KL-divergence to the KL-divergence of a Bernoulli distribution with  $p = 0.5$  from a Bernoulli with  $p = b(\delta) = \frac{1 + \sqrt{1 - e^{-2\delta}}}{2}$ . Table 2.1 shows a range of correspondences.

So the *selecting rule* can be expressed as follows. *Consult the expert  $Q_{j;k}^*$  such that*

$$E[KL(h_{j^*;k}, h_{k-1})] \geq E[KL(h_{j;k}, h_{k-1})] \quad j \neq j^* \quad (2.19)$$

*on condition that*

$$E[KL(h_{j^*;k}, h_{k-1})] \geq \delta \quad (2.20)$$

*If  $Q_{j;k}^*$  does not satisfy (2.20), then proceed to a second order analysis: that is, consult the pair  $(Q_{j;k}, Q_{u;k})^*$  presenting the greatest expected KL-divergence, provided that it is  $E[KL(h_{(j,u)^*;k}, h_{k-1})] \geq \delta$ .*

**Table 2.1** Large or small KL-divergences? Relation between  $\delta$  and  $b(\delta)$  values

$\delta$	0	0.0001	0.001	0.005	0.010	0.020	0.050	0.090	0.10	0.14
$b(\delta)$	0.50	0.51	0.52	0.55	0.57	0.60	0.65	0.70	0.71	0.75
$\delta$	0.22	0.34	0.51	0.83	1	2	> 3			
$b(\delta)$	0.80	0.85	0.90	0.95	0.96	0.99	$\approx 1.00$			

Otherwise contact a new set of experts and perform a new process by using the posterior  $h_{k-1}(\theta)$  as a new prior  $h'_0(\theta)$ .

The expert  $Q_{j;k}^*$  satisfying both the Eqs. (2.19) and (2.20) becomes just  $Q_k$ , the “ $k$ -stage expert.” By consulting him,  $A$  learns the location  $m_k$  of the density  $g_k(\cdot)$ : now, the  $k$ -stage calibration function  $C_k(\theta)$  is univocally defined, and consequently, the likelihood function  $\ell_k(\theta)$  and the posterior density  $h_k(\theta)$  too.

It is intuitive, as well as reasonable, that the investigator stops the process only when the knowledge about  $\theta$ , expressed by the posterior density, is ‘inertially stable’: i.e., only when additional experts, even if jointly considered, are not able to modify the synthesis distribution appreciably, on the contrary they contribute to its inertiality. So, the *stopping rule* can be defined as follows. *Stop the consulting process at stage  $k^*$  at which none of the remaining experts satisfies condition (2.20).*

If too many experts are needed for realizing such a stopping condition, it can be weakened by just requiring the knowledge about  $\theta$  deriving from expert answers to be enough for  $A$ ’s purposes. A measure of the experimental data strength in determining a preference ordering among ‘infinitesimally close’ values of  $\theta$  is Fisher’s notion of information. The value of the *observed information*  $I(\cdot)$  at the maximum of the log-likelihood function

$$I_k(\theta_{\max}) := -\partial^2/\partial\theta^2 \ln \ell_k(\theta_{\max}) \quad (2.21)$$

is a second-order estimate of the spherical curvature of the function at its maximum: within a second-order approximation, it corresponds to the KL-divergence between two distributions that belong to the same parametric family and differ infinitesimally over the parameter space.

So, an alternative *stopping rule* can be defined as follows. *Stop the consulting at stage  $k^*$  at which a preset observed curvature  $\lambda$  of the log-likelihood valued at  $\theta := \theta_{\max}$  has been achieved,*

$$I_{k^*}(\theta_{\max}) \geq \lambda \quad (2.22)$$

In order to decide whether a curvature value  $I(\theta_{\max}) = \lambda$  is a large or a small one, a device could be the following. Let’s think of a binomial experiment where a number  $x = n/2$  of successes is observed in  $n$  trials and find  $x$  such that  $I(\hat{p}_{\text{ML}} = 0.5) = \lambda$ , where  $\hat{p}_{\text{ML}} = 0.5$  is the maximum likelihood estimate of the binomial parameter  $p$ . Table 2.2 shows a range of  $x$  values with the corresponding curvature values. The simple relation  $x = \lambda/8$  holds: so, for example, if  $\lambda = 120$ , the width of the curve  $\ln \ell_k(\theta)$  near  $\theta := \theta_{\max}$  is the same as the curve  $\ln \ell(p)$  at  $\hat{p}_{\text{ML}} = 0.5$  when  $x = 15$  and  $n = 30$ .

