XX IMEKO World Congress
Metrology for Green Growth
September 9–14, 2012, Busan, Republic of Korea

MEASUREMENT UNCERTAINTY AND SUMMARISING MONTE CARLO SAMPLES

A. B. Forbes

National Physical Laboratory, Teddington, UK, alistair.forbes@npl.co.uk

Abstract: Many uncertainty evaluation applications require summary information about the distribution associated with the measurand. This paper looks at summarizing distributions on the basis of Monte Carlo samples and describes how relative entropy can be used as a measure of the effectiveness of the summary.

Keywords: kernel density estimation, measurement uncertainty, mixture distribution, relative entropy

1. INTRODUCTION

The Guide to the Expression of Uncertainty in Measurement (GUM [1]) describes how information about a quantity, gained from measurement, can be represented in terms of a probability density function (PDF). Summary information about the quantity can be derived from the PDF. In particular, the mean of the distribution is taken as the estimate of the quantity and the standard deviation is taken to be the standard uncertainty associated with the estimate. To determine further summary information, such as a coverage interval, the GUM assumes that the PDF is either a Gaussian or a *t*-distribution, with the degrees of freedom parameter providing additional shape information. Thus, the output from the GUM uncertainty framework is condensed into two or three parameters.

Supplement 1 to the GUM (GUMS1 [2]) uses a Monte Carlo method (MCM) to produce a random sample from the distribution associated with the measurand. In this case, the distribution is represented by, perhaps, tens of thousands of draws but there is still the requirement to provide summary information about the measurand derived from the sample. Similar remarks apply to Bayesian interpretations of the GUM uncertainty framework [3,4,5,6,7] in which a random sample from the posterior distribution is generated using a Markov chain Monte Carlo (MCMC) algorithm [8,9]. Thus, for the MCM and MCMC approaches, the question arises of how to provide summary information about the distribution associated with the measurand.

In this paper, we are concerned with approximating distributions on the basis of Monte Carlo samples and measuring the information loss associated with the approximation. The information loss has two possible sources, firstly the form of the approximating distribution and secondly the fact that information about the distribution is represented by a finite sample. In section 2, we consider a class of approximating distributions based on mixture distributions involving standard distributions, rectangular or Gaussian distributions, and determining an approximation from a sample. In particular, we provide a formulation of the problem of finding a maximum likelihood estimate of a mixture of Gaussians that leads to a relatively simple optimization problem involving only linear inequality and equality constraints. In section 3, we consider the use of relative entropy as a measure of the quality of approximation or information loss. In section 4, we analyse the behavior of the approximation schemes using a simple numerical example. A discussion and concluding remarks are given in section 5.

2. APPROXIMATING A DISTRIBUTION FROM A SAMPLE

Suppose $\{x_i, i = 1, \dots, M\}$ is a random sample from a distribution with density $p(\xi)$. We look for ways of deriving a density $q(\xi)$ from the sample such that inferences based on $q(\xi)$ will be as reliable as possible, i.e., as close as possible to those that would be made if $p(\xi)$ was available. We consider four general approaches to generating the approximating distribution. First, determine the mean $\bar{x}$ and standard deviation s of the sample and assign the PDF $q_{\mathrm{N}}(\xi|\bar{x}, s^2)$, the PDF associated with the Gaussian distribution $\mathrm{N}(\bar{x}, s^2)$. This approach can be generalized to associate to the measurand a particular member of a parametric distribution based on information (moments, etc.) derived from the sample [10]. Second, assign a PDF $q_{\mathrm{H}}(\xi)$ derived from a histogram of the sampled points with bin edges given by $\boldsymbol{z} = (z_1, \dots z_n)^{\mathrm{T}}$ with $-\infty \le z_1 < z_2 < \cdots < z_n \le \infty$. Third, use kernel density estimation methods [11]

$$q_{\mathrm{K}}(\xi) = \frac{1}{hM}\sum_{i=1}^{M} K\left(\frac{\xi - x_i}{h}\right),$$

where K is a symmetric function integrating to 1. Here, we set $K(\xi) = q_{\mathrm{N}}(\xi|0,1)$, the PDF associated with a standard Gaussian. The bandwidth or scale parameter h above plays a similar role to that of the bin width using a histogram approximation.

