Bayes' Rule of Information

Spencer Graves
PDF Solutions, Inc.
333 West San Carlos, Suite 700
San José, CA 95126
spencerg@pdf.com

ABSTRACT

This chapter discusses a duality between the addition of random variables and the addition of information via Bayes' theorem: When adding independent random variables, variances (when they exist) add. With Bayes' theorem, defining "score" and "observed information" via derivatives of the log densities, the posterior score is the prior score plus the score from the data, and observed information similarly adds. These facts make it easier to understand and use Bayes' theorem. They also provide tools for easily deriving approximate posteriors in particular families, especially normal. Other tools can then be used to evaluate the adequacy of naive use of these approximations. Even when, for example, a normal posterior is not sufficiently accurate for direct use, it can still be used as part of an improved solution obtained via adaptive Gauss-Hermite quadrature or importance sampling in Monte Carlo integration and Markov Chain Monte Carlo, for example.

One important realm for application of these techniques is with various kinds of (extended) Kalman / Bayesian filtering following a 2-step Bayesian sequential updating

cycle of (1) updating the posterior from the previous observation to model a possible change of state before the current observation, and (2) using Bayes' theorem to combine the current prior and observation to produce an updated posterior. These tools provide easy derivations of the posterior and of approximations, especially normal approximations. Another application involves mixed effects models outside the normal linear framework. This chapter includes derivations of Bayesian exponentially weighted moving averages (EWMAs) for exponential family / exponential dispersion models including gamma-Poison, beta-binomial and Dirichlet-multinomial. Pathologies that occur with violations of standard assumptions are illustrated with an exponential-uniform model.

1. INTRODUCTION

Many tools are available for deriving and easily understanding sums of random variables. This chapter presents two comparable (dual) properties of Bayes' theorem. These results concern the "score" and the "information", where the score = the first derivative of the log(likelihood) [3], extended here to include log(prior) and log(posterior); differentiation is with respect to parameter(s) of the distribution of the observations, which are therefore the random variables of the prior and posterior. Similarly, the "observed information" = the negative of the second derivatives. With these definitions, (*a*) the posterior score is the prior score plus the score from the data, and (b) the posterior observed information is the prior information plus the information from the data. Previous Bayesian analyses have used this mathematics (e.g., [6], [7]) but without recognizing it as having sufficient general utility to merit a name like "Bayes' Rule of Information".

These tools provide relatively easy derivations of extended Kalman filter / Bayesian filtering approximations and simple Laplace / saddle point approximations for mixed models outside the normal linear case (e.g., [16], which includes software for S-Plus and R). The adequacy of these approximations can then be evaluated using techniques like importance sampling with Monte Carlo integration (including, e.g., importance weighted marginal posterior density estimation within Markov Chain Monte Carlo [5]) or in low dimensions adaptive Hermite quadrature [8], [22]. The error in the simple approximation can then be used to decide if the additional accuracy provided by the more sophisticated methods is worth the extra expense.

By defining score and observed information in this way, we get the same answer whether we process *n* observations into the posterior all at once or one at a time. We therefore focus on the power and simplicity obtainable from "keeping score with Bayes' theorem" and accumulating observed information from prior to posterior.

In Sections 2 and 3, we derive the properties of interest by factoring the joint distribution of observations **y** and parameters **x** in two ways: (predictive) × (posterior) = (observation) × (prior):

$$\begin{array}{ccccccccc} p(\mathbf{y},\mathbf{x}) & = & p(\mathbf{y}) & \times & p(\mathbf{x}\,|\,\mathbf{y}) & = & p(\mathbf{y}\,|\,\mathbf{x}) & \times & p(\mathbf{x}) \\ (\text{joint}) & = & (\text{predictive}) & \times & (\text{posterior}) & = & (\text{observation}) & \times & (\text{prior}), \end{array} \tag{1}$$

where *p*(.) = probability density of observations or parameters as indicated. In Kalman or more general Bayesian filtering applications, we want to track the evolution of the unknown or latent parameters **x** over time through their influence on the observations.

The predictive distribution does not appear in the score and information equations, but can be useful for evaluating if it is plausible to assume that **y** came from this model; if **y** seems inconsistent with that model, the posterior computation might be skipped and other action taken [36].

Beta-binomial, gamma-Poisson, and other conjugate exponential family applications appear in Section 2. In Section 4 (and the appendix), we keep score with Bayes' theorem and apply Bayes' rule of information with normal priors and posteriors. The results are specialized further to normal observations including linear regression in Section 5. Section 6 reviews the connection between Bayes' and central limit theorems. The relationships between alternative definitions of information in statistics are reviewed in Section 7, and concluding remarks appear in Section 8.

2. FACTORING JOINT PROBABILITY AND KEEPING SCORE

Taking logarithms of (1), letting $l(.)$ = $\log[p(.)]$ = the logarithm of the corresponding probability density, we get the following:

$$\begin{array}{ccccccc} l(\mathbf{y}) & + & l(\mathbf{x} \mid y) & = & l(\mathbf{y} \mid \mathbf{x}) & + & l(\mathbf{x}) \\ (\text{predictive}) & + & (\text{posterior}) & = & (\text{observation}) & + & (\text{prior}). \end{array}$$

R. A. Fisher described the first derivative of the log(density) as the "efficient score" [3], [21]. In this sense, the "score" from *n* independent observations is the sum of the scores from the individual observations, and with regular likelihood, prior and posterior, the likelihood is maximized or the posterior mode is located where the applicable score (i.e., the first derivative of the log density) "balances" at 0.

In particular, the posterior score is the prior score plus the score from the data:

$$\frac{\partial l(\mathbf{x}\mid\mathbf{y})}{\partial\mathbf{x}}=\frac{\partial l(\mathbf{y}\mid\mathbf{x})}{\partial\mathbf{x}}+\frac{\partial l(\mathbf{x})}{\partial\mathbf{x}}\ , \qquad (2)$$

As explained in the rest of this chapter, expression (2) is a powerful tool for computing Bayesian posteriors, especially when a normal distribution is an adequate approximation for both prior and posterior or when a normal distribution is used as a kernel for adaptive Hermite quadrature or for importance sampling in Monte Carlo. As a mnemonic device to make it easier to remember, it describes how to keep score with Bayes' theorem.

Before taking the second derivative, we illustrate the use of (2) in examples.

Example 1: Gamma-Poisson. Consider the gamma-Poisson conjugate pair. In this case, the gamma prior $p(\lambda)$ = $\beta^{\alpha}\lambda^{\alpha-1}e^{-\lambda\beta}/\Gamma(\alpha)$, so the prior score for λ is $\partial l(\lambda)/\partial\lambda$ = $\{[(\alpha-1)/\lambda]-\beta\}$. Meanwhile, the observation density is $p(y\mid\lambda)$ = $\lambda^{y}e^{-\lambda}/y!$, so the score of the data is $\partial l(y\mid\lambda)/\partial\lambda$ = $\{[y/\lambda]-1\}$. Whence, the posterior score is $\partial l(\lambda\mid y)/\partial\lambda$ = $\{[(\alpha_1-1)/\lambda]-\beta_1\}$, where α_1 = $\alpha + y$ and β_1 = β + 1. Since this has the same form as the prior score, the posterior is also gamma. Thus, Bayes' theorem tells us to keep score in the gamma-Poisson model by adding y to α and 1 to β.