**Table 2.2** Large or small curvature values? Relation between  $x$  and  $\lambda$  values

$x$	1	2	5	10	15	20	25	30	40	50
$\lambda$	8	16	40	80	120	160	200	240	320	400

## 2.4 Selecting and Stopping Rules Implemented: A Mathematica Code

The procedure implemented in the notebook file `EXPS.nb` allows to select the ‘best’ expert to consult at each stage of a sequential process and decide the ‘best’ stage at which the process can be stopped. The program computes the expected KL-divergencies between posterior densities at two contiguous stages, for any number of experts: it uses condition (2.20) as stop criterion.

The choice of Mathematica package is due to complexity and accuracy necessary in this recursive procedure [2]. The code begins by importing package Statistics ‘NormalDistribution’ and opening declarations.

```
<<Statistics`NormalDistribution`;
PDF[NormalDistribution[mu_,sigma_],x_]:=1/(sigma*Sqrt[2 Pi])
                                         Exp[-((x-mu)/sigma)^2/2];
CDF[NormalDistribution[mu_,sigma_],x_]:= (Erf[(x-mu)/(Sqrt[2] sigma)]
                                         +1)/2;
```

The program needs the following input quantities:

- ‘Initialdata’: it is the matrix which contains the mean  $m_j$ , the variance  $v_j$ , the calibration parameters  $t_j$  and  $s_j^2$  of each expert  $Q_j$  ( $j = 1, \dots, n$ );
- ‘R’: it is the correlation matrix. Its elements are the calibration parameters  $r_{ji}$ , denoting the correlation degree between the performances of the experts  $Q_i$  and  $Q_j$ ;
- ‘delta’: it is the threshold for the Kullback–Leibler divergence;
- ‘v0, m0’: they are the variance and the mean of the prior  $h_0(\theta)$ ;
- ‘mmin, mmax’: they are the lower and the upper limits for the unknown  $\theta$ ;
- ‘stagemax’: it is the size  $n$  of the pool of experts.

The code restores the input matrices: the process starts out.

```
numexpert=Dimensions[Initialdata][[2]];
expert=Table[k,{k,numexpert}];
Matdata=Transpose[Join[{exp,m,v,t,s2},
                        Transpose[Join[{expert},Initialdata]]]];
{rin,sin}=Dimensions[Matdata];
M=Initialdata[[1]];V=Initialdata[[2]];
T=Initialdata[[3]];S=Initialdata[[4]];
Cova=Table[Sqrt[S[[i]]*S[[j]]]*R[[i,j]],{i,numexpert},{j,numexpert}];

nextmax=Table[0,{i,stagemax}];
next[n_]:=Table[nextmax[[j]],{j,n-1}];
nextadd[n_,k_]:=Join[next[n],{k}];
mattV[n_]:=Table[V[[nextadd[n,k]]],{k,numexpert}];
mattT[n_]:=Table[T[[nextincr[n,k]]],{k,numexpert}];
mattM[n_]:=Table[M[[nextadd[n,k]]],{k,numexpert}];
mattS[n_]:=Table[S[[nextadd[n,k]]],{k,numexpert}];
mattCova[n_]:=Table[Cova[[nextadd[n,k],nextadd[n,k]]],{k,numexpert}];
Covamin[n_]:=Table[Cova[[next[n],next[n]]]];
nextH=Table[0,{i,stagemax}];
```

```

g[x_, i_] := PDF[NormalDistribution[0, Sqrt[mattV[n][[k]][[i]]]], x]
G[x_, i_] := CDF[NormalDistribution[0, Sqrt[mattV[n][[k]][[i]]]], x]
g0[x_] := PDF[NormalDistribution[m0, Sqrt[v0]], x]

```

```

MatrixForm[Matdata]
MatrixForm[R]

