Open Journal of Statistics, **2013, 3, 1-8**
Published Online December 2013 (http://www.scirp.org/journal/ojs)
http://dx.doi.org/10.4236/ojs.2013.36A001

Scientific Research

Subjectivity in Application of the Principle of Maximum Entropy

Jan Peter Hessling
Measurement Technology, SP Technical Research Institute, Borås, Sweden
Email: peter.hessling@sp.se

Received September 24, 2013; revised October 24, 2013; accepted November 1, 2013

ABSTRACT

Complete prior statistical information is currently required in the majority of statistical evaluations of complex models. The principle of maximum entropy is often utilized in this context to fill in the missing pieces of available information and is normally claimed to be fair and objective. A rarely discussed aspect is that it relies upon *testable information*, which is never known but estimated, *i.e.* results from processing of raw data. The subjective choice of this processing strongly affects the result. Less conventional posterior completion of information is equally accurate but is computationally superior to prior, as much less information enters the analysis. Our recently proposed methods of lean deterministic sampling are examples of very few approaches that actively promote the use of minimal incomplete prior information. The inherited subjective character of maximum entropy distributions and the often critical implications of prior and posterior completion of information are here discussed and illustrated, from a novel perspective of consistency, rationality, computational efficiency and realism.

Keywords: Maximum; Entropy; Bayes; Monte Carlo; Uncertainty; Covariance; Deterministic; Sampling; Testable; Information; Model; Calculation; Simulation

1. Introduction

The principle of maximum entropy (PME) can be utilized to determine a probability density distribution function (pdf) from incomplete statistical information. The approach is not limited to determination of prior pdfs in Bayesian estimation, even though that is a common application. It is rather a general recipe how to make known but incomplete statistical information complete with the most 'fair' or the weakest possible hypotheses. As such it fits very well into our practice of statistics, where the lack of complete information is the rule rather than the exception. For instance, knowing only the mean and the variance of an uncertain parameter PME results in a normal distribution [1]. The known information must be well defined in a statistical sense, *i.e.* be formulated in terms of statistical expectations $\langle f(\theta)\rangle$ of functions $f(\theta)$ of the considered random quantity θ. For instance, $f(\theta)=\theta$ yields the mean and $f=(\theta-\langle\theta\rangle)^q$ produces the q-th moment around the mean. Their expectations can be estimated from any set of observations $\{\theta_k\}$ with estimators [2] $\hat{f}$ to give *testable information* $\phi\equiv\langle f(\theta)\rangle\approx\hat{f}(\{\theta_k\})$ [3].

Bayesian estimation generalizes traditional approaches limited to observations by inclusion of prior knowledge. Its claimed advantage or superiority relies heavily upon fair and truthful assignment of the prior pdf. If the applied method (like PME) to determine the prior pdf turns out to be subjective it would degrade the legitimacy of the approach. Indeed, PME is often motivated by its fairness or objectivity [1]. A minimum of supplementary (unknown) statistical information is imposed by, loosely speaking, maximizing the residual randomness as measured with the information entropy introduced by Shannon [4] and further explored by Jaynes [3]. In practice, the procedure does not provide a complete recipe. By starting with known testable information, PME avoids to define its the best form. There is a widely spread practice though, which probably originates from the ubiquitous use of Taylor expansions. Testable information is usually considered in a hierarchy starting from the mean, the covariance, the skewness, and the

kurtosis etc. Various statistical moments around the mean are tested, as if they were terms of a Taylor series. The functions $g_q = \left(\theta - \langle\theta\rangle\right)^q$ are indeed the q-th order monomial related to a Taylor expansion around the mean. The expectation $\langle g_q \rangle$ will contribute to the non-linear displacement $\langle h(\theta)\rangle - h(\langle\theta\rangle)$, or *scent* [5] of q-th order, of the model h. For any other statistic like e.g. covariance, no such simple direct relation holds [6]. Another line of reasoning might be that the mean describes the location of the distribution, the second moment the width, the third the lowest order asymmetry, while the fourth is the lowest order shape indicator etc. We might subjectively claim that these properties (location-width-asymmetry-shape) provide a hierarchy of testable information. This does not imply that the moments themselves drop in magnitude or relevance, with their order. On the contrary, the linearly scaled even moments $M_{2q} = \langle g_{2q}(\theta)\rangle^{1/2q}$ usually *increase* with the order q. In the limit $q \to \infty$, $M_{2q} \to \max_k |\theta_k|$, where θ_k are the observations of θ.

Given a set of observations it is not evident how they should be processed to provide testable information of the highest possible quality, *i.e.* with minimal residual uncertainty. For instance, which moments $\langle g_q \rangle$ should be estimated with a quality (variance of estimator) that usually decrease with the order q? As will be shown, the selection will directly influence the PME distribution function. It is also not evident why PME should be restricted to prior application, as usually practiced. Posterior utilization might in fact simplify the analysis greatly. After all, PME is a general method with no explicit reference to what the distribution describes, the input or the output of the analysis.