The three approaches can be seen as approximating the distribution $p(\xi)$ by a mixture distribution

$$q(\xi) = \sum\nolimits_{j=1}^{N} w_j q_j(\xi),$$

where each $q_j(\xi)$ is a well-known PDF and the nonnegative weights w_j sum to 1. The first approach involves only one Gaussian summand, the histogram approach uses uniform distributions on the intervals $[z_j, z_{j+1}]$, while the kernel density method uses Gaussians. The number of summand functions for the three approaches are 1 (the minimum), N, and M (the maximum).

The fourth approach used here explicitly involves a mixture of Gaussians

$$q_{\mathrm{M}}(\xi|\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\sigma}) = \sum_{j=1}^{N} w_j q_{\mathrm{N}}(\xi|\mu_j,\sigma_j{}^2),$$

The mean μ and variance σ^2 of such a mixture distribution are given by

$$\mu = \sum_{j=1}^{N} w_j\mu_j, \;\; \sigma^2 = \sum_{j=1}^{N} w_j\left(\left(\mu_j-\mu\right)^2+\sigma_j^2\right).$$

Given a sample $\{x_i, i = 1, \ldots, M\}$, optimal estimates of the parameters $\boldsymbol{w}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ can be determined by maximizing the likelihood

$$l(\boldsymbol{x}|\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\sigma}) = \prod_i q_{\mathrm{M}}(x_i|\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\sigma})$$

subject to non-negativity constraints on the weights and the additional necessary equality constraint $\sum_j w_j = 1$. This optimization problem is likely to be ill-conditioned if there are more than a few summand distributions. To remove such ill-conditioning, we consider a number of constraints on the mixture distribution. We first assign the same dispersion parameter σ_0 to each summand distribution, i.e., $\sigma_j = \sigma_0$. We also apply the constraints that the mixture distribution has the same mean and standard deviation as the sample, so that

$$\sum_j w_j\mu_j = \bar{x}, \qquad \sigma_0^2 = s^2 - \sum_{j=1}^{N} w_j\left(\mu_j-\bar{x}\right)^2.$$

In the presence of the first constraint above, the second constraint can be written as

$$\sigma_0^2 = s^2 + \bar{x}^2 - \sum_{j=1}^{N} w_j\mu_j^2.$$

By subtracting the mean from the data we can assume that $\bar{x} = 0$, so that the constraints on the problem can be written as

$$w_j \geq 0\,, \quad \sum_j w_j = 1, \qquad \sum_j w_j\mu_j = 0$$

and

$$\sigma_0^2 = s^2 - \sum_{j=1}^{N} w_j\mu_j^2.$$

This latter constraint implies an additional linear inequality constraint on the weights, namely,

$$\sum_{j=1}^{N} w_j\mu_j^2 \leq s^2.$$

If the means μ_j are fixed then all constraints are linear equality or inequality constraints on the weights w_j and the variance term σ_0^2. An additional degree of freedom can be included by introducing a single scale parameter $\lambda > 0$ associated with the mean locations. In this case the constraints can be written as

$$\widetilde{w}_j \geq 0\,, \quad \sum_j \widetilde{w}_j = \lambda^2, \qquad \frac{1}{\lambda}\sum_j \widetilde{w}_j\mu_j = 0,$$

and

$$\sigma_0^2 = s^2 - \sum_{j=1}^{N} \widetilde{w}_j\mu_j^2, \quad \sum_{j=1}^{N} \widetilde{w}_j\mu_j^2 \leq s^2,$$

where $\widetilde{w}_j = w_j\lambda^2$. Thus, the problem can be posed in terms of the un-normalised weights $\widetilde{w}_j$. The effect of introducing the scale parameter λ is to modify the constraint on the sum of these scaled weights. The normalised weights are given by $w_j = {}^{\widetilde{w}_j}/_{\lambda^2}$, with $\lambda = \left(\sum_{j=1}^{N}\widetilde{w}_j\right)^{1/2}$, and the summand Gaussians for the mixture distribution have means $\tilde{\mu}_j = \lambda\mu_j$. In the numerical example discussed in section 4, we take for an initial set of means the centres of N bins determined from the sample adjusted so that the centre of one bin coincides with the sample mean.