```

The code calculates the expected KL-divergences at stage 1 and put them in a vector.

```

Expect1:=(n=1; Print["Stage K =", " ", n];
  Kull=Table[0, {k, numexpert}];
  For[k=1, k<=numexpert, k++, Kull[[k]] =
    NIntegrate[(F1[m] - H1[m] * Log[H1[m]]) / K1[m], {m, -5, 8}] /
    NIntegrate[H1[m] / K1[m], {m, -5, 8}];
  Print[Kull[[k]]];
];
MatKull=Join[Matdata, {Join[{K}, Kull]}];
max=Position[Kull, Max[Kull]][[1, 1]]; k=max; H11=H1[M[[max]]];
nextmax[[1]]=max; nextH[[1]]=H11;
Return[MatrixForm[MatKull]]
)
likelihood1[x_, m_] := Block[{G1x, g0x, B1x, cx, gauX, factx},
  G1x=CDF[NormalDistribution[0, Sqrt[V[[k]]]], x];
  g0x=PDF[NormalDistribution[m0, Sqrt[v0]], x];
  B1x=(G1x+10^-20)/(1-G1x+10^-20)*(1-T[[k]])/T[[k]];
  cx=Log[B1x];
  gauX=Exp[-1/2*(1/S[[k]]*cx^2+1/V[[k]]*x^2)];
  factx=1/((G1x+10^-20)(1-G1x+10^-20));
  likeK1=gauX*factx;
  likeH1=likeK1*g0x*Exp[-1/2*m^2+1/v0-((x-m0)/v0)*m];
  likeF1=likeH1*Log[likeK1];
  Return[{likeK1, likeH1, likeF1}]
]
K1[m_] := NIntegrate[likelihood1[x, m][[1]], {x, mmin-m, mmax-m}]
H1[m_] := NIntegrate[likelihood1[x, m][[2]], {x, mmin-m, mmax-m}]
F1[m_] := NIntegrate[likelihood1[x, m][[3]], {x, mmin-m, mmax-m}]

```

The code computes the expected KL-divergences at a generic stage and put them in a vector.

```

ExpectKull[stage_] := (n=Rationalize[stage]; Print["Stage K =", " ", n];
  row=Dimensions[MatKull][[1]];
  If[rin+n-row>1, Print["former stage not perfoms"];
  Break[]];
  rownum>Delete[MatKull[[row]], 1];
  max=Position[rownum, Max[rownum]][[1, 1]];
  If[n==2, Hprec=H11, k=max; n=n-1;
  matrainv=Inverse[mattCova[n][[k]]];
  Cov=Inverse[Covamin[n]];
  Hprec=H[M[[max]], n];
  nextmax[[n]]=max; nextH[[n]]=Hprec; n=n+1
];
  kullback=Table[0, {k, numexpert}];
  For[k=1, k<=numexpert, k++,
    If[Abs[Det[mattCova[n][[k]]]]>10^-6,

```

```

matrainv=Inverse[mattCova[n][[k]]];
Cov=Inverse[Covamin[n]];
den=NIntegrate[H[m,n]/K[m,n],{\{ }m,-5,8{\} }];
num=NIntegrate[F[m,n]*H[m,n]/K[m,n],
{\{ }m,-5,8{\} }];
kullback[[k]]=num/den;
];
Print[kullback[[k]]]
];
MatKull=Join[MatKull,{Join[{K},kullback]}];
Return[MatrixForm[MatKull]]
)
likelihood[x_,m_,stage_] := (
vetG=Join[Table[G[x+m-mattM[n][[k]][[i]],i,{i,n-1}],{G[x,n]}];
vetg=Join[Table[g[x-mattM[n][[k]][[i]],i,{i,n-1}],{g[x,n]}];
vetB=Table[(vetG[[i]]+10^-20)/(1-vetG[[i]]+10^-20)*
(1-mattT[n][[k]][[i]])/mattT[n][[k]][[i]],{i,n}];
vetC=Log[vetB];
form=-1/2*vetC.matrainv.vetC;
expo=form-Sum[(m^2/2*1/mattV[n][[k]][[i]]+
m*(x-mattM[n][[k]][[i]])/mattV[n][[k]][[i]]},{i,n-1}];
fact=Product[vetg[[i]]/((vetG[[i]]+10^-20)*
(1-vetG[[i]]+10^-20)),{i,n}];
likelihoodK=Exp[expo]*fact;
likelihoodH=likelihoodK*Exp[-1/2*m^2*1/v0-m*(x-m0)/v0]*g0[x];
vetridC=Delete[vetC,n];
expogam=form+1/2*vetridC.Cov.vetridC;
Gam=vetg[[n]]/((vetG[[n]]+10^-20)*(1-vetG[[n]]+10^-20))*
Exp[expogam];
Return[{likelihoodK,likelihoodH,Gam}]
)
H[m_,n_] := NIntegrate[likelihood[x,m,n][[2]],{x,mmin-m,mmax-m}]
K[m_,n_] := NIntegrate[likelihood[x,m,n][[1]],{x,mmin-m,mmax-m}]
F[m_,n_] := (Hm=H[m,n];Fstage=1/Hm*NIntegrate[
likelihood[x,m,n][[2]]*Log[likelihood[x,m,n][[3]]],{x,mmin-m,
mmax-m}]-Log[Hm/nextH[n-1]]]; Return[Fstage]
)
Kullmax[stages_] := (Expect1;
maxim = {Position[Kull, Max[Kull]][[1, 1]], Max[Kull]};
MaxKull = maxim;
Print["Kmax :", maxim];
For[n = 2, n <= stages, n++, ExpectKull[n];
maxim={Position[kullback,Max[kullback]][[1, 1]],Max[kullback]};
Print["delta value: ", delta]; Print["Kmax ", maxim];
If[maxim[[2]] < delta, Break[], MaxKull = maxim];
]; jump = n - 1;
Return[{MatrixForm[MatKull], jump, MaxKull}]
)
Kullmax[stagemax]