2. Method of Maximum Entropy

For a continuous sample space Ω, the entropy [4] functional to be maximized in the method of maximum entropy is given by the Lebesgue integral,

$$S = -\int_{\bar{\Omega}} \bar{p}(\bar{\theta}) \log\left(\bar{p}(\bar{\theta})\right) \mathrm{d}\bar{\theta}. \tag{1.1}$$

The bar-symbol restricts the integration over *distinguishable* outcomes $\bar{\theta} \in \bar{\Omega} \subset \Omega \ni \theta$. That accounts for our possible ignorance[1] of not distinguishing distinct outcomes. As there is an obvious contradiction of being aware of ignorance, it is better translated into irrelevance (for our stated problem). The integration over $\bar{\Omega}$ can be extended to Ω, by *locally* measuring the relative density of $\bar{\Omega}$ with the Lebesgue measure $m(\theta)$, $\mathrm{d}\bar{\theta} = m(\theta)\mathrm{d}\theta$. Consistency requires *transformation invariance* [1] of the probability $\mathrm{d}P$ in the interval $\bar{\theta}, \bar{\theta} + \mathrm{d}\bar{\theta}$. That is, $\bar{p}(\bar{\theta})\mathrm{d}\bar{\theta} = p(\theta)\mathrm{d}\theta$, giving the substitution $\bar{p}(\bar{\theta}) = p(\theta)/m(\theta)$. This invariance is equivalent of requiring independence [3] of $\mathrm{d}P(\theta)$ on the subjectively chosen parameterization. In fact, that constraint provides a direct method of determining the Lebesgue measure: If $\theta \to \psi$, $m(\theta)\mathrm{d}\theta = m(\psi)\mathrm{d}\psi$. For instance, ignorance to change of units correspond to the transformation $\psi(\theta) = \beta\theta$, where β describes the conversion factor between different units. It yields $m(\theta)\mathrm{d}\theta = m(\psi)\mathrm{d}\psi = m(\beta\theta)\mathrm{d}(\beta\theta) = \beta m(\beta\theta)\mathrm{d}\theta$ or $m(\theta) \propto 1/\theta$.

Utilizing $m(\theta)$ the integral in Equation (1.1) is converted into a Riemann integral,

$$S = -\int_{\Omega} p(\theta) \log\left(\frac{p(\theta)}{m(\theta)}\right) \mathrm{d}\theta. \tag{1.2}$$

The optimization is subject to all known testable information $\phi_i, i = 0,1,2,\cdots,N$,

$$\phi_i \equiv \langle f_i(\theta)\rangle = \int \bar{p}(\bar{\theta})\, f_i(\bar{\theta})\mathrm{d}\bar{\theta} = \int p(\theta)\, f_i(\theta)\mathrm{d}\theta, \tag{1.3}$$

$$f_0 = 1, \phi_0 = 1. \tag{1.4}$$

The functions f_k will for convenience be denoted *test functions*. The mandatory zeroth constraint ϕ_0 (Equation (1.4)) is the normalization condition. As pointed out [3], for discrete sample spaces and no degeneracy (m =*const*), there is a general expression for the maximum of Equation (1.1) in terms of a partition function Z. It can be generalized to non-constant measures $m(\theta)$, and continuous distributions using calculus of variations [7],

$$Z = \int_{\Omega} m(\theta) \mathrm{e}^{-\sum_{i\geq 1} \lambda_i f_i(\theta)} \mathrm{d}\theta. \tag{1.5}$$

The Lagrange multipliers of optimization are implicitly given by the testable information,

$$-\frac{\partial}{\partial \lambda_i} \log Z + \phi_i = 0, \; i = 1,2,\cdots,N, \tag{1.6}$$

$$\lambda_0 = \log Z. \tag{1.7}$$

The maximum entropy pdf is then given by,

$$p(\theta) = m(\theta) \cdot \mathrm{e}^{-\sum_{i=0}^{N} \lambda_i f_i(\theta)}, \theta \in \Omega. \tag{1.8}$$

2.1. Testable Information

The PME solution (Equation (1.8)) does not specify the test functions $f_k, k \geq 1$. They are of major importance though since the solution is directly expressed in them. As stated in the introduction there is a convention of setting $f_1 = \theta$, $f_2 = \left(\theta - \langle\theta\rangle\right)^2$ etc. That is a habit, not a prescription. The choice of test functions must consequently be considered to be at our free disposal.

The difficulty or accuracy of estimating $\langle f_k(\theta)\rangle$ is

[1]Remark: "Ignorance" will here refer to the aspect of counting outcomes [1], not the concept of entropy [3].

dependent on the explicit form of f_k. Also, the information contained in the observation set $\{\theta_k\}$ is to a variable degree transferred to the estimate of $\langle f_k(\theta)\rangle$. For instance, if $f(\theta)=\theta$ all observations are equally weighted, but if $f_q = g_q$,

$$g_q(\theta) \equiv (\delta\theta)^q,\ \delta\theta \equiv \theta - \langle\theta\rangle, \tag{1.9}$$

the exponent q determines how much different observations contribute. In the asymptotic limit $q\to\infty$, the observation with the largest deviation is much more important than any other (which only contributes to the estimate of the mean). A lot of information is obviously disregarded. The estimator covariance will accordingly be large. Nevertheless, in many situations the range [6] $\lim_{k\to\infty}\langle g_{2k}(\theta)\rangle^{(1/2k)} = \max_k|\theta-\langle\theta\rangle|$ is of much larger interest than any other information. The confidence interval is a more general statistic allowing for sample spaces without bound. From the perspective of objectivity, it thus appears difficult to prescribe any specific set $f_k, k=1,2,\cdots,N$, or even their number (*N*). Clearly, the larger amount of data or independent information that is available, the larger N is allowed without resulting in unacceptable estimator quality, as expressed with its bias and covariance.