With these additional constraints on the mean and standard deviation of the mixture distribution, the determination of the maximum likelihood estimate amounts to maximizing a nonlinear function subject to linear equality and inequality constraints. This formulation of the problem avoids nonlinear constraints that would make the optimization problem much more difficult to solve.

3. RELATIVE ENTROPY

We look for a measure of how well the distribution derived from the sample approximates the distribution generating the sample. Given two probability distributions with densities $p(\xi)$ and $q(\xi)$, the relative entropy (or Kullback–Leibler divergence) is given by [12]

$$D_{\mathrm{KL}}(p||q) = \int p(\xi)\log\left(\frac{p(\xi)}{q(\xi)}\right)d\xi.$$

The relative entropy is not a metric since it is not symmetric, but the Gibbs inequality can be used to show that the relative entropy is zero only if the distributions are essentially the same [12]. If p is close to q in the sense that $1 - \frac{p}{q} \approx 0$, then $p\log(p/q) \approx p - q$. The relative entropy is an important concept in information theory. In Bayesian inference, the relative entropy $D_{\mathrm{KL}}(p(\xi|\boldsymbol{y})||p(\xi))$ of the posterior distribution $p(\xi|\boldsymbol{y})$, relative to the prior distribution $p(\xi)$, is a measure of the information gain about ξ derived from the measurement data $\boldsymbol{y}$.

If $p(\xi)$ is the density associated with the Gaussian distribution $\mathrm{N}(\alpha_1, \sigma_1^2)$ and $q(\xi)$ that associated with $\mathrm{N}(\alpha_2, \sigma_2^2)$, then

$$D_{\mathrm{KL}}(p||q) = \frac{1}{2}\left[\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right)+\frac{\sigma_1^2}{\sigma_2^2}-1\right]+\frac{1}{2}\left(\frac{\alpha_1-\alpha_2}{\sigma_2}\right)^2 \quad .$$

The first term depends only on the standard deviations of the distributions while the second term also takes into account the difference in their means. Similarly, if $p(\xi)$ is the density associated with the rectangular distribution R(A,B) and $q(\xi)$ that associated with $\mathrm{N}(\alpha,\sigma^2)$, then

$$D_{\mathrm{KL}}(p||q) = \log\left[\frac{(2\pi\sigma^2)^{\frac{1}{2}}}{B-A}\right] + \frac{\sigma}{6(B-A)}\left[\left(\frac{B-\alpha}{\sigma}\right)^3 - \left(\frac{A-\alpha}{\sigma}\right)^3\right] \ .$$

If $p(\xi)$ has mean μ and standard deviation σ, it is straightforward to show that the Gaussian distribution $p_{\mathrm{N}}(\xi|\mu',\sigma')$ that minimises

$$D_{\mathrm{KL}}(p(\xi)||p_{\mathrm{N}}(\xi|\mu',\sigma'))$$

is $p_{\mathrm{N}}(\xi|\mu,\sigma)$. Thus, if we are going to approximate a distribution with a Gaussian distribution, then the best such distribution to choose according to the relative entropy criterion is that with the same mean and standard deviation.

If a sample is drawn from $p(\xi)$ and used to construct an approximating distribution $q(\xi)$, we use the relative entropy $D_{\mathrm{KL}}(p||q)$ as a measure of the quality of the approximation derived from the sample.

4. NUMERICAL EXAMPLE

This example concerns the distribution $p(\xi)$ associated with the random variable $X = \exp Y/5$ where $Y \sim \mathrm{N}(0,1)$. Using the transformation of variables rule [13], it is straightforward to write down an analytical formula for $p(\xi)$, namely

$$p(\xi) = \frac{5}{\xi} p_{\mathrm{N}}(5\log\xi|0{,}1), \quad \xi > 0,$$

where $p_N(\xi|0{,}1)$ is the standard Gaussian density.

If $\{y_i, i = 1,\dots,M\}$ is a sample from N(0,1) then $\{x_i = \exp y_i/5, i = 1,\dots,M\}$ is a sample according to $p(\xi)$. Figure 1 graphs $p(\xi)$ (blue, solid) along with the Gaussian approximant (green, long dash) and the approximant derived from a histogram with 9 bins (red, short dash) derived from M = 10,000 samples. Figure 2 graphs $p(\xi)$ (blue, solid) along with approximants derived using the kernel density approach (green, long dash) and a Gaussian mixture (red, short dash) with 5 components also derived from M = 10,000 samples. It is seen that the approximations in Figure 2 are much better than those in Figure 1.