Suppose now that we have a series of Poisson observations y_t with prior distribution for λ_t of $\Gamma(\alpha_{t|t-1},\beta_{t|t-1})$. Then keeping score with Bayes' theorem tells us that the posterior is $\Gamma(\alpha_{t|t},\beta_{t|t})$ with $\alpha_{t|t}$ = $\alpha_{t|t-1}$ + y_t and $\beta_{t|t}$ = $\beta_{t|t-1}$ + 1. Let's model a possible migration over time in $\lambda = \lambda_t$ with a discount factor θ ($0 < \theta < 1$), as $\alpha_{t+1|t}$ = $\theta\alpha_{t|t}$ and $\beta_{t+1|t}$ = $\theta\beta_{t|t}$. Thus, $\alpha_{t+1|t}$ = $\theta(\alpha_{t|t-1}+y_t)$ = θy_t + $\theta^2 y_{t-1}$ + ..., and $\beta_{t+1|t}$ = $\theta(\beta_{t|t-1}+1)$ = θ + θ^2 + ... $\cong$ $\theta/(1-\theta)$, if t = 0 is sufficiently far in the past to be irrelevant. In that case, $\beta_{t+1|t}$ is constant, and $\alpha_{t+1|t}$ = $\theta\tilde{y}_t/(1-\theta)$, where $\tilde{y}_t$ = $\theta\tilde{y}_{t-1}$ +

$(1-\theta)y_t$ = an exponentially weighted moving average (EWMA) of the observations y_t. In essence, Bayes' theorem tells us to track the gamma scale parameter α by keeping score with an EWMA. For an EWMA application with a somewhat different gamma-Poisson model, see [20].

Example 2: Beta-Binomial. Consider the beta-binomial pair with observation y ~ bin(p, m) and prior p ~ beta(α,β). The same logic as for gamma-Poisson tells us that keeping score with Bayes' theorem produces a posterior that is beta(α_1, β_1) with α_1 = α + y and $\beta_1 = \beta + m - y$. With a sequence y_t ~ bin(p_t, m_t), and prior p_t ~ beta$\left(\alpha_{t|t-1}, \beta_{t|t-1}\right)$, we keep score with $\alpha_{t+1|t}$ = $\theta\left(\alpha_{t|t-1}+y_t\right)$ and $\beta_{t+1|t}$ = $\theta\left[\beta_{t|t-1}+\left(m_t-y_t\right)\right]$. If $m_t = m$ is constant and $t = 0$ is sufficiently far in the past to be negligible, then $\alpha_{t+1|t}$ = $\theta\tilde{y}_t/(1-\theta)$, where $\tilde{y}_t$ is the EWMA of the observations as before, and $\beta_{t+1|t}$ = $\theta(m-y_t)$ + $\theta\beta_{t|t-1}$ = $\theta\left(m-\tilde{y}_t\right)/(1-\theta)$. Yousry et al. [37] discuss the use of this kind of EWMA in manufacturing.

Example 3: Conjugate Updating an Exponential Dispersion Model. Examples 1 and 2 can be generalized to an arbitrary exponential family or exponential dispersion model [15], with

$$p\left(\mathbf{y}\mid\boldsymbol{\eta},\phi\right)=\exp\left\{\phi\left[\mathbf{y}'\boldsymbol{\eta}-b\left(\boldsymbol{\eta}\right)\right]-c\left(\mathbf{y},\phi\right)\right\}, \tag{3}$$

for some $\phi > 0$. The multinomial distribution with $(k+1)$ categories can be written in this form, with the k-vector $\boldsymbol{\eta}$ being the logistic transformation of the probabilities, so p_i = $\exp\left(\eta_i\right)/\left\{1-\sum\exp\left(\eta_i\right)\right\}$, and with $\phi\mathbf{y}$ being nonnegative integers whose sum never exceeds another integer N.

For this distribution, consider a conjugate prior, CP($\boldsymbol{\alpha}$, s), on the natural parameter $\boldsymbol{\eta}$ with density

$$p(\boldsymbol{\eta}) = \exp\{s[\boldsymbol{\alpha}'\boldsymbol{\eta} - b(\boldsymbol{\eta})] - d(\boldsymbol{\alpha}, s)\}, \tag{4}$$

where $b(\boldsymbol{\eta})$ is the same as in (3), and $s>0$ and $\boldsymbol{\alpha}$ are known.

The gamma-Poisson model of Example 1 can be written in the form (3)-(4) with $\eta = \log(\lambda)$. The beta-binomial of Example 2 can also be expressed in this form with $\eta = \log[p/(1-p)]$. If the two possible outcomes of the beta-binomial are further subdivided binomially to $(k+1) > 2$ possible outcomes, we get a Dirichlet-multinomial model.

The "scores" required for (2) are simple:

$$d\,l(\mathbf{y} \mid \boldsymbol{\eta})/d\,\boldsymbol{\eta} = \phi[\mathbf{y} - db(\boldsymbol{\eta})/d\,\boldsymbol{\eta}],$$

and (5)

$$d\,l(\boldsymbol{\eta})/d\,\boldsymbol{\eta} = s[\boldsymbol{\alpha} - db(\boldsymbol{\eta})/d\,\boldsymbol{\eta}].$$

Then the posterior score is

$$d\,l(\boldsymbol{\eta} \mid y)/d\,\boldsymbol{\eta} = (s\boldsymbol{\alpha} + \phi\mathbf{y}) - (s+\phi)db(\boldsymbol{\eta})/d\,\boldsymbol{\eta}.$$

If we know from other sources that CP($\boldsymbol{\alpha}$, s) is conjugate for (3), this score equation gives us the values of the parameters of that conjugate posterior CP($\boldsymbol{\alpha}_1$, β_1), where

$$\boldsymbol{\alpha}_1 = \boldsymbol{\alpha} + \kappa(\mathbf{y} - \boldsymbol{\alpha}) \text{ with } \kappa = \phi/(s+\phi),$$

and (6)

$$s_1 = s + \phi.$$

For an exponential family with a conjugate prior that can be written in the form (3)-(4), these results can be obtained from standard exponential family properties without "keeping score" in this way. Specifically, the product of (3) and (4) gives us the joint distribution, also in exponential family form:

$$p(\mathbf{y}\,|\,\boldsymbol{\eta})\,p(\boldsymbol{\eta}) = \exp\left\{(s\boldsymbol{\alpha}+\phi\mathbf{y})'\boldsymbol{\eta} - (s+\phi)b(\boldsymbol{\eta}) - c(\mathbf{y},\phi) - d(\boldsymbol{\alpha},s)\right\} \; . \tag{7}$$

Since the prior density (4) must integrate to 1 for any $s>0$ and $\boldsymbol{\alpha}$, it must also integrate to 1 for $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}+\kappa(\mathbf{y}-\boldsymbol{\alpha})$ and $s_1 = s+\phi$. This property allows us to easily integrate out $\boldsymbol{\eta}$ to get the predictive distribution:

$$p(\mathbf{y}) = \exp\{d(s\boldsymbol{\alpha}+\phi\mathbf{y}, s+\phi) - d(\boldsymbol{\alpha},s) - c(\mathbf{y},\phi)\}.$$

or (8)

$$p(\mathbf{y}) = \exp\{d(\boldsymbol{\alpha}_1, s_1) - d(\boldsymbol{\alpha},s) - c(\mathbf{y},\phi)\}.$$

This predictive distribution can be used to evaluate the consistency of each new observation with this model. New observations that seem implausible relative to this predictive distribution (8) should trigger further study to determine if these observations (*a*) might suggest improvements to the model or to the data collection methodology or (*b*) are honest rare events that deserve to be incorporated into the posterior with other observations or (*c*) are outliers that should not be incorporated into the posterior.