To illustrate how subjectivity enters in practice, assume we have gathered a set of prior observations $\{\theta_k\}$ of a phenomenological constant θ contained in a computer model. To calibrate the model [8], *i.e.* determine the optimal values of its parameters and their uncertainties (θ being one of them) from observations, Bayesian estimation is applied. That requires complete prior information, *i.e.* knowledge of the prior pdf $p(\theta)$. Applying PME starting from $\{\theta_k\}$, test functions $f_k(\theta)$ must be selected. With the choice of $f_1 = g_1(\theta)$, $f_2 = g_q(\theta)$, the partition function will according to Equation (1.5) read,

$$Z = \int_{-\infty}^{\infty} m(\theta)\mathrm{e}^{-\lambda_1 g_1(\theta)-\lambda_2 g_q(\theta)}\mathrm{d}\theta.$$

Since we are not aware of any ignorance, we set $m(\theta)=1$. The condition Equation (1.6) on $g_1(\theta)$ implies,

$$0 = Z^{-1}\int_\Omega \delta\theta\,\mathrm{e}^{-\lambda_1\delta\theta-\lambda_2(\delta\theta)^q}\mathrm{d}\theta.$$

A fair amount of subjective pragmatism now suggests q to be limited to even numbers and the support Ω to extend to the whole real axis, since that allows us to determine λ_1 with ease. The symmetry of the integrand then implies $\lambda_1 = 0$. If $\theta<0$ is prohibited, $\lambda_1\neq 0$. The current assignment then translates into an approximation of the support Ω as well as the factor $\mathrm{e}^{-\lambda_1\cdot(\theta-\langle\theta\rangle)}\approx 1$ of $p(\theta)$. The remaining Lagrange multiplier λ_2 is obtained by re-scaling,

$$Z = \int_{-\infty}^{\infty}\mathrm{e}^{-\lambda_2(\delta\theta)^q}\mathrm{d}\theta = \lambda_2^{-1/q}\int_{-\infty}^{\infty}\mathrm{e}^{-t^q}\mathrm{d}t,$$

$$\langle g_q(\theta)\rangle = -\frac{\partial\log Z}{\partial\lambda_2} = \frac{1}{\lambda_2 q}.$$

The resulting pdf,

$$p(\theta) \propto \mathrm{e}^{-g_q(\theta/\Delta)/q},\ \Delta \equiv \langle g_q(\theta)\rangle^{1/q}, \tag{1.10}$$

where Δ describes the q-th order "width" of the pdf $p(\theta)$, is for different q displayed in **Figure 1**. Clearly, the PME pdf is to a significant extent controlled by our subjective choice of test function, *i.e.* q. The result varies between normal $(q=2)$ and uniform $(q=\infty)$. More generally, with the understanding of the dependence on the test functions f_i (Equation (1.8)), almost any PME pdf can be generated. It is only a matter of formulating the question (selecting f_i) to obtain the answer $(p(\theta))$ on virtually any desired form.

2.2. Quality of Estimation

In practice, a testable piece of information ϕ_i is estimated from raw observations with quality depending on the test function f_i. The maximum entropy pdf $p(\theta)$ can be directly formulated in our observations $\{\theta^{(k)}\}$, using a specific estimator of the q-th moment around the mean. An obvious estimator is given by,

$$\hat{S}^{(q)} = B^{(q)}\sum_{k=1}^{n}\left(\theta^{(k)} - \frac{1}{n}\sum_{k=1}^{n}\theta^{(k)}\right)^q, \tag{1.11}$$

where $B^{(q)}$ is an unknown constant which hopefully

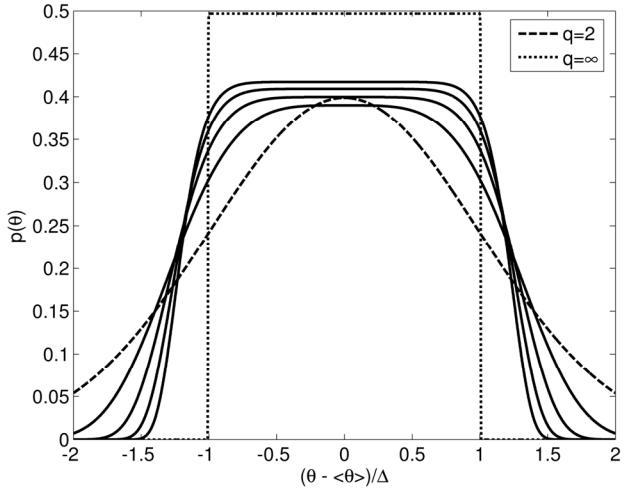

Figure 1. The maximum entropy pdf $p(\theta)$ for different kinds $(q=2,4,6,8,10,\infty)$ of testable information $\langle f(\theta)\rangle$, see Equation (1.3). Case *q* corresponds to knowledge of the mean $(f=\theta)$ and the *q*-th statistical moment around the mean ($f=g_q(\theta)$, see Equation (1.9)).