A measure of the quality of the approximation is given by the relative entropy. Table 1 shows the relative entropy calculations $D_{\mathrm{KL}}(p||q_{\mathrm{H}})$ and $D_{\mathrm{KL}}(p||q_{\mathrm{M}})$ for the histogram and mixture approximations, respectively, for various numbers of components. Each calculation involved the same sample of 10,000 random draws. For comparison, the relative entropy $D_{\mathrm{KL}}(p||q_{\mathrm{N}})$ and $D_{\mathrm{KL}}(p||q_{\mathrm{K}})$ for the Gaussian and kernel density approximations, q_{N} and q_{K}, are 0.0298 and 0.0011, respectively. (Both the Gaussian and kernel approximations do not depend on the number of bins, etc.)

We also calculate the relative entropy $D_{\mathrm{KL}}(p||q_{\mathrm{M}}^*)$ of $p(\xi)$ relative to q_{M}^*, the mixture distribution that minimises the relative entropy derived directly from $p(\xi)$ rather than from the sample. Table 1 also shows the relative entropy associated with the mixture approximation to the kernel approximant. Again we consider two mixture approximants, the first derived from the sample, leading to the calculation of $\mathrm{D}_{\mathrm{KL}}(q_{\mathrm{K}}||q_{\mathrm{M}})$, and, the second, the mixture distribution that minimises the relative entropy derived directly from q_{K}, leading to $D_{\mathrm{KL}}(q_K||q_{\mathrm{M}}^*)$. We are interested in these calculations because for most practical applications, it is not possible to provide an analytical definition of $p(\xi)$ but the kernel density approximant can be calculated instead. However, we would like to represent this approximant more compactly, for example, using a mixture distribution with a modest number of components.

From table 1, we note:

- The histogram approximation q_{H} of p is much less accurate than the corresponding mixture distribution.
- The mixture distribution approach provides successful approximations for N = 5 or more. Figure 3 shows a graph of $p\log(p/q)$ where q is the kernel density q_{K} approximant (blue, solid), the Gaussian mixture approximant $q_{\mathrm{M,5}}$ with 5 summand functions determined from the data using maximum likelihood estimation (green, long dash), and the Gaussian mixture approximant $q_{\mathrm{M,5}}^*$ with 5 summand functions (red, short dash) determined by minimising the relative entropy $D_{\mathrm{KL}}(q_K||q_{\mathrm{M}}^*)$ directly. The figure provides evidence that the kernel density approximant is slightly better. Figure 4 shows the same information but for Gaussian mixture approximants with 7 summand distributions. In this case it can be argued that the Gaussian mixture distributions provide better approximants.
- Comparing the values of $D_{\mathrm{KL}}(p||q_{\mathrm{M}})$ with $D_{\mathrm{KL}}(p||q_{\mathrm{M}}^*)$ and also those of $\mathrm{D}_{\mathrm{KL}}(q_{\mathrm{K}}||q_{\mathrm{M}})$ with $D_{\mathrm{KL}}(q_K||q_{\mathrm{M}}^*)$, we see that the quality of the approximant determined from the sample data using maximum likelihood estimation is similar to that calculated directly from p or q_{K}. (Compare the green, long-dashed curves with the red, short-dashed curves in figures 3 and 4). In other words, the sample contains essentially the same information as p itself as far as determining the Gaussian mixture approximant.

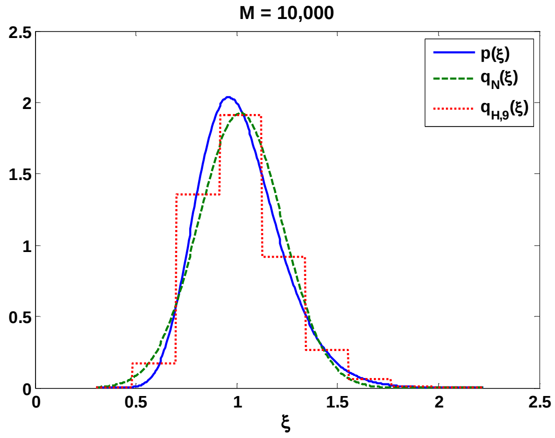

Figure 1. Gaussian and nine-bin histogram approximations to $p(\xi)$ based on 10,000 samples.