The standard application of Bayes' theorem in this context proceeds by dividing the joint density (7) by this predictive density $p(\mathbf{y})$ to get a posterior of the form (4) with parameters (6). However, if we use anything other than a conjugate prior like (4), the posterior might not be obtained so easily. It is precisely for such situations that more general tools like keeping score using (2) are most useful; see also [16].

Before leaving this example, suppose we have a series of observations $\mathbf{y}_t$ with density (3) and prior CP($\boldsymbol{\alpha}_{t|t-1}$, $s_{t|t-1}$). Then the posterior is CP($\boldsymbol{\alpha}_{t|t}$, $s_{t|t}$), where

$$\boldsymbol{\alpha}_{t|t} = \boldsymbol{\alpha}_{t|t-1} + \kappa_t\left(\mathbf{y}_t - \boldsymbol{\alpha}_{t|t-1}\right), \tag{8.5}$$

with $\kappa_t = \phi/\left(s_{t|t-1}+\phi\right)$ and $s_{t|t} = s_{t|t-1}+\phi$ (with ϕ constant). Similar to examples 1 and 2, we model a possible change in $\boldsymbol{\eta}_t$ between the current and the next observations with a discount factor θ on s:

$$s_{t+1|t} = \theta s_{t|t} = \theta\left(s_{t|t-1}+\phi\right) = \theta\left[\phi+\theta\left(s_{t-1|t-2}+\phi\right)\right] .$$

Moreover, if t = 0 is sufficiently remote to be negligible, we substitute this expression into itself repeatedly to get $s_{t+1|t} \cong \phi\theta/(1-\theta) = s_+$, say, which makes it essentially constant over time. This gives us the following:

$$\boldsymbol{\alpha}_{t+1|t} = \boldsymbol{\alpha}_{t|t-1}+\kappa\left(\mathbf{y}_t-\boldsymbol{\alpha}_{t|t-1}\right),$$

where

$$\kappa = 1-\theta . \tag{9}$$

In sum, a standard EWMA of random variable $\mathbf{y}_t$ of an exponential family (3) estimates the prior location parameter $\boldsymbol{\alpha}_t$ of a standard conjugate prior (4) of the location $\boldsymbol{\eta}_t$ of $\mathbf{y}_t$. as $\boldsymbol{\eta}_t$ evolves over time as modeled by the discount factor θ on the prior information s per (8.5). This provides a deeper understanding of the gamma-Poison and beta-binomial models of Examples 1 and 2.

This exponential family EWMA has been discussed, applied, and generalized by West and Harrison [36, sec. 14.2], Grigg and Spiegelhalter [14], Klein [16] and others. We will interpret κ in (9) using "Bayes' rule of Information" in the next section. Before that, however, we note that this exponential family EWMA can be applied in a quasi-likelihood context [21], assuming only that (4) with parameter values (6) provides a reasonable approximation to the posterior. We could check the adequacy of these assumptions using Markov Chain Monte Carlo (MCMC) with a sample of such data. This could be quite valuable in engineering applications where MCMC might be used during engineering design to evaluate whether a much cheaper EWMA would be adequate for routine use where MCMC would not be feasible.

3. BAYES' RULE OF INFORMATION

We return now to (2) and take another derivative to get the following:

$$\frac{\partial^2 l(\mathbf{x} \mid \mathbf{y})}{\partial \mathbf{x}\, \partial \mathbf{x}'} = \frac{\partial^2 l(\mathbf{y} \mid \mathbf{x})}{\partial \mathbf{x}\, \partial \mathbf{x}'} + \frac{\partial^2 l(\mathbf{x})}{\partial \mathbf{x}\, \partial \mathbf{x}'} \,. \tag{10}$$

In this article, we let **J**(.) denote the *observed information*, which we define here as the negative of the matrices of second partials in (10). Then (10) becomes

$$\begin{array}{ccccc} \mathbf{J}(\mathbf{x} \mid \mathbf{y}) & = & \mathbf{J}(\mathbf{y} \mid \mathbf{x}) & + & \mathbf{J}(\mathbf{x}) \\ \begin{pmatrix} \text{posterior} \\ \text{information} \end{pmatrix} & = & \begin{pmatrix} \text{information from} \\ \text{observation(s)} \end{pmatrix} & + & \begin{pmatrix} \text{prior} \\ \text{information} \end{pmatrix} . \end{array} \tag{11}$$

We call this "Bayes' Rule of Information", as it quantifies in many applications the accumulation of information via Bayes' theorem. If **y** ~ $N_k(\mathbf{x}, \boldsymbol{\Sigma}_y)$, we get **J**(**y**|**x**) = $\boldsymbol{\Sigma}_y^{-1}$. Since **J**(**y**|**x**) is constant independent of **x** in this case, it is also the Fisher (expected) information, though that is not true in other applications. Similarly, with a prior **x** ~ $N_k(\boldsymbol{\theta}, \boldsymbol{\Sigma}_x)$, we have **J**(**x**) = $\boldsymbol{\Sigma}_x^{-1}$. Then (11) tells us that **J**(**x**|**y**) = $\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{\Sigma}_y^{-1}$. Since we know from other arguments that the posterior is also normal, this gives us the posterior variance in the form of its inverse, the "information".

In the normal case, the information terms in (11) are also called precision parameters [4], being the inverse of variances (or covariance matrices); this case is considered further in Section 5. In Section 4, we assume that the prior is normal and the observed information can be adequately approximated by a constant in **x**, though it may depend on the observation **y**; this will support using a normal approximation for the posterior.

With non-normal observations, the information may not be approximately constant. In extreme examples, the observed distribution may even be multimodal. In such cases, the information from the observation(s) $[-\partial^2 l(\mathbf{y}\,|\,\mathbf{x})/\partial\mathbf{x}\partial\mathbf{x}']$ can even have negative eigenvalues in a certain region between modes. Fortunately, many such examples are still sufficiently regular that standard results can be used to show that observations with indefinite or even negative definite information are so rare that their impact on the posterior vanishes almost surely as more data are collected. If this is not adequate, we could handle mixtures by computing the posterior as a mixture and then deleting components with negligible posterior mixing probabilities as suggested by West and Harrison [36, ch. 12]. (For more on finite mixtures, see [35] and [24].)