can be selected to eliminate the bias, e.g. $B^{(2)}=1/(n-1)$. To express the bias of $\hat{S}^{(q)}$ in terms of $B^{(q)}$, expand and calculate its expectation over observations $\{\theta^{(k)}\}$,

$$\left\langle \hat{S}^{(q)} \right\rangle = B^{(q)} n \sum_{\sum_{j=0}^{n} k_j = q} \frac{q!}{\prod_{j=0}^{n} k_j!} \left(-\frac{1}{n}\right)^{q-k_0} \times \left\langle \theta^{k_0+k_1} \right\rangle \prod_{j=2}^{n} \left\langle \theta^{k_j} \right\rangle . \tag{1.12}$$

On the other hand,

$$\left\langle S^{(q)} \right\rangle = \left\langle (\delta\theta)^q \right\rangle = \sum_{k=0}^{q} \frac{(-1)^{q-k} q!}{k!(q-k)!} \left\langle \theta^k \right\rangle \left\langle \theta \right\rangle^{q-k} . \tag{1.13}$$

Clearly, the number of terms of $\left\langle S^{(q)} \right\rangle$ is much less than for $\left\langle \hat{S}^{(q)} \right\rangle$ for $q>3$. Elimination of bias by proper selection of $B^{(q)}$ requires proportionality between $\left\langle S^{(q)} \right\rangle$ and $\left\langle \hat{S}^{(q)} \right\rangle$. That cannot in general be achieved since one of them have much more terms than the other. Thus, no scaling of $\left\langle \hat{S}^{(q)} \right\rangle$ can make it universally unbiased. Evaluating the first two coefficients of the expansion in Equation (1.12),

$$\left\langle \hat{S}^{(q)} \right\rangle \equiv B^{(q)} n \left(c_q^{(q)} \left\langle \theta^q \right\rangle - q \cdot c_{q-1}^{(q)} \left\langle \theta^{q-1} \right\rangle \left\langle \theta \right\rangle \cdots \right)$$

$$c_q^{(q)} = c_{q-1}^{(q)} = \sum_{k=0}^{q-2} \frac{q!}{k!(q-k)!} \left(-\frac{1}{n}\right)^k + (q-1) \cdot \left(-\frac{1}{n}\right)^{q-1} . \tag{1.14}$$

Surprisingly, these two terms satisfies the proportionality required to eliminate the bias. Bias can thus be eliminated by rescaling up to $q=3$, but not for $q \geq 4$ (see estimation of kurtosis in [9]). This suggests the normalization,

$$B^{(q)} \to \left\{ n \cdot \left[\sum_{k=0}^{q-2} \frac{q!}{k!(q-k)!} \left(-\frac{1}{n}\right)^k + (q-1) \cdot \left(-\frac{1}{n}\right)^{q-1} \right] \right\}^{-1} . \tag{1.15}$$

While $B^{(2)}=1/(n-1)$, $B^{(4)}=1/\left(n-4+6/n^2-3/n^3\right)$. Clearly, $B^{(4)}$ bears little resemblance to the conventional form $1/(n-z)$ of $B^{(2)}$, where z is the number of degrees of freedom.

After failing to obtain an unbiased estimator of the q-th moment around an unknown mean, we may lower the ambition by assuming the mean $\langle\theta\rangle$ is known. The corresponding estimator reads,

$$\hat{T}^{(q)} = B^{(q)} \sum_{k=1}^{n} \left(\theta_k - \langle\theta\rangle\right)^q , \tag{1.16}$$

This is a considerably simpler situation for which the normalization $B^{(q)}=1/n$ makes the estimator unbiased for all q. The expected precision of any estimator $\hat{M}$ describing its typical relative variation may be defined by,

$$\Upsilon\left(\hat{M}\right) \equiv \frac{\sqrt{\mathrm{var}\left(\hat{M}\right)}}{\left\langle \hat{M} \right\rangle} . \tag{1.17}$$

The least possible variance of any unbiased estimator $\tilde{S}^{(q)}$ of $\left\langle g_q(\theta) \right\rangle$ is given by the Cramer-Rao lower bound [2] $\Lambda^{(q)}$, evaluated in Appendix A,

$$\Upsilon\left(\tilde{S}^{(q)}\right) \geq \frac{q}{\sqrt{2n}} \equiv \Lambda^{(q)} . \tag{1.18}$$

Since $\Lambda^{(q)}$ increases with q, it is indeed more difficult to determine higher order moments accurately on an absolute scale, with *any* estimator. The efficiency η of our specific estimator $\hat{T}^{(q)}$ measures its relative quality,

$$\eta^{(q)} \equiv \frac{\Lambda^{(q)}}{\Upsilon\left(\hat{T}^{(q)}\right)} = \frac{q(q-1)!!}{\sqrt{2\left[(2q-1)!!-(q-1)!!\right]}} \leq 1, \tag{1.19}$$

where $m!!=(m)(m-2)(m-4)\cdots 1$, q is even and the precision $\Upsilon\left(\hat{T}^{(q)}\right)$ is evaluated in Appendix B. For an *efficient* [2] estimator, $\eta=1$. A low value of η means that the potential for improvement of our estimator is large. Since $\eta^{(q)}$ decreases rapidly with q, $\hat{T}^{(q)}$ also becomes worse on a relative scale as q increases. Nevertheless, the actual precision Υ may or may not be satisfactory. The performance of the estimators $\hat{S},\hat{T}$ is illustrated and compared to the Cramer-Rao lower bound in **Figure 2**, by evaluating their relative variation Υ and bias β numerically with multiple Monte Carlo ensembles.

