Recursive Inference for Inverse Problems using Variational Bayes Methodology

Invited Paper

Anthony Quinn*
Department of Electronic and Electrical Engineering
Trinity College Dublin
Ireland
aquinn@tcd.ie

ABSTRACT

The epistemic inverse problem is treated via Bayesian inference in this paper. In particular, the conditions for recursive computation of the required posterior inference are recalled. The emphasis is on appropriate structure in the observation model and choice by the Bayesian designer of a prior matched to that structure. Bayesian conjugate inference for the exponential family (EF) of observation models is recalled. This inspires progress with design of recursive algorithms for the time-variant (Bayesian filtering) case, using the variational Bayes (VB) approximation. A rich class of augmented observation models is defined, for which the posterior inference is closed under a local VB approximation, a principle known as *VB-conjugacy*. The key mathematical object is the VB-observation model, arising from application of VB in each data step. We force this to be an EF member. The theory is specialized to finite mixtures of heterogenous components, requiring recursive evaluation of one sufficient statistic per component, with the posterior component weights (*i.e.* filtering distribution) evaluated in a principled way. Further specialization to signal-independent system modelling is also considered. An extended case study in decoding and synchronization for the phase-uncertain digital receiver is presented as a key application of VB-conjugate recursive inference.

1. THE EPISTEMIC INVERSE PROBLEM

At its most fundamental, the inverse problem concerns the inference of an unknown object of interest, θ, based on related observations, x. In the Bayesian framework, we never model unknowns themselves, but, rather, model our beliefs about propositions formulated in terms of the unknowns. As Bayesian modellers or designers, we must specify the probability triple, $(\Omega, \mathcal{A}, \Pr[\cdot])$, induced by couple, (x, θ). We will make the usual, but reasonably flexible, assumptions [7] that $\mathcal{A}$ is a σ-algebra of Borel sets on $\Omega \subseteq \mathbb{R}^m$, and that $\Pr[\cdot]$ is a *known* mapping from $\mathcal{A}$ to $[0, 1]$. In this case, we can posit a *known* probability density function, $f(x, \theta)$, on Ω. Since the implied $\Pr[\cdot]$ is a quantfier of the observer's belief, *i.e.* epistemic (subjective) probability [2, 5, 22], we refer to the elicitation of $f(x, \theta)$ as *Bayesian* probability modelling. We recognize no restriction to aleatory probabilities derived from frequency-based experience of random (x, θ).

The epistemic inverse problem refers to inference of θ given x, *i.e.* computation of $f(\theta|x)$. Direct application of probability calculus provides the solution

$$f(\theta|x) = \frac{f(x, \theta)}{\int f(x, \theta) d\theta}.$$

In standard settings of the inverse problem [11], the observer's belief structure is not expressed via the joint pdf, $f(x, \theta)$, but via an observation model, $f(x|\theta)$, and a prior model, $f(\theta)$, for θ. In this sense, θ is the *parameter* of the observation model, and the task is one of parametric inference. Then, by Bayes' rule,

$$f(\theta|x) \propto f(x|\theta) f(\theta), \tag{1}$$

with, once again, the constant of proportionality being given by $K(x)^{-1}$, where $K(x) = f(x) = \int f(x|\theta) f(\theta) d\theta$. It is because of the factorization in (1) that the epistemic inverse problem is referred to as *Bayesian parametric inference*.

In many engineering contexts, x is observed as the output of a system with unknown parameters, θ, driven by input, u. Once again, we eschew any notion of modelling the system directly, in favour of modelling our belief about the system output, x, with θ the parameter of the epistemic model, now also involving u:

$$f(x, \theta, u) = f(x|\theta, u) f(\theta, u).$$

Here, u is called a *hidden variable* or field. Another interpretation follows by noting

$$f(x|\theta) = \int_u f(x|\theta, u) f(u) du. \tag{2}$$

In this context, u can be interpreted as the *auxiliary* or *state variable* in an *augmented model* for x; *i.e.* the modelling of our belief about x, via θ, is facilitated by knowledge of u,