Example 3 (cont.): EWMA for Exponential Dispersion Data. What does "Bayes' Rule of Information" tell us about processing data from a (possibly overdispersed) generalized linear model (3) with a conjugate prior (4)? To find out, we differentiate (5):

$$\mathbf{J}(\mathbf{y}\,|\,\boldsymbol{\eta}) = \phi\left[\frac{d^2 b(\boldsymbol{\eta})}{d\boldsymbol{\eta}\,d\boldsymbol{\eta}'}\right], \text{ and } \mathbf{J}(\boldsymbol{\eta}) = s\left[\frac{d^2 b(\boldsymbol{\eta})}{d\boldsymbol{\eta}\,d\boldsymbol{\eta}'}\right]. \tag{12}$$

To help build our intuition about this, we use dimensional analysis assuming **y** has "*y* units", and **η** has "η units". Then *b*(**η**) has $(y\eta)$ units. If the exponent in (3) is dimensionless, ϕ must have $(y\eta)^{-1}$ units. For a normal distribution, "η units" are "*y* units", so ϕ has y^{-2} units. For a Poisson distribution, **y** is counts of events, and **η** is in log(counts). Then ϕ can be said to have $(\text{count}\times\log(\text{count}))^{-1}$ units, though counts and log(counts) could also be considered dimensionless themselves. A similar analysis applies to binomial or multinomial observations, where $\boldsymbol{\eta}$ is in logits and $\mathbf{y}$ is either counts or proportions; in the latter case, ϕ is in $\left(\text{counts}\times\text{logits}\right)^{-1}$ [or in $\left(\text{counts}\right)^{-1}$ if logits are considered dimensionless].

This tells is that $d\,b(\boldsymbol{\eta})/d\,\boldsymbol{\eta}$ has "*y* units", which it must have since a standard exponential family property makes $E\mathbf{y}$ = $d\,b(\boldsymbol{\eta})/d\,\boldsymbol{\eta}$. Similarly, $d^2b(\boldsymbol{\eta})/(d\boldsymbol{\eta}\,d\boldsymbol{\eta}')$ has $\left(y\eta^{-1}\right)$ units. Then by (12), $\mathbf{J}\left(\mathbf{y}\,|\,\boldsymbol{\eta}\right)$ has η^{-2} units, which it must have, because the inverse of observed and Fisher information is variance (of $\boldsymbol{\eta}$ in this case). Note also that another standard exponential family property has

$$\operatorname{var}\left(\mathbf{y}\,|\,\boldsymbol{\eta}\right)=\phi^{-1}\left[\frac{d^2b\left(\boldsymbol{\eta}\right)}{d\boldsymbol{\eta}\,d\boldsymbol{\eta}'}\right].$$

This is the same as $\mathbf{J}\left(\mathbf{y}\,|\,\boldsymbol{\eta}\right)$ except that the scale factor ϕ is inverted, which change the units from η^{-2} to y^2, as required for $\operatorname{var}\left(\mathbf{y}\,|\,\boldsymbol{\eta}\right)$.

In Section 4, we will assume that the posterior information is always positive (or nonnegative definite) and can be adequately approximated by a constant in a region of sufficiently high probability near the posterior mode. In this case, with a normal prior, a normal posterior also becomes a reasonable approximation. Before turning to that common case, we first illustrate pathologies possible with irregular likelihood when the range of support depends on a parameter of interest.

Example 4. Exponential - Uniform. Pathologies with likelihood often arise with applications where the range of support of a distribution involves parameter(s) of interest. For example, consider y ~ Uniform(0, e^{γ}). We take as a prior for γ a 2-parameter exponential with mean ν^{-1} and support on $\left(\gamma_0,\infty\right)$; this is equivalent to the Pareto prior for e^{γ} considered by Rossman, Short and Parks [30]. We denote this by Exp(ν^{-1}, γ_0); its density is as follows:

s,

where *I*(*A*) is the indicator function of the event *A*. Then the log(density) is as follows:

$$l(\gamma) = \ln(\nu) - \nu(\gamma - \gamma_0), \text{ for } (\gamma > \gamma_0). \tag{13}$$

Also, the density for *y* is as follows:

$$f(y \mid \gamma) = e^{-\gamma} I\left(0 < y < e^{\gamma}\right),$$

so

$$l(y \mid \gamma) = (-\gamma), \text{ for } \left(0 < y < e^{\gamma}\right). \tag{14}$$

Therefore, the support for the joint distribution has $\gamma >$ $\max\{\gamma_0, \ln(y)\}$. To keep score with Bayes' theorem, we need the prior score and the score from the data. We get the prior score by differentiating (13):

$$\frac{\partial l(\gamma)}{\partial \gamma} = (-\nu), \text{ for } (\gamma > \gamma_0). \tag{15}$$

For the data, by differentiating (14) we see that the score function is a constant (-1):

$$\frac{\partial l(y \mid \gamma)}{\partial \gamma} = (-1), \text{ for } \left(0 < y < e^{\gamma}\right), \text{ i.e., } \{\ln(y) < \gamma\}. \tag{16}$$

We add this to (15) to get the posterior score:

$$\frac{\partial l(\gamma \mid y)}{\partial \gamma} = (-1 - \nu) = (-\nu_1), \text{ for } (\gamma > \gamma_1 = \max\{\gamma_0, \log(y)\}\}),$$

where ν_1 = $\nu + 1$. By integrating the posterior score over $(\gamma > \gamma_1)$, the range of support for *γ*, we find that the posterior is Exp(ν_1^{-1}, γ_1). Thus, the 2-parameter exponential is a conjugate prior for the uniform distribution considered here. With repeated data collection, ν_1 increases by 1 with each observation pulling $E(\gamma) = \gamma_1 + \nu_1^{-1}$ ever closer to the lower limit γ_1. (Alternative conjugate priors for this uniform distribution include a Pareto and a truncated normal. Both exhibit pathologies similar to but different from the ones discussed here.)

To get the observed information, we differentiate (15) and (16) a second time to get

$$J(\gamma) = J(\gamma \mid y) = J(y \mid \gamma) = 0 .$$

Thus, in this example, the observed information from prior, data, and posterior are all 0. Clearly, the posterior gets sharper with additional data collection. This reflects an accumulation of knowledge, even though there is no "observed information" in anything!

The problems in this case arise because the parameter of interest defines a boundary, which means that many of the standard properties of "regular likelihood" do not hold. In this example, both prior and observation densities have a point of discontinuity, but the score and information equations (2) and (11) are still valid everywhere else.

If we change the parameterization, we get different pathologies, For example consider $y \sim U(0, b)$ [e.g., with b following a Pareto distribution]. Then the score from the data is $(-1/b)$ if $0 < y < b$, so the Fisher information defined as the variance of the score is 0. The observed information, however, is not zero; it's negative = $(-1/b^2)$! The usual equality between the Fisher information and the expected observed information assumes that the order of differentiation and expectation can be interchanged, which does not hold in this case. Fisher information may not be useful in such irregular situations, but we can still keep score and accumulate observed information using (2) and (11).