To conclude, the determination of the PME pdf involved several subjective choices of processing the observations. The shape was here restricted by using monomial test functions (g_q). The order q was limited to be even, to simplify the calculation of the integrals. Likewise, an infinite symmetric sample space Ω was assigned. The difficulty of estimating various statistical moments around the mean increases rapidly with the order. If there are not more than a few samples it is in most cases impossible to reliably estimate any other moments than the first and the second. The application of PME thus relied upon rational selection rather than objective deduction.

2.3. Prior vs Posterior Application of PME

The PME constructs complete pdfs from incomplete testable information. It can be applied to the input (PME prior), or the output (PME post(-erior)) of the analysis. As no restriction or preference is stated in PME, both alternatives deserves to be considered and evaluated for efficiency and accuracy. The default method of linearization (LIN) [10] as well as the unscented Kalman filter (UKF) [11,12] and deterministic sampling (DS) [6]

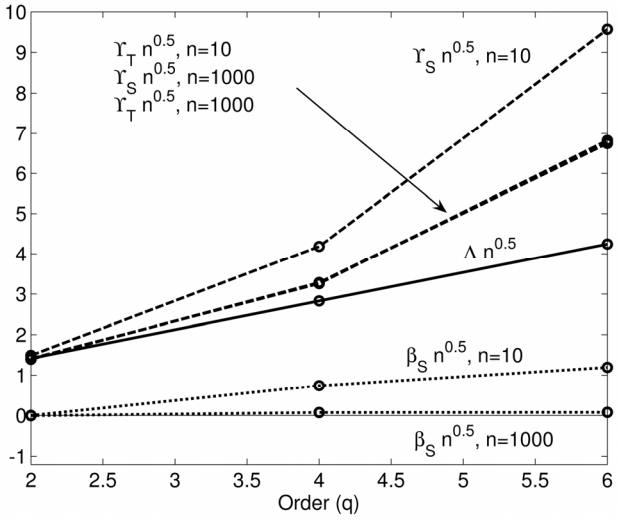

Figure 2. The relative precision Υ_s and Υ_T (dashed, Equation (1.17)) for the estimators $\hat{S}^{(q)}$ (Equation (1.11)) and $\hat{T}^{(q)}$ (Equation (1.16)), respectively, compared to the Cramer-Rao lower bound Λ (solid, Equation (1.18)), for $q = 2,4,6$ (connected by lines for clarity) and $n = 10,\ 1000$ samples. The relative bias $\beta^{(q)}$ of $\hat{S}^{(q)}$ (dotted, $B^{(q)}$ given by Equation (1.15)) is also shown. The calculations were made for 10^7 randomly generated samples of a normal distributed variable, equally split into ensembles of n samples. Scaling with $n^{0.5}$ account for the main dependences on the number n of samples (Equation (1.18)).

in general propagate statistics like covariance which later can be expanded to information related to entire pdfs, e.g. confidence intervals. That is, LIN, UKF, and DS promotes PME post. Monte Carlo simulations [13] or random sampling (RS) must start with complete statistical information to determine its random generator, *i.e.* RS requires PME prior.

An elementary example illustrates the principal differences between PME prior and PME post. Assume it is of interest to calculate the 95% confidence interval $[h_a, h_b]$ of an uncertain model $h(\theta)$, dependent on one parameter θ. Let the model uncertainty [8] be given by the mean $\langle\theta\rangle = 0$ and the variance $\langle(\delta\theta)^2\rangle = 1$. PME prior determines the input pdf $f(\theta)$ to be normal distributed. Repeated generation of RS ensembles estimates their variation as function of their size: To achieve a standard deviation of the input 95% confidence interval $[\theta_a, \theta_b]$ for θ of about 5%, no less than around 3000 randomly generated samples of θ are required. In RS, the model is evaluated for all these samples of θ and $[h_a, h_b]$ is found by evaluating the results $h(\theta^{(k)}), k = 1,2,\cdots,3000$, ordering them, and extracting the 95% percentiles. The known statistics $\langle\theta\rangle = 0$ and $\langle(\delta\theta)^2\rangle = 1$ can however be encoded *exactly* in no more than two(!) deterministic (calculated with a rule) samples $\theta^{(1)} = -1, \theta^{(2)} = 1$. Assuming the model is close to linear-in-parameters, the model variance is mainly determined by the parameter variance, here represented by the parameter ensemble $\theta^{(1)}, \theta^{(2)}$. In DS, the model variance is then with the argument of consistency given by the second moment around the mean of the model ensemble $h(\theta^{(k)}), k = 1,2$, *i.e.*

$$\langle(\delta h)^2\rangle \equiv \left\langle\left(h - \langle h\rangle\right)^2\right\rangle = \left[(\delta h)^2(\theta^{(1)}) + (\delta h)^2(\theta^{(2)})\right]/2.$$