```

Finally, the output shows:

- The expected KL-divergences of each density  $h_{j;k}(\theta)$  from the posterior  $h_{k-1}(\theta)$  obtained at the previous stage;

- The label  $j$  of the expert that satisfies the selecting criterion;
- A matrix containing the input data, KL-divergences at each stage, the number of experts to consult, and the label of the last expert to consult together with the corresponding expected KL divergence.

## 2.5 A Case-Study

The behavior of the algorithms and rules proposed in the previous sections, and implemented in *Mathematica*, has been investigated in simulation and experimental studies. Here the results from medical data are synthetically presented to exemplify how the selecting and stopping rules work.

An orthopedist  $A$  is uncertain about the long-term failure log-odds  $\theta$  of a new hip prosthesis. Therefore, he decides to consult a number  $K$  of colleagues, sequentially selected from a pool of size  $n = 7$ . He learns the variance  $v_j$  from each orthopedist  $Q_j$  ( $j = 1, \dots, n$ ) and assesses all the calibration parameters, without modifying them in proceeding from a stage to the successive one: these data are shown in Table 2.3. Subject  $A$  has also (subjectively) assessed  $m_0 = -1$ ,  $v_0 = 1$ , and set the threshold  $\delta = 0.035$ : the choice of this value means that, at stage  $k$ , the most informative expert  $Q_{j^*,k}^*$  will be consulted only if the expected KL-divergence of  $h_{j^*,k}(\theta)$  with respect to  $h_{k-1}(\theta)$  will be no less than the KL-divergence of a Bernoulli distribution  $B(p)$  with  $p = 0.5$  from a Bernoulli distribution with  $p = 0.63$ ; or, in other words, only if stopping the process at stage  $k - 1$  instead of proceeding to stage  $k$  involves, on average, an information loss larger than that one yielded by using a  $B(0.63)$  instead of a  $B(0.5)$ .

Conditions *i*), *ii*), and *iii*) in Sect. 2.2 are assumed to be satisfied, so that the combining algorithm outlined in Sect. 2.2 can be fairly applied. In fact: *i*) as confirmed by experts, it rests on empirical evidence that the failure logodds  $\theta$  can be supposed as Gaussian; *ii*) it is reasonable to think the probability the orthopedist  $A$  assigns to the event  $v^{(k)}$  is the same for all  $\theta$  values: the surgeons' variances alone give no information able to change the subject  $A$ 's beliefs about  $\theta$ ; *iii*) it is reasonable as well to assume the conditional probability  $A$  assigns to the event “the

**Table 2.3** Initial data for the case-study

$Q_j$	$v_j$	$t_j$	$s_{jj}$	$r_{j1}$	$r_{j2}$	$r_{j3}$	$r_{j4}$	$r_{j5}$	$r_{j6}$	$r_{j7}$
$Q_1$	0.90	0.35	3.93	1						
$Q_2$	0.40	0.60	4.86	+0.20	1					
$Q_3$	2.00	0.42	1.80	-0.1	-0.60	1				
$Q_4$	2.25	0.54	1.11	0	+0.30	0	1			
$Q_5$	1.80	0.50	1.31	0	+0.10	0	0	1		
$Q_6$	2.92	0.75	3.81	0	+0.20	-0.10	0	0	1	
$Q_7$	2.35	0.60	4.53	+0.20	0	-0.60	+0.30	+0.10	0	1