A primary area for application of Bayes' Rule of Information (11) and the companion scoring rule (2) is for Kalman filtering, especially nonlinear extended Kalman filtering and for more general Bayesian sequential updating ([36]; [26]; [13]). Such cases involve repeated applications of Bayes' theorem, where the information from the data arriving with each cycle accumulates in the posterior, summarizing all the relevant information in the data available at that time, which then with a possible transition step becomes the prior for the next cycle.

Another important area of application is for deriving importance weighting kernels for Monte Carlo integration with random effects and / or Bayesian mixed effect models outside of the normal linear paradigm. Beyond providing a first order approximation, which may not be adequate, they provide a tool for handling relatively easily the "curse of dimensionality," which says roughly that almost everything is sparse in high enough dimensions. For example, Evans and Schwartz [8] note that the volume of a k-dimensional unit sphere as a proportion of the circumscribing unit cube, $[-1, 1]^k$, goes to zero as k increases without bounds. Thus, if we try to estimate the volume of this sphere via Monte Carlo sampling from a uniform distribution on $[-1, 1]^k$, we would need ever larger Monte Carlo samples as k increases just to maintain an fixed probability of getting at least one observation in this sphere!

However, if we know that most of the mass of the distribution is close to the coverage of the corresponding normal approximation, most of the k-dimensional pseudo-random normal variates we generate will also be relevant to the non-normal distribution of interest. This makes importance sampling a simple yet valuable tool for evaluating the adequacy of a normal approximation and for improving upon it when it is not adequate.

4. NORMAL PRIOR AND POSTERIOR

We assume in this and the next sections that the prior and posterior are both adequately approximated by normal distributions, $N_p(\mathbf{x}_0, \mathbf{\Sigma}_0)$ and $N_p(\mathbf{x}_1, \mathbf{\Sigma}_1)$, respectively. Then

$$l(\mathbf{x}) = c_0 - \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)' \mathbf{\Sigma}_0^{-1}(\mathbf{x} - \mathbf{x}_0),$$

and

$$l(\mathbf{x} \mid \mathbf{y}) = c_1 - \frac{1}{2}(\mathbf{x} - \mathbf{x}_1)' \mathbf{\Sigma}_1^{-1}(\mathbf{x} - \mathbf{x}_1),$$

where c_0 and c_1 are appropriate constants (relative to **x**). We'd like to use (11) to compute $\mathbf{\Sigma}_1$ and (2) to get $\mathbf{x}_1$. For this, we need following:

$$\frac{\partial l(\mathbf{x})}{\partial \mathbf{x}} = \left[-\mathbf{\Sigma}_0^{-1}(\mathbf{x} - \mathbf{x}_0)\right]; \quad \frac{\partial l(\mathbf{x} \mid \mathbf{y})}{\partial \mathbf{x}} = \left[-\mathbf{\Sigma}_1^{-1}(\mathbf{x} - \mathbf{x}_1)\right], \tag{17}$$

and

$$\mathbf{J}(\mathbf{x}) = \left[-\frac{\partial^2 l(\mathbf{x})}{\partial \mathbf{x}\, \partial \mathbf{x}'}\right] = \mathbf{\Sigma}_0^{-1}; \quad \mathbf{J}(\mathbf{x} \mid \mathbf{y}) = \left[-\frac{\partial^2 l(\mathbf{x} \mid \mathbf{y})}{\partial \mathbf{x}\, \partial \mathbf{x}'}\right] = \mathbf{\Sigma}_1^{-1}. \tag{18}$$

To keep things simple, we substitute (18) into (11) evaluating $\mathbf{J}(\mathbf{y} \mid \mathbf{x})$ at the prior mode **x** = $\mathbf{x}_0$ to get the following (provided only that the likelihood for **y** is regular):

$$\mathbf{\Sigma}_1^{-1} = \mathbf{J}(\mathbf{y} \mid \mathbf{x} = \mathbf{x}_0) + \mathbf{\Sigma}_0^{-1}. \tag{19}$$

We assume in this section that variations in $\mathbf{J}(\mathbf{y} \mid \mathbf{x})$ are so small that a normal approximation with mean at the posterior mode $\mathbf{x}_1$ and "information" $\mathbf{\Sigma}_1^{-1}$ per (19) provides an adequate approximation to the posterior. If that is not appropriate, but replacing $\mathbf{x}_0$ by $\mathbf{x}_1$ in (19) would produce an adequate approximation to the posterior, then we can iterate to obtain $\mathbf{x}_1$ and $\mathbf{\Sigma}_1^{-1}$ simultaneously, as discussed in the Appendix.

If we now use (17) to compute the score (2) at the prior mode $\mathbf{x} = \mathbf{x}_0$, we get the following:

$$-\boldsymbol{\Sigma}_1^{-1}\left(\mathbf{x}_0 - \mathbf{x}_1\right) = \left[\frac{\partial l\left(\mathbf{y} \mid \mathbf{x} = \mathbf{x}_0\right)}{\partial \mathbf{x}}\right] + \mathbf{0},$$

so

$$\mathbf{x}_1 = \mathbf{x}_0 + \boldsymbol{\Sigma}_1\left[\frac{\partial l\left(\mathbf{y} \mid \mathbf{x} = \mathbf{x}_0\right)}{\partial \mathbf{x}}\right], \tag{20}$$

assuming the posterior information matrix $\boldsymbol{\Sigma}_1^{-1}$ is of full rank. Thus, when the normal distribution with information $\boldsymbol{\Sigma}_1^{-1}$ computed via (19) is an adequate approximation to the posterior, (20) provides a simple way to obtain the posterior mean $\mathbf{x}_1$. If in addition the observations are linear in **x** plus normal error, $\mathbf{J}\left(\mathbf{y} \mid \mathbf{x}\right)$ is constant in **x**, and the posterior is exactly normal, as we explain in the next section.

With a series of observations, possible changes of state between them are typically modeled by a random walk, possibly added to a deterministic change. Special consideration must be given to cases where the posterior information $\boldsymbol{\Sigma}_{t-1|t-1}^{-1}$ from the previous observation is singular; we consider this issue further in the next section.

5. NORMAL OBSERVATIONS

In this section, we first assume that **y** ~ N_p(**x**, **V**) and later that **y** ~ N_k(**Zx**, **V**). In the first case, the log(likelihood) is as follows:

$$l\left(\mathbf{y} \mid \mathbf{x}\right) = c_y - \frac{1}{2}\left(\mathbf{y} - \mathbf{x}\right)' \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{x}\right).$$

Then the score from the data is

$$\frac{\partial l\left(\mathbf{y} \mid \mathbf{x}\right)}{\partial \mathbf{x}} = \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{x}\right). \tag{21}$$

Taking second derivatives gives us

$$\mathbf{J}(\mathbf{y} \mid \mathbf{x}) = \mathbf{V}^{-1}.$$

We now substitute this into (19) to get

$$\mathbf{\Sigma}_1^{-1} = \mathbf{V}^{-1} + \mathbf{\Sigma}_0^{-1}. \tag{22}$$

We substitute (21) into (20) to get

$$\mathbf{x}_1 = \mathbf{x}_0 + \mathbf{\Sigma}_1 \mathbf{V}^{-1}(\mathbf{y} - \mathbf{x}_0). \tag{23}$$

Now consider applying (22) and (23) *n* times to a series of *n* numbers starting with a non-informative prior $\mathbf{\Sigma}_0^{-1}$ = **0**. We can show by induction that the final $\mathbf{x}_1$ will be the arithmetic average of the *n* numbers or vectors assuming $\mathbf{V}^{-1}$ is nonsingular. This provides a way to compute an average without storing all the numbers. Alternating these computations with a migration following a normal random walk produces from (23) a Bayesian EWMA [12].