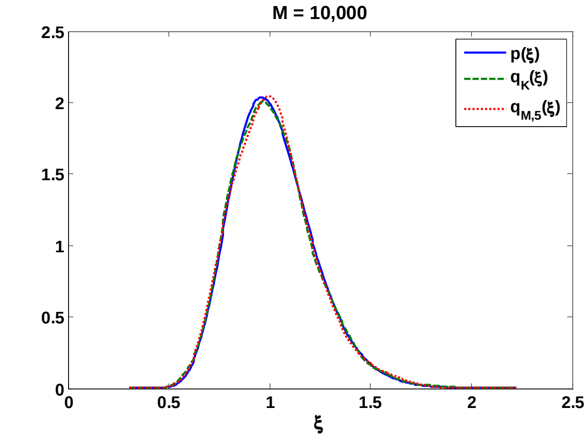

Figure 2. Graph of p and approximating distributions calculated from kernel density and Gaussian mixture distributions (with 5 summand functions) based on 10,000 samples.

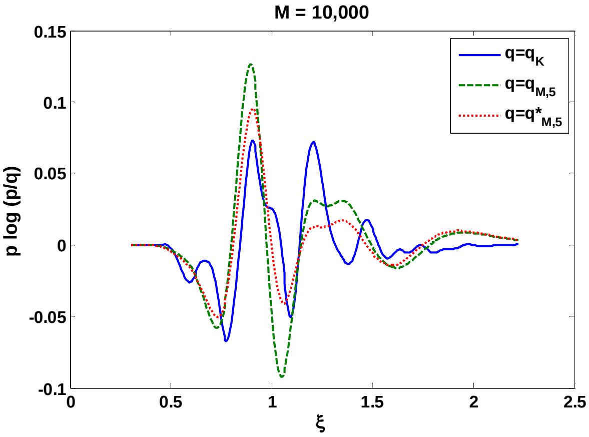

Figure 3. Graph of $p\log(p/q)$ where q is calculated from kernel density and Gaussian mixture distributions with 5 summand functions based on 10,000 samples.

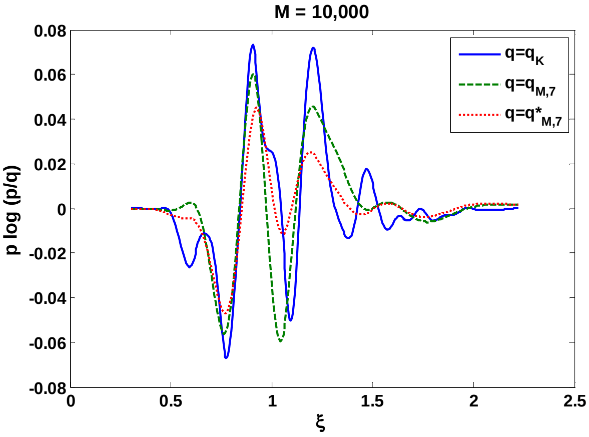

Figure 4. Graph of $p\log(p/q)$ where q is calculated from kernel density and Gaussian mixture distributions with 7 summand functions based on 10,000 samples. The blue, solid curves in figures 3 and 4 are the same.

Table 1. Relative entropy associated with histogram and mixture approximations to $p(\xi)$ for different numbers of components N, 10,000 samples.

N	$p\|\|q_{\mathrm{H}}$	$p\|\|q_{\mathrm{M}}$	$p\|\|q^*_{\mathrm{M}}$	$q_{\mathrm{K}}\|\|q_{\mathrm{M}}$	$q_K\|\|q^*_{\mathrm{M}}$
1	0.7165	0.0298	0.0298	0.0328	0.0328
3	0.1506	0.0243	0.0242	0.0253	0.0252
5	0.1475	0.0032	0.0029	0.0042	0.0040
7	0.0791	0.0009	0.0006	0.0012	0.0010
9	0.0509	0.0005	0.0003	0.0009	0.0006