Applying PME post to the result $\langle h\rangle$ and $\langle(\delta h)^2\rangle$ implies that the resulting pdf is normal distributed, analogously to the input pdf in RS. That implies a coverage factor of $k = 1.96$ resulting in a confidence interval,

$$[h_a, h_b] = \langle h\rangle + k\sqrt{\frac{(\delta h)^2(\theta^{(1)}) + (\delta h)^2(\theta^{(2)})}{2}}[-1,1]. \quad (1.20)$$

For estimating confidence intervals, PME does not in fact need to be utilized at all. By using another DS technique of ours, sampling on confidence boundaries [5], the desired interval can be found directly as,

$$[h_a, h_b] = \left[h\left(\langle\theta\rangle - k\sqrt{\langle(\delta\theta)^2\rangle}\right), h\left(\langle\theta\rangle + k\sqrt{\langle(\delta\theta)^2\rangle}\right)\right]. \quad (1.21)$$

Consistency in evaluating confidence intervals of models here lead to the related concept of confidence boundaries in parameter space. For multivariate models with non-linear dependencies on parameters, both these results needs to be properly generalized [5,6]. Propagating a PME prior probability density *function* thus typically require several thousands of model samples [13,14] or complex algorithms [15]. The *number* of model samples propagated with DS for final completion with PME post can be as few as $n+1$ for n parameters and increases with the known statistical information, the complexity of the model and acceptable accuracy of evaluation [16].

PME prior requires much more information to be analyzed, than PME post. For the example above, it means 3000 or 2 model evaluations. Plain rationality thus strongly favors PME post. The reason for requesting complete prior information is likely an ambition to find unique unquestionable answers, even if that in practice requires an extensive amount of blind assignments or assumptions. Unique is not equivalent to accurate however, and the quality of any assignment should be critically judged. The loss in efficiency of PME prior compared to PME post is not compensated by superior accuracy. In a number of cases, DS produces the correct result without any error: Any moment of any model

given by a finite Taylor expansion can be exactly calculated with stratified DS [16]. Similarly, the confidence interval of any model $h(\theta)=\eta(\beta^{\mathrm{T}}\theta)$ of any number (n) of parameters can be evaluated exactly [5], where $\beta\equiv(\beta_1\ \ \beta_2\ \ \cdots\ \ \beta_{n_T})^{\mathrm{T}}$, are constants, $\theta\equiv(\theta_1\ \ \theta_2\ \ \cdots\ \ \theta_n)^{\mathrm{T}}$ all parameters, and $\eta(x)$ is any non-linear monotonic function. For comparison, we are not aware of any general analytic method to propagate a univariate pdf determined by PME prior through any non-linear model. In this case, RS is a numerical method that yield arbitrarily small errors, but at very high computational cost. For "genuinely" multivariate problems RS seizes to be accurate. Multivariate refers to non-trivial finite dependencies of any order, as required for optimal modeling [16]. As far as we know, higher order dependencies (beyond second moments and not normal distributions) can only be implemented in RS by excluding samples. Exceptionally dense sampling is then required to accurately represent sampling *density* over the entire n-dimensional sampling domain. Nevertheless, it is straight-forward to represent any arbitrary mixed moment in stratified DS [16]. The difficulty is just that the more requirements, the more samples are required. If not enough samples can be afforded it is possible to find the best approximating ensemble where different requirements are given different weights of importance. There are thus several reasons (as exemplified here) for PME post methods to not only be superior to PME prior approaches in efficiency, but also accuracy. The lean set of known input information in PME post is much simpler to analyze or propagate through any model than the complete information in PME prior.

With these observations, the traditional preference to prior instead of posterior completion of information with PME appears biased and strongly subjective. It is highly questionable if PME post methods like DS [6,11,16,17] can be claimed less accurate than state-of-the-art RS relying upon PME prior per se. As both practice *statistical sampling* once defined by Enrico Fermi [18], they are fully comparable. The choice is critical for complex models, as the low efficiency of RS easily render an impossible numerical task. Indeed, that is the main current obstacle for wide utilization of uncertainty quantification with RS.

Bayesian estimation is formulated for PME prior. The combination of the maximum likelihood function and the prior distribution makes Bayes approach superior to traditional approaches limited to the former. In practice though, two *functions* (PME prior) or two *discrete* sets of statistics (PME post) derived from observations are fused. Indeed, the common assumption of Gaussian noise and PME prior derived from the mean and covariance results in combination of covariance matrices (numbers) and not pdfs [8]. A non-trivial posterior distribution *function* requires non-Gaussian and/or correlated noise and utilization of no less than an infinite set of testable *reliable* information in the PME prior. If not so, Bayes estimation can without any loss of accuracy and relevance be made with PME post instead of PME prior, e.g. with DS instead of RS [8].