In a regression situation, **y** ~ N_k(**Zx**, **V**), this same analysis gives us

$$\mathbf{\Sigma}_1^{-1} = \mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z} + \mathbf{\Sigma}_0^{-1},$$

and

$$\mathbf{x}_1 = \mathbf{x}_0 + \mathbf{\Sigma}_1 \mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Z}\mathbf{x}_0). \tag{24}$$

Kalman filtering can be derived by repeated use of (24), obtaining the prior covariance matrix for the each observation $\mathbf{\Sigma}_{t|t-1}$ by adding a covariance matrix $\mathbf{W}_t$ to model a random walk between (*t*–1) and *t* to the posterior $\mathbf{\Sigma}_{t-1|t-1}$ from the previous observation [8, Sections 3-6, possibly after some deterministic change]. However, if the posterior information from the previous step $\mathbf{\Sigma}_{t-1|t-1}^{-1}$ is singular, we must consider this fact in handling the migration. In such cases, we use the information matrix rather than the covariance matrix as the primary representation of the variability of the distribution, because it is easier computationally to handle zero information than infinite variance. Let $\mathbf{Q}_0\mathbf{\Lambda}_0^{-1}\mathbf{Q}_0' = \mathbf{\Sigma}_{t-1|t-1}^{-1}$ denote the eigenvalue decomposition of $\mathbf{\Sigma}_{t-1|t-1}^{-1}$ omitting its null space. Then $\mathbf{Q}_0\mathbf{\Lambda}_0\mathbf{Q}_0'$ = the Moore-Penrose pseudo-inverse of $\mathbf{\Sigma}_{t-1|t-1}^{-1}$ and is therefore a reasonable representation of the singular covariance matrix $\mathbf{\Sigma}_{t-1|t-1}$. To get $\mathbf{\Sigma}_{t|t-1}$, we can't just add $\mathbf{W}_t$ to this $\mathbf{\Sigma}_{t-1|t-1}$, because that would make $\mathbf{\Sigma}_{t|t-1}$ nonsingular, ignoring the infinite variance in the orthogonal space of $\mathbf{Q}_0$. Instead, we compute $\mathbf{\Sigma}_{t|t-1}$ = $\mathbf{Q}_0\mathbf{\Lambda}_0\mathbf{Q}_0'$ + $\mathbf{Q}_0\mathbf{Q}_0'\mathbf{W}_t\mathbf{Q}_0\mathbf{Q}_0'$ = $\mathbf{Q}_0\left(\mathbf{\Lambda}_0 + \mathbf{Q}_0'\mathbf{W}_t\mathbf{Q}_0\right)\mathbf{Q}_0'$ and $\mathbf{\Sigma}_{t|t-1}^{-1}$ = $\mathbf{Q}_0\left(\mathbf{\Lambda}_0 + \mathbf{Q}_0'\mathbf{W}_t\mathbf{Q}_0\right)^{-1}\mathbf{Q}_0'$. If we do this starting with zero information, $\mathbf{\Sigma}_{1|0}^{-1}$ = $\mathbf{0}$, and ignore the migration by letting $\mathbf{W}_t = \mathbf{0}$, we can get ordinary least squares regression.

6. BAYES AND THE CENTRAL LIMIT THEOREM

Expression (19) relates to a more general result, namely that the sampling distribution of maximum likelihood estimators (MLEs) is, under very general regularity conditions, approximately normal with covariance matrix being the inverse of the information (e.g., [27]). Even with non-normal observations, $\mathbf{J}(\mathbf{y} \mid \mathbf{x})$ (under suitable regularity conditions) generally acts like precision parameter(s), being the inverse of variance-covariance matrices. These results are typically derived by writing the vector of MLEs as a weighted sum of Fisher's efficient scores and assuming that variations in $\mathbf{J}(\mathbf{y} \mid \mathbf{x})$ are sufficiently small (and the dominating measure for the prior sufficiently flat) that that the posterior is adequately approximated by a normal distribution with information $\mathbf{\Sigma}_1^{-1}$ and mean $\mathbf{x}_1$ computed via (19) and (20). Under suitable regularity conditions, we have exactly this structure in (20) and in the somewhat more general situations discussed in the appendix. In such cases, a Bayesian posterior will generally be more nearly normal than either the prior or the score from the data ([1]; [28]). Edgeworth correction terms could also be obtained to quantify rates of convergence to the central limit theorem, following [29], [32], [33], and [11], and the relative magnitude of such correction terms typically declines with the accumulation of posterior information.

Central limit convergence of MLEs has been proven with otherwise adequately behaved multimodal distributions with occasionally negative observed information $\mathbf{J}(\mathbf{y} \mid \mathbf{x})$. This property rests on the fact that observations with negative information are so relatively rare that they disappear almost surely with increasing numbers of observations. Alternatively, finite mixtures in prior and observation distributions can often be adequately approximated by the obvious finite mixtures in the posterior, dropping all but the dominant components as describe by West and Harrison [27, ch. 12].

7. ALTERNATIVE DEFINITIONS OF INFORMATION IN STATISTICS

Several different types of "information" have been defined and used in statistical work (see, e.g., [34]). The Fisher information is a tool of choice for developing approximate sampling distributions for maximum likelihood estimates, as discussed in the previous section. The observed information is also sometimes used for this purpose.

Shannon [31] argued that the information contained in a "message" (observation) **y** is the number of bits required to produce the equivalent reduction in uncertainty, which is $E\{-\log_2[f(\mathbf{y})]\}$. For example, if **y** is the outcome of the toss of an unbiased coin, then $E\{-\log_2[f(\mathbf{y})]\}$ = $0.5[-\log_2(0.5)]$ + $0.5[-\log_2(0.5)]$ = $\log_2(2)$ = 1. Important results in modern communication theory are based on Shannon's concept of information.