*This work was partially supported by SFI grant 08/RFP/MTH1710

NCMIP 2011, May 25, Paris, France

DOI 10.4108/icst.valuetools.2011.246122

as is naturally the case in system modelling, and in applications such as computed tomography [11]. (2) is an uncountably infinite mixture model with mixing density $f(u)$. An important special case arises when u is discrete, with possible states, $\{u_j\}$. Using the Radon-Nykodym theorem, $f(u) = \sum_j \Pr[u = u_j]\delta(u - u_j)$, where $\delta(\cdot)$ is the Dirac δ-function. Inserting into (2) and using the sifting property of $\delta(\cdot)$:

$$f(x|\theta) = \sum_j \Pr[u = u_j] f(x|\theta, u_j), \tag{3}$$

where, in general, $f(x|\theta, u_j) = f_j(x|\theta_j)$, where $\theta = \cup_j \theta_j$, and we believe that x is conditionally independent of $\theta_{\setminus j}$ (the complement of θ_j in θ), given $u = u_j$. Here, $f_j(x|\theta_j)$ are the *components* of the countable (possibly finite) mixture model (3). This case is of overwhelming importance in engineering applications. Examples include (i) phoneme-based modelling of speech using hidden Markov chains (typically called *hidden markov modelling (HMM)* [16]),where u is the phoneme sequence; (ii) texture-based modelling of images using hidden Markov random fields [17], where u is the spatial map of texture classes; (iii) tomographic reconstruction of digital imagery [11], where u is the grey-scale volumetric image to be reconstructed; and (iv) modelling of digital receiver data [8], where u is the digital sequence injected at the transmitter (Sections 2.3 and 4, to follow).

Returning to the system context, with u as input, each of the following tasks can be formulated as a particular epistemic inverse problem.

Blind deconvolution: the observation model describing our belief about the system (*i.e.* about u-x interactions) has parametric uncertainty, and the aim is to infer u:

$$f(u|x) \propto \int_\theta f(x, \theta, u) d\theta.$$

This is synonymous with Bayesian filtering [22] in the time-variant case of u (Section 3).

Non-blind deconvolution: the observation model is known. Then

$$f(u|x) \propto f(x, u).$$

Blind equalization: the aim is to quantify beliefs about the system, parameterized by θ, when driven by unknown input, u. Then

$$f(\theta|x) \propto \int_u f(x, \theta, u) du.$$

Non-blind equalization: the input is known. Then

$$f(\theta|x, u) \propto f(x, \theta|u).$$

1.1 Hidden Label Field

Consider the case where u is a finite-state scalar variable, with ordered states (alphabet), $\mathbf{u} = [u_1, \ldots, u_M]'$, with prior probabilities, $p = [p_1, \ldots, p_M]'$, respectively. In this case, it is convenient to posit a *label* or *pointer* variable, $l \in \mathbb{E}_M$, where $\mathbb{E}_M$ is the elementary basis in $\mathbb{R}^M$, such that $u = l'\mathbf{u}$. It follows that $l \sim \mathcal{M}u_l(1, p) \equiv l'p$ *a priori*, a *multinomial* variable. Hence:

$$f(x|\theta, u) = f(x|\theta, l) = \prod_{j=1}^{M} [f_j(x|\theta_j)]^{l_j} \tag{4}$$

$$f(x|\theta, p) = \sum_{j=1}^{M} p_j f_j(x|\theta_j), \tag{5}$$

a *finite mixture model (FMM)* [19] of M (signal-dependent) components, $f(x|\theta, l = e_j) \equiv f_j(x|\theta_j)$, where $\theta_j \subseteq \theta$.