4.1 Effect of sample size on the quality of the approximants

The effect of sample size can be examined by performing the same calculations for different sample sizes M. Tables 2 and 3 provide the same information as table 1 but for sample sizes of M = 3,000 and M = 1,000, respectively. (The values of $D_{\mathrm{KL}}(p||q^*_{\mathrm{M}})$ differ across tables 1, 2 and 3 due to the fact that the means for the summand Gaussians are determined from histograms derived from the sample.) For the Gaussian approximation, $D_{\mathrm{KL}}(p||q_{\mathrm{N}})$ = 0.0298, 0.0301 and 0.0310 for M = 10,000, 3,000 and 1,000, respectively while for the kernel density estimate, the corresponding values are $D_{\mathrm{KL}}(p||q_{\mathrm{K}})$ = 0.0011, 0.0029, and 0.0088. Thus, there is some evidence that a sample size of 1,000 is not sufficiently large to capture the information represented by the distribution p (for this example). In terms of the quality of the histogram approximation, the smaller sample sizes have no significant effect on the relative entropy values. This reflects the fact that the loss of information about p arises not from the finite sampling but from the form of the approximating distribution. For the Gaussian mixture approximations, comparing table 3 with tables 1 and 2, it is seen that the smaller sample size is losing pertinent information about p as far as approximations with 5 or more summand functions.

Figures 5 and 6 correspond to figure 3 but for sample size of 3,000 and 1,000 respectively, and show, for the case of M = 1,000 especially, the effect of the small sample size on the quality of the maximum likelihood approximation determined from the sample.

Table 2. Relative entropy associated with histogram and mixture approximations to $p(\xi)$ for different numbers of components N, 3,000 samples.

N	$p\|\|q_{\mathrm{H}}$	$p\|\|q_{\mathrm{M}}$	$p\|\|q^*_{\mathrm{M}}$	$q_{\mathrm{K}}\|\|q_{\mathrm{M}}$	$q_K\|\|q^*_{\mathrm{M}}$
1	0.7165	0.0301	0.0301	0.0309	0.0309
3	0.1500	0.0247	0.0245	0.0242	0.0241
5	0.1521	0.0034	0.0031	0.0077	0.0072
7	0.0816	0.0015	0.0008	0.0034	0.0027
9	0.0532	0.0013	0.0005	0.0020	0.0013

Table 3. Relative entropy associated with histogram and mixture approximations to $p(\xi)$ for different numbers of components N, 1,000 samples.

N	$p\|\|q_{\mathrm{H}}$	$p\|\|q_{\mathrm{M}}$	$p\|\|q^*_{\mathrm{M}}$	$q_{\mathrm{K}}\|\|q_{\mathrm{M}}$	$q_K\|\|q^*_{\mathrm{M}}$
1	0.7165	0.0310	0.0310	0.0340	0.0340
3	0.1532	0.0256	0.0251	0.0311	0.0308
5	0.1591	0.0085	0.0041	0.0074	0.0045
7	0.0890	0.0074	0.0017	0.0053	0.0016
9	0.0628	0.0074	0.0014	0.0053	0.0016

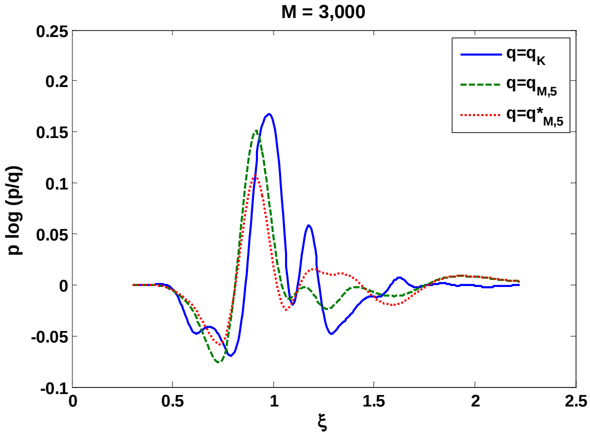

Figure 5. Graph of $p\log(p/q)$ where q is calculated from kernel density and Gaussian mixture distributions with 5 summand functions based on 3,000 samples. The red, dotted curves in figures 3, 5 and 6 are the same curve.