Using natural rather than base 2 logarithms, Kullback and Leibler [17] (see also [23]) quantified the information in an observation **y** for discriminating a probability density $f(\mathbf{y})$ from $g(\mathbf{y})$ as $E\{\log[f(\mathbf{y})/g(\mathbf{y})] | f\}$; they called this a measure of "distance" or "divergence" between *f* and *g*. With $I(\mathbf{x}, \mathbf{x}+\boldsymbol{\delta})$ = $E\{\log[f(\mathbf{y} | \mathbf{x})/ f(\mathbf{y} | \mathbf{x}+\boldsymbol{\delta})] | \mathbf{x}\}$, Kullback and Leibler showed that under suitable regularity conditions, the Fisher expected information was twice the second derivative of their "divergence" with respect to a perturbation:

$$E[J(\mathbf{y} | \mathbf{x}) | \mathbf{x}] = 2\left[\frac{\partial^2 I(\mathbf{x}, \mathbf{x}+\boldsymbol{\delta})}{\partial \boldsymbol{\delta}\ \partial \boldsymbol{\delta}'}\right].$$

To help educate our intuition about this, consider **y** ~ $N(\mathbf{x}, \boldsymbol{\Sigma})$. Then

$$\begin{aligned} I(\mathbf{x},\mathbf{x}+\boldsymbol{\delta}) &= \frac{1}{2} E\left\{[\mathbf{y}-(\mathbf{x}+\boldsymbol{\delta})]' \boldsymbol{\Sigma}^{-1}[\mathbf{y}-(\mathbf{x}+\boldsymbol{\delta})]-[\mathbf{y}-\mathbf{x}]' \boldsymbol{\Sigma}^{-1}[\mathbf{y}-\mathbf{x}]\right\} \\ &= E\,\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}[0.5\boldsymbol{\delta}+\mathbf{x}-\mathbf{y}] = 0.5\,\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}. \end{aligned} .$$

Since the Fisher information in this context is $\boldsymbol{\Sigma}^{-1}$, we find that the Fisher information here is precisely $2\left[\partial^2 I(\mathbf{x}, \mathbf{x}+\boldsymbol{\delta})/\partial \boldsymbol{\delta}\ \partial \boldsymbol{\delta}'\right]$, consistent with Kullback and Leibler's general result.

For a more general review of these and other types of "information" used in statistics, see [34], [10], [19], and [9].

In sum, several different concepts of "information" have been discussed in the statistics literature, with each serving different purposes. The focus of this article has been Fisher's efficient score and the observed information, which provide powerful tools for deriving exact and approximate posterior distributions.

8. SUMMARY

We discussed Bayes' rule of information generally in (11) and in approximate and exact normal applications in (19), (22) and (24). We also showed how keeping score with Bayes' theorem provides easy derivations of the posterior for the gamma-Poisson, beta-binomial, and exponential-uniform conjugate pairs. These tools have long been used when prior and observations are normal (e.g., [25] and [18]), but without substantive consideration of their more general utility. Yousry et al. [37] describe the use in quality control of an EWMA for binomial data with a beta prior. Their derivation is similar to the discussion in Example 2, Section 2 above, but without the convenience of using the concept of Fisher's efficient score or of Bayesian sequential updating, promoted as a general foundation for monitoring [13].

The results here are related to but different from the traditional frequentist result that the Fisher information for the joint distribution of two independent random variables is the sum of the Fisher information for each marginal [19, sec. 5a.4]. For example, with non-normal observations where normal distributions provide acceptable approximations to prior and posterior, it is sometimes appropriate to further simplify the posterior information computation in (19) by replacing the observed information $\mathbf{J}(\mathbf{y} \mid \mathbf{x})$ with its expectation over $\mathbf{y}$ given $\mathbf{x}$. If we do this twice starting from a noninformative prior with $\mathbf{J}(\mathbf{x}) = \mathbf{0}$, we get the result mentioned by Rao [17, sec. 5.a4].

In many cases, a normal distribution provides an adequate approximation to the posterior, even with nonlinear or non-normal likelihood. When it is not convenient to compute derivatives analytically, the score function and information from the data can be estimated by numerical differentiation.

After the posterior mode and information ($\mathbf{x}_1$, $\mathbf{\Sigma}_1^{-1}$) are found by iterating with (27) and (28), the adequacy of the normal approximation might be checked using importance sampling, computing, e.g., the difference between $l(\mathbf{x}|\mathbf{y})$ and the normal approximation at a sample of pseudo-random normal deviates following the approximating normal distribution. Of course, we must also assure ourselves that the posterior does not have another substantive mode that might be completely missed with this importance sampling. If substantive discrepancies are found, they can be reported with profile confidence intervals, marked to highlight the discrepancies between the profile and the normal approximation. Certain likelihoods (e.g., mixtures; see [35] or [24]) are known to have potential difficulties. These cases might be identified by excessive variability in the observed information from the data. Once identified, special procedures can be developed appropriate to the situation.

ACKNOWLEDGEMENTS

The author wishes to express appreciation to the PDF Solutions management team for their support and especially to George Cheroff, whose assistance with library research has been quite valuable.

REFERENCES

[1] Bernardo, J. M., Smith, A.F. M. (2000) *Bayesian Theory* (NY: Wiley, prop. 5.14).

[2] Box, G., and G. M. Jenkins (1970) *Time Series Analysis, Forecasting and Control* (San Francisco, Holden Day, sec. 4.3.1)

[3] Box, G., and Luceño, A. (1997) *Statistical Control by Monitoring and Feedback Adjustment* (NY: Wiley, ch. 10-11).

[4] DeGroot, M. H. (1970) *Optimal Statistical Decisions* (NY: McGraw-Hill, p. 39).

[5] Dey, D. K, Ghosh, S. K., and Mallick, B. K. (2000) *Generalized Linear Models: A Bayesian Perspective* (NY: Marcel Dekker, esp. ch. 3, p. 50, by Ibrahim and Chen)

[6] Durbin, J. (2004) "Introduction to State Space Time Series Analysis", ch. 1 in A. Harvey, S. J. Koopman, and N. Shephard, *State Space and Unobserved Component Models* (Cambridge, UK: Cambridge U. Pr., pp. 3-25, esp. p. 22)

[7] Durbin, J., and Koompan, S. J. (2002) *Time Series Analysis by State Space Methods*, corrected ed. (Oxford, UK: Oxford U. Pr., sec. 8.2, p. 157)

[8] Evans, M., and Schwartz, T. (2000) *Approximating Integrals via Monte Carlo and Deterministic Methods* (Oxford, UK: Oxford U. Pr.)

[9] Goel, P. K., and M. H. DeGroot (1979) "Comparison of Experiments and Information Measures", *Annals of Statistics*, 7: 1066-1077.

[10] Good, I. J. (1960) "Weight of Evidence, Corroboration, Explanatory Power, Information and the Utility of Experiments", *Journal of the Royal Statistical Society, series B*, 22: 319-331.

[11] Graves, S. B. (1983) *Edgeworth Expansions for Discrete Sums and Logistic Regression* (Ph.D. Dissertation, University of Wisconsin-Madison).

[12] _______, Bisgaard, S., and Kulahci, M. (2002) "Designing Bayesian EWMA Monitors Using Gage R & R and Reliability Data" (technical report downloadable from www.prodsyse.com).

[13] _______, Bisgaard, S., Kulahci, M., Van Gilder, J., Ting, T., Marko, K., James, J., Zatorski, H., Wu, C. (2001) *Foundations of Monitoring Dynamic Systems* (technical report downloadable from www.prodsyse.com).