expert  $Q_{j;k}$  will give the variance  $v_k$ ," given the values of the shape and location parameters provided by the  $k - 1$  experts previously consulted, is the same for all  $\theta$  values. In order to perform an efficient sequential consulting, input data are entered into Mathematica notebook EXPS.nb:

```
Initialdata={{-2, -0.8, -1, -1.5, -1.3, -1.4, -1.35},
             {0.9, 0.4, 2, 2.25, 1.8, 2.92, 2.35},
             {0.35, 0.6, 0.42, 0.54, 0.5, 0.75, 0.6},
             {3.93, 4.86, 1.80, 1.11, 1.31, 3.81, 4.53}};
R={{1, 0.2, -0.1, 0, 0, 0, 0.2},{0.2, 1, -0.6, 0.3, 0.1,
   0.2, 0},{-0.1, -0.6, 1, 0, 0, -0.1, -0.6},{0, 0.3, 0,
   1, 0, 0, 0.3},{0, 0.1, 0, 0, 1, 0, 0.1},{0, 0.2, -0.1,
   0, 0, 1, 0},{0.2, 0, -0.6, 0.3, 0.1, 0, 1}};

delta=0.035;          v0=1; m0=-1;
mmmin=-8; mmax=11;    stagemax=7;
```

Mathematica output is the following:

Stage k = 1	Stage k = 2	Stage k = 3	Stage k = 4
0.340594	1.36673	1.57449	1.70397
0.519248	0	0	0
0.299469	1.65218	0	0
0.373058	1.37703	1.27123	1.40743
0.394185	1.47585	1.63163	0
0.102336	0.80258	0.42429	0.45441
0.121147	0.90094	1.56959	1.75234
Kmax:{2,0.5192}	delta value:0.035	delta value:0.035	delta value:0.035
	Kmax:{3,1.65218}	Kmax:{5, 1.63163}	Kmax:{7, 1.75234}

Stage k = 5	Stage k = 6	Stage k = 7
2.14744	2.17423	0
0	0	0
0	0	0
2.62677	0	0
0	0	0
0.310055	0.0533287	0.034978
0	0	0
delta value:0.035	delta value:0.035	delta value:0.035
Kmax:{4,2.62677}	Kmax:{1, 2.17267}	Kmax:{6, 0.034978}

The output shows that, at stage  $k = 1$ , expert  $Q_2$  is selected, due to an expected KL-divergence equal to 0.519248. At stage  $k = 2$ , the maximum expected KL-divergence corresponds to expert  $Q_3$ , characterized by a high negative correlation with  $Q_2$  ( $r = -0.60$ ) together with a low bias ( $t = 0.42 \cong 0.5$ ). Expert  $Q_5$  is, on average, the most informative at stage  $k = 3$ , and so on. At stage  $k = 7$ , the expected KL-divergence for the last expert,  $Q_6$ , is 0.034978, that is less than the threshold  $\delta$ : since  $Q_6$  does not involve a knowledge expected gaining judged as relevant, subject  $A$  will not consult him. So, as the output matrix shows too,

the sequential consulting stops at stage  $k = 6$ : the last expert who enters into the process is  $Q_1$ , with an expected KL-divergence equal to 0.0533287.

exp	1	2	3	4	5	6	7	
$m$	-2	-0.8	-1	-1.5	-1.3	-1.4	-1.35	
$v$	0.9	0.4	2	2.25	1.8	2.92	2.35	
$t$	0.35	0.6	0.42	0.54	0.5	0.75	0.6	
$s_2$	3.93	4.86	1.80	1.10	1.31	3.81	4.53	
$K$	0.34059	0.51925	0.29947	0.37306	0.39419	0.10234	0.12115	
$K$	1.36673	0	1.65218	1.37703	1.47585	0.80258	0.90094	$, 6, \{1, 2.17267\}$
$K$	1.57449	0	0	1.27123	1.63163	0.42429	1.56959	
$K$	1.70397	0	0	1.40743	0	0.45441	1.75234	
$K$	2.14744	0	0	2.62677	0	0.310155	0	
$K$	2.17423	0	0	0	0	0.053329	0	
$K$	0	0	0	0	0	0.034978	0	