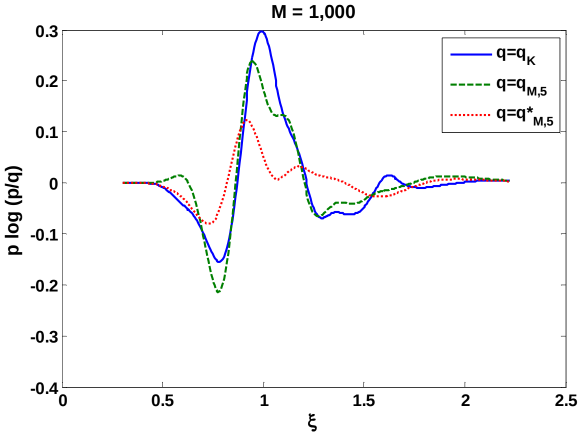

Figure 6. Graph of $p\log(p/q)$ where q is calculated from kernel density and Gaussian mixture distributions with 5 summand functions based on 1,000 samples. The red, dotted curves in figures 3, 5 and 6 are the same curve.

5. DISCUSSION AND CONCLUDING REMARKS

The kernel density approximation derived from a sample can usually be regarded as a reasonably accurate approximation to the underlying distribution that gave rise to the sample, assuming the sample size M is large enough. However, the kernel density estimate is defined in terms of M summands making it potentially unwieldy for further inferences. Our approach to determining a mixture distribution involving a modest number N of component functions as discussed here has some advantages. The distribution is defined by $2N$ items of information: N means and N weights with the common standard deviation determined by the constraint that the mixture distribution has the standard deviation as that of the sample. Our formulation also ensures that the mixture distribution has the same mean as that of the sample. These constraints can be implemented in terms of linear inequality and equality constraints, thereby reducing the computational complexity of the associated optimisation problem. For large samples, initial estimates of the optimisation parameters can be determined from a subsample, so that only a few iterations are required involving the complete sample.

All probabilities involving the mixture distribution are easily calculated in terms of the standard component distributions, and the mean and standard deviation of the mixture distribution can be calculated exactly.

Relative entropy can be used as a measure of the quality of an approximating distribution and can be used to determine an appropriate number of summand functions to represent the distribution directly or through its proxy, a kernel density estimate.

6. ACKNOWLEDGEMENTS

This work was supported by the UK's National Measurement Office Mathematics and Modelling for Metrology programme. Colleagues Maurice Cox, Peter Harris and Clare Matthews made helpful comments on drafts of this paper.

7. REFERENCES

[1] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, Guide to the expression of uncertainty in measurement, (GUM: 1995, with minor corrections), Bureau International des Poids et Mesures, JCGM 100: 2008.

[2] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, Guide to the expression of uncertainty in measurement – Supplement 1: Propagation of distributions using the Monte Carlo method, Bureau International des Poids et Mesures JCGM 101:2008.

[3] M G Cox, A B Forbes, P M Harris and C E Matthews, Numerical aspects in the evaluation of measurement uncertainty, in *Uncertainty Quantification* (IFIP 2011), Springer, to appear.

[4] C Elster, W Wöger and M G Cox. Draft GUM Supplement 1 and Bayesian analysis, *Metrologia*, 44, L31–L32, 2007.

[5] C Elster and B Toman, Bayesian uncertainty analysis under prior ignorance of the measurand versus analysis using Supplement 1 to the Guide: a comparison. Metrologia, 46, 261–266, 2009.

[6] A B Forbes and J A Sousa, The GUM, Bayesian inference and forward and inverse uncertainty evaluation. *Measurement*, 44, 1422–1435, 2011.

[7] A B Forbes, An MCMC algorithm based on GUM Supplement 1 for uncertainty evaluation. *Measurement*, online DOI: 10.1016/j.measurement.2012.01.018, 2012.

[8] D Gamerman, *Markov Chain Monte Carlo*, Chapman & Hall/CRC, Boca Raton, 1999.

[9] A Gelman, J B Carlin, H S Stern and D B Rubin, *Bayesian Data Analysis*, 2nd Edn. Chapman & Hall/CRC, Boca Raton, 2004.

[10] R Willink, Transferring Monte Carlo distributions in the evaluation of uncertainty, in Advanced Mathematical and Computational Tools for Metology VIII, F Pavese, et al. eds. World Scientific, Singapore, pp 351–356, 2008.

[11] B W Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman&Hall, London, 1986.

[12] D J C MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, Cambridge, 2003.

[13] G Grimmett and D Stirzaker, *Probability and Random Processes*, 3rd Edn. Oxford University Press, Oxford, 2001.