PME post (DS) makes it evident that results seldom are unique, while PME prior (RS) conceals ambiguities in more or less dubious or blind assignments in the problem set up. The indefiniteness of the result is an unavoidable consequence of starting with incomplete information, not the analysis. For PME post analyses (DS) this can easily be illustrated, for instance using different ensembles [6]. It is considerably more difficult for PME prior analyses (RS) due to their prohibiting numerical complexity. The contradictory consequence is that PME prior generally are considered more credible than PME post analyses, even if the latter are more honest and realistic as their flaws easily can be illustrated.

2.4. Towards Consistency and Rationality

The PME test functions should be selected according to the evaluation of the model, its behavior, and the quality of estimating the corresponding testable information. If we are interested in propagating the covariance of a signal processing model [6], we should consequently use a distribution which has been tested for representation of covariance. That is a *consistent* rather than an objective choice. Objectivity is an impossible target—selecting *any* method is a subjective choice. Our ignorance of $p(\theta)$ is rather reflecting relevance in perspective of our primary interest, or consistency throughout the analysis. It appears that consistency summarizes the goals of objectivity as well as our ignorance [1,3,4] and emphasizes the context. In contrast to objectivity, consistency expresses a relativity that can indeed be achieved in many situations, as well as measured, questioned and criticized.

The ubiquitous sparsity of statistical information in virtually all practical problems needs to be seriously addressed. If the realism of two approaches are comparable but their efficiency distinctively different, the choice should be based on *rationality* rather than consistency. Realism is distinct from resolution. Any complete set of assumptions will result in arbitrary high resolution, no matter how realistic the assumptions are. As the fidelity of the result never can be enhanced with blind assignments, it is questionable when statistical analyses should be made with distribution functions, rather than the testable information (statistics) these are derived from with PME. If desired, both approaches results in distribution functions. The completion of information is just made before (PME prior) or after (PME post) the analysis. The rational choice is PME post, as it propagates a

minute fraction of the statistical information used in PME prior.

3. Conclusions

The character of maximum entropy (PME) distribution functions has been discussed. There are two principal innovations of our study. The PME solutions can to some extent be controlled by how primary data is processed. The prevailing preference for applying PME to the input (prior) and not the output (posterior) of statistical analyses is difficult to justify, as the their accuracy is comparable but the latter is computationally superior. The choice appears governed by the method of analysis. Hence, subjectivity enters into the processing of data as well how the analysis is made.

A non-trivial selection enters when observations are processed into testable information. A simple example illustrated common subjective choices, giving different results for identical observations.

Redirecting the focus from the treatment of known information to the targeted evaluation of the analysis emphasizes *consistency*, rather than objectivity. When consistency is indecisive, *rationality* or efficiency of the analysis provides obvious guidance. PME can be applied to find prior as well as posterior distributions. Its unconventional posterior application deserves to be seriously considered, as the analysis involves much less statistical information and is correspondingly more effective, than the prior. Consistency and rationality thus fundamentally questions the prevailing method of completing statistical information prior to the analysis, as in e.g. Monte Carlo simulations.

Maximal consistency and rationality are indeed primary goals of all our proposed methods of deterministic sampling. For complex models, such lean and customized approaches are often required to obtain any measure of modeling quality at all (within acceptable computational time). Without assessment of quality, any (modeling) result is of no value. These aspects are thus of paramount importance to our society where complex calculations (technical, physical, econometrical etc.) are rapidly increasing due to the fast development of computers.

REFERENCES

[1] D. S. Sivia and J. Skilling, "Data Analysis—A Bayesian Tutorial," Oxford University Press, Oxford, 2006.

[2] S. M. Kay, "Fundamentals of Statistical Signal Processing, Estimation Theory," Volume 1, Prentice Hall, Upper Saddle River, 1993.

[3] E. T. Jaynes, "Information Theory and Statistical Mechanics," *The Physical Review*, Vol. 106, No. 4, 1957, pp. 620-630. http://dx.doi.org/10.1103/PhysRev.106.620

[4] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, Vol. 27, No. 3, 1948, pp. 379-423.

[5] J. P. Hessling and T. Svensson, "Propagation of Uncertainty by Sampling on Confidence Boundaries," *International Journal for Uncertainty Quantification*, Vol. 3, No. 5, 2013, pp. 421-444.

[6] J. P. Hessling, "Deterministic Sampling for Propagating Model Covariance," *SIAM/ASA Journal on Uncertainty Quantification*, Vol. 1, No. 1, 2013, pp. 297-318.

[7] G. M. Ewing, "Calculus of Variations with Applications," Dover, New York, 1985.

[8] J. P. Hessling, "Identification of Complex Models," in Review, 2013.

[9] L. Råde and B. Westergren, "Mathematics Handbook," 2 Edition, Studentlitteratur, Lund, 1990.

[10] ISO GUM, "Guide to the Expression of Uncertainty in Measurement," Technical Report, International Organisation for Standardisation, Geneva, 1995.

[11] S. Julier and J. Uhlmann, "Unscented Filtering and Nonlinear Estimation," *Proceedings IEEE*, Vol. 92, No. 3, 2004, pp. 401-422.

[12] S. Julier, J. Uhlmann and H. Durrant-Whyte, "A New Approach for Filtering Non-Linear Systems," *American Control Conference*, Seattle, 21-23 June 1995, pp. 1628-1632.

[13] R. Y. Rubenstein and D. P. Kroese, "Simulation and the Monte Carlo Method," 2 Edition, John Wiley & Sons Inc., New York, 2007.