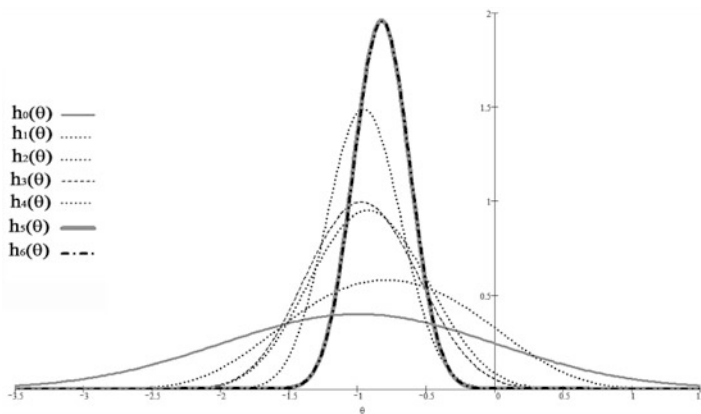
Table 2.4 shows the expected KL-divergences for each expert at each stage (the maximum expected KL-divergence is displayed in bold), as well as the location parameters supplied by the selected experts. Posterior distributions from stage 0 (the prior) to stage 6 are shown in Fig. 2.3. The ‘final’ pdf on the unknown log-odds  $\theta$  has mean, median, and mode equal to  $-0.826$  and standard deviation equal to 0.202: it can be regarded as the synthesis representation of the expert knowledge about the long-term failure log-odds of the new hip prostheses.

The behavior of the proposed rules and algorithms in the case-study appears to be coherent with the intuition and gives an empirical support about the efficiency of the selecting and stopping criteria.

Table 2.4 Results of the sequential process

	Stage	Stage	Stage	Stage	Stage	Stage
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
$Q_j$	$E[KL(h_{j:1}, h_0)]$	...	...	...	...	$E[KL(h_{j:6}, h_5)]$
$Q_1$	0.34059	1.36673	1.57449	1.70397	2.14744	<b>2.17423</b>
$Q_2$	<b>0.51925</b>	—	—	—	—	—
$Q_3$	0.29947	<b>1.65218</b>	—	—	—	—
$Q_4$	0.37306	1.37703	1.27123	1.40743	<b>2.62677</b>	—
$Q_5$	0.39419	1.47585	<b>1.63163</b>	—	—	—
$Q_6$	0.10234	0.80258	0.42429	0.45441	0.310155	0.053329
$Q_7$	0.12115	0.90094	0.56959	<b>1.75234</b>	—	—
	↓	↓	↓	↓	↓	↓
$Q_{j:k}^*$	$Q_2$	$Q_3$	$Q_5$	$Q_7$	$Q_4$	$Q_1$
$m_k$	-0.8	-1	-1.3	-1.35	-1.5	-2





**Fig. 2.3** Posterior distributions at stages 0 (i.e., the prior), 1, 2, ..., 6 of the sequential procedure

## References

1. Agati, P., Calò, D.G., Stracqualursi, L.: A joint calibration model for combining predictive distributions. *Statistica* **2**, 203–212 (2007)
2. Brown, J.R., Harvey, M.E.: Arbitrary precision mathematica functions to evaluate the one-sided one sample K–S cumulative sampling distribution. *J. Stat. Softw.* **26**(3), 1–55 (2008). <http://www.jstatsoft.org/v26/i03>
3. Kullback, S.: *Information Theory and Statistics*. Wiley, New York (1959)
4. Lindley, D.V.: The 1988 Wald Memorial Lectures: the present position in Bayesian statistic. *Stat. Sci.* **5** (1), 44–89 (1990)
5. McCulloch, R.: Local model influence. *J. Am. Stat. Assoc.* **84**, 473–478 (1989)
6. Monari, P., Agati, P.: Fiducial inference in combining expert judgements. *J. Ital. Stat. Soc.* **84**, 81–97 (2001)
7. Morris, P.A.: Combining expert judgments: a Bayesian approach. *Manag. Sci.* **23**, 679–693 (1977)
8. Wolfram, S.: *The Mathematica Book*. 5th edn. Wolfram Media, Champaign, USA (2003)

Topics in Statistical Simulation

Research Papers from the 7th International Workshop  
on Statistical Simulation

Melas, V.B.; Mignani, S.; Monari, P.; Salmaso, L. (Eds.)

2014, XVIII, 539 p. 85 illus., 34 illus. in color., Hardcover

ISBN: 978-1-4939-2103-4