[14] Grigg, O. A., and Spiegelhalter, D. J. (2005) "A Simple Risk-Adjusted Exponentially Weighted Moving Average", MRC Biostatistics Unit: Technical report 2005/2, Medical Research Council of the Laboratory of Molecular Biology, Cambridge, UK (http://www.mrc-bsu.cam.ac.uk/BSUsite/Publications/pp+techrep.shtml, accessed 1 August 2005)

[15] Jørgensen, B. (1987) "Exponential Dispersion Models" (with discussion), *Journal of the Royal Statistical Society*, B-49: 127-162

[16] Klein, B. M. (2003) "State Space Models for Exponential Family Data" (http://www.stat.sdu.dk/publications/monographs/m001/KleinPhdThesis.pdf for the report and http://genetics.agrsci.dk/~bmk/index.html for the software 2005/07/17)

[17] Kullback, S., and Leibler, R. A. (1951) "On information and sufficiency", *Annals of Mathematical Statistics*, 22, 79-86.

[18] Kwon, I. (1978) *Bayesian Decision Theory with Business and Economic Applications* (NY: Petrocelli / Charter, pp. 214-215).

[17] Lindley, D. V. (1972) *Bayesian Statistics: A Review* (Philadelphia, PA: Society for Industrial and Applied Mathematics, sec. 12.6).

[20] Martz, H. F., Parker, R. L., and Rasmuson, D. M. (1999) "Estimation of Trends in the Scram Rate at Nuclear Power Plants", *Technometrics*, 41: 352-364.

[21] McCullagh, P., and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd ed. (NY: Chapman & Hall, p. 470)

[22] McCulloch, C. E., and Searle, S. R. (2001) *Generalized, Linear, and Mixed Models* (NY: Wiley)

[23] McCulloch, R. E. (1989) "Local Model Influence", *Journal of the American Statistical Association*, 84: 473-478.

[24] McLachlan, G., and Peel, D. (2000) *Finite Mixture Models* (NY: Wiley).

[25] Morgan, B. W. (1968) *An Introduction to Bayesian Statistical Decision Processes* (Englewood Cliffs, NJ: Prentice-Hall, pp. 63-67).

[26] Pole, A., West, M., and Harrison, H. (1994) *Applied Bayesian Forecasting and Time Series Analysis* (NY: Chapman & Hall).

[27] Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*, 2nd ed. (NY: Wiley).

[28] Press, S. J. (1972) *Applied Multivariate Analysis* (NY: Holt, Rinehart and Winston, theorem 4.6.1).

[29] Robert, C. P. (2001) *The Bayesian Choice* (NY: Springer, sec. 3.5.5).

[30] Rossman, A. J., Short, T. H., and Parks, M. T. (1998) "Bayes Estimators for the Continuous Uniform Distribution", *Journal of Statistics Education*, 6(3), http://www.amstat.org/publications/jse/v6n3/rossman.html.

[31] Shannon, C. E. (1948) "A Mathematical Theory of Communication", *Bell System Technical Journal*, 27, pp. 379-423; pp. 623-656.

[32] Skovgaard, I. M. (1981) "Edgeworth Expansions of the Distribution of the Maximum Likelihood Estimators in the General (non i.i.d.) Case", *Scandinavian Journal of Statistics*, 8, 227-236.

[33] _______ (1986) "On Multivariate Edgeworth Expansions", *International Statistical Review*, 54: 29-32.

[34] Soofi, E. S. (2000) "Principal Information Theoretic Approaches", *Journal of the American Statistical Association*, 95: 1349-1353.

[35] Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions* (NY: Wiley).

[36] West, M. and Harrison, P. J. (1999) *Bayesian Forecasting and Dynamic Models*, 2nd ed., corrected 2nd printing (NY: Springer).

[37] Yousry, M. A., Sturm, G. W., Felitz, C. J., and Noorossana, R. (1991) "Process Monitoring in Real Time: Empirical Bayes Approach -- Discrete Case", *Quality and Reliability Engineering International*, 7: 123-132.

APPENDIX: NON-CONSTANT OBSERVED INFORMATION

In this appendix, we develop an iteration to an approximate normal posterior $N_p(\mathbf{x}_1, \mathbf{\Sigma}_1)$ from a normal prior $N_p(\mathbf{x}_0, \mathbf{\Sigma}_0)$ and either non-normal data or data with normal errors nonlinearly related to the parameters of interest **x**. We shall not prove here anything about the convergence of our iteration; such a proof would follow the lines of comparable results on convergence of MLEs.

The iteration will ultimately require keeping score at the posterior mode $\mathbf{x} = \mathbf{x}_1$, rather than the prior mode as with (20), substituting (17) into (2) to obtain the following:

$$\mathbf{0} = \left[\frac{\partial l(\mathbf{y} \mid \mathbf{x} = \mathbf{x}_1)}{\partial \mathbf{x}}\right] - \boldsymbol{\Sigma}_0^{-1}(\mathbf{x}_1 - \mathbf{x}_0). \tag{25}$$

Since $\mathbf{x}_1$ is initially unknown, we expand the score from the data in a Taylor approximation about an arbitrary point $\mathbf{x} = \xi$, beginning from $\xi = \mathbf{x}_0$, as follows:

$$\left[\frac{\partial l(\mathbf{y} \mid \mathbf{x} = \mathbf{x}_1)}{\partial \mathbf{x}}\right] = \left[\frac{\partial l(\mathbf{y} \mid \mathbf{x} = \xi)}{\partial \mathbf{x}}\right] - \mathbf{J}(\mathbf{y} \mid \mathbf{x} = \xi)(\mathbf{x}_1 - \xi).$$

We substitute this into (25) to get the following:

$$0 = \left[\frac{\partial l(\mathbf{y} \mid \mathbf{x} = \xi)}{\partial \mathbf{x}}\right] - \mathbf{J}(\mathbf{y} \mid \mathbf{x} = \xi)(\mathbf{x}_1 - \xi) - \boldsymbol{\Sigma}_0^{-1}(\mathbf{x}_1 - \mathbf{x}_0). \tag{26}$$

We begin each iteration by evaluating (11) at $\mathbf{x} = \xi$ using (18) as follows:

$$\boldsymbol{\Sigma}_{1(\xi)}^{-1} = \mathbf{J}(\mathbf{y} \mid \mathbf{x} = \xi) + \boldsymbol{\Sigma}_0^{-1}. \tag{27}$$

By substituting this into (26), we get the following:

$$\boldsymbol{\Sigma}_{1(\xi)}^{-1}\mathbf{x}_1 = \left[\frac{\partial l(\mathbf{y} \mid \mathbf{x} = \xi)}{\partial \mathbf{x}}\right] + \mathbf{J}(\mathbf{y} \mid \mathbf{x} = \xi)\xi + \boldsymbol{\Sigma}_0^{-1}\mathbf{x}_0. \tag{28}$$

Each iteration involves solving (28) for $\mathbf{x}_1$. If the difference between $\mathbf{x}_1$ and ξ is not sufficiently small, we replace ξ by the latest estimate of $\mathbf{x}_1$ in (27) and (28) and repeat the operation; if convergence is not obviously monotonic, then we may employ some form of step size control, replacing ξ by an appropriate linear interpolation between the previous ξ and the latest estimate of $\mathbf{x}_1$.