[14] J. C. Helton and F. J. Davis, "Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems," *Reliability Engineering and System Safety*, Vol. 81, No. 1, 2003, pp. 23-69. http://dx.doi.org/10.1016/S0951-8320(03)00058-9

[15] T. Lovett, "Polynomial Chaos Simulation of Analog and Mixed-Signal Systems: Theory, Modeling Method, Application," Lambert Academic Publishing, Saarbrücken, 2006.

[16] J. P. Hessling, "Stratified Deterministic Sampling of Multivariate Statistics," in Preparation, 2013.

[17] J. P. Hessling, "Deterministic Sampling for Quantification of Modeling Uncertainty of Signals," In: F. P. G. Márquez and N. Zaman, Eds., *Digital Filters and Signal Processing*, INTECH, Rijeka, 2012.

[18] N. Metropolis, "The Beginning of the Monte Carlo Method," *Los Alamos Science Special Issue*, Vol. 15, 1987, pp. 125-130.

Appendix A Cramer-Rao Lower Bounds of Statistical Moments of a Normal Distributed Variable

Assume a random variable θ with zero mean is normal distributed, $\theta \sim N(0,\sigma)$. By integrating by parts it can be shown that $M^{(q)} = \sigma^q (q-1)!!$, where $(q-1)!! = (q-1)(q-3)(q-5)\cdots 1$ is the semi-factorial function. The probability distribution function for θ can then be expressed in $M^{(q)}$,

$$p(\theta) \sim N(0,\sigma) \propto \sqrt[q]{\frac{(q-1)!!}{M^{(q)}}} \exp\left\{-\theta^2 \frac{\left[(q-1)!!\right]^{2/q}}{2\left[M^{(q)}\right]^{2/q}}\right\}. \tag{1.22}$$

For n independent observations $\{\theta_k\}, k = 1,2,\cdots,n$, the likelihood function is given by,

$$\begin{aligned} & p\left(\{\theta_k\}\middle| M^{(q)}\right) \\ & \propto \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}} \sqrt[q]{\frac{(q-1)!!}{M^{(q)}}} \exp\left\{-\theta_k^2 \frac{\left[(q-1)!!\right]^{2/q}}{2\left[M^{(q)}\right]^{2/q}}\right\} \\ & \propto \left(\frac{(q-1)!!}{M^{(q)}}\right)^{(n/q)} \exp\left\{-\frac{\left[(q-1)!!\right]^{2/q}}{2\left[M^{(q)}\right]^{2/q}} \sum_{k=1}^{n} \theta_k^2\right\}. \end{aligned} \tag{1.23}$$

Cramer-Rao lower bound ([2]) now states that for any estimator $\hat{M}^{(q)}$ of $M^{(q)}$,

$$\mathrm{var}\left(\hat{M}^{(q)}\right) \geq F^{-1}(\theta), \tag{1.24}$$

where $F(\theta)$ is the Fisher information matrix (scalar for one parameter),

$$F(\theta) = -\left\langle \frac{\partial^2 \ln p\left(\{\theta_k\}\middle| M^{(q)}\right)}{\partial \theta^2} \right\rangle = \frac{2n}{q^2 \left(M^{(q)}\right)^2}. \tag{1.25}$$

The expected precision Υ (Equation (1.17)) of the estimator $M^{(q)}$ hence satisfies,

$$\Upsilon \equiv \frac{\sqrt{\mathrm{var}\left(\hat{M}^{(q)}\right)}}{\left\langle M^{(q)} \right\rangle} \geq \frac{q}{\sqrt{2n}}. \tag{1.26}$$

Appendix B Variance of Estimator of Statistical Moments around Given Mean

An estimator of the q-th statistical moment $M^{(q)} \equiv \left\langle \left(\theta - \langle\theta\rangle\right)^q \right\rangle$ of θ around a known mean $\langle\theta\rangle$ from a set of n independent observations $\{\theta_k\}$ is given by,

$$\hat{M}^{(q)} = B \sum_{k=1}^{n} \left(\theta_k - \langle\theta\rangle\right)^q, \tag{1.27}$$

where the normalization constant B is chose to minimize its bias $\left\langle \hat{M}^{(q)} \right\rangle - M^{(q)}$. Since $\langle\theta\rangle$ is a known constant it is trivially found that $B = 1/n$ eliminates all bias, for all values of q. Its variance is found to be,

$$\mathrm{var}\left(\hat{M}^{(q)}\right) = \left\langle \left[M^{(q)}\right]^2 \right\rangle - \left\langle M^{(q)} \right\rangle^2 = \frac{M^{(2q)}}{n} - \frac{\left[M^{(q)}\right]^2}{n}. \tag{1.28}$$

An explicitly value can be found for a normal distributed parameter, $\theta \sim N(0,\sigma)$. Then, by recursively integrating by parts it is found that $M^{(q)} = \sigma^q (q-1)!!$, where $(q-1)!! = (q-1)(q-3)(q-5)\cdots 1$ is the semi-factorial function, giving

$$\mathrm{var}\left(\hat{M}^{(q)}\right) = \sigma^q \frac{(2q-1)!! - (q-1)!!}{n}. \tag{1.29}$$