2. RECURSIVE BAYESIAN INFERENCE VIA CONJUGACY

Given a sequence of observations, $\mathbf{x}_n = [x_1, \ldots, x_n]'$, our aim in signal processing [22] is to design a computationally tractable recursive algorithm for processing $\mathbf{x}_n$,

$$s_n = g(x_n, \sigma(\mathbf{x}_n), s_{n-1}) \tag{6}$$

for some time-invariant, finite-dimensional mapping, $g(\cdot)$, where $\sigma(\cdot)$ is a finite memory of $\mathbf{x}_n$. In classical signal processing, s_n provides an estimate of the quantity of interest, θ. In contrast, Bayesian signal processing is concerned with the computation of the epistemic probability function, $f(\theta|\mathbf{x}_n)$ (Section 1), and so s_n must be a *sufficient statistic*, as follows:

$$f(\theta|\mathbf{x}_n) \equiv f(\theta|s_n).$$

A sufficient condition is that the observation model, $f(x|\theta)$, belong to the *exponential family (EF)* of distributions [9], so that x and θ have separate kernels:

$$f(x|\theta) = a(\theta) b(x) \exp\left[\langle c(\theta), h(x)\rangle\right]. \tag{7}$$

$\langle \cdot, \cdot \rangle$ denotes a scalar function, linear in the first argument. If, in this case, the prior is *designed* as

$$f(\theta|s_0) \propto a(\theta)^{\nu_0} \exp\left[\langle c(\theta), v_0\rangle\right], \tag{8}$$

with $s_0 = \{v_0, \nu_0\}$, and with finite normalizing constant, $K(s_0)$, then, by Bayes' rule:

$$f(\theta|x, s_0) = f\left(\theta|(v_0 + h(x), \nu_0 + 1) \equiv f\left(\theta|s_1\right).\right.$$

Thus, the conjugate distribution is functionally invariant under Bayesian inference, *i.e.* it is the *learning invariant* for the observation model (7). Given exchangeable [5] (i.e. *conditionally independent, identically distributed (ciid)*) data, $\mathbf{x}_n$, the sufficient statistics, $s_n = \{v_n, \nu_n\}$, are updated in a simple additive recursive way:

$$\left.\begin{array}{rcl} v_n &=& v_{n-1} + h(x_n), \\ \nu_n &=& \nu_{n-1} + 1, \end{array}\right\} \quad n = 1, 2, 3, \ldots \tag{9}$$

Computationally, Bayesian inference requires maintenance of a counter, ν, and a finite-length array, v, accumulating time-invariant mapping, $h(\cdot)$, of each observation. Two important applications are in estimation and prediction.

2.1 Bayesian Estimation

The minimum-risk estimate of $\widehat{g(\theta)}$ requires that a loss function be defined on the product-space of $\widehat{g(\theta)}$ and θ [1]. Nonlinear moments are common projections of $f(\theta|\mathbf{x}_n)$:

$$\widehat{g(\theta)}_n \equiv \mathsf{E}_{f(\theta|\mathbf{x}_n)}\left[g(\theta)\right] = \int_\theta g(\theta) f(\theta|\mathbf{x}_n) d\theta.$$

Note that the estimate is a function only of the sufficient statistics, s_n, in the conjugate learning context above. A characteristic, then, of Bayesian estimation is that recursion always happens at the level of s_n (9), and not, generally, at the level of $\widehat{g(\theta)}_n$. In contrast, classical estimators are typically recursive at the level of the estimator itself.

2.2 Bayesian Conjugate Prediction

The prior predictor of x in the conjugate learning context is

$$f(x|s_0) \equiv \int_\theta f(x|\theta) f(\theta|s_0) d\theta = b(x) \frac{K(s_1)}{K(s_0)},$$

following immediately from (7), (8), and the definition of the normalizing constant, $K(\cdot)$. Of particular interest is the posterior predictor:

$$f(x_{n+1}|s_n) = b(x_{n+1}) \frac{K\left(v_n + h(x_{n+1}), \nu_n + 1\right)}{K(s_n)}.$$

Once again, $\mathbf{x}_n$ is recursively and finitely encoded via s_n, but x_{n+1} enters the distribution via the normalizing constant, $K(\cdot)$. In many cases, the highly nonlinear or open form of $K(\cdot)$ renders the posterior predictor intractable, in the sense that its moments are not available analytically.

2.3 Example: Bayesian Binary Learning

The canonical example of learning via binary observations can be treated insightfully using Bayesian inference. We consider two cases: (i) a binary memoryless (*i.e.* ciid) source, u, is observed directly; and (ii) it is observed via outputs, x, from a binary, memoryless, noisy channel. From Section 1.1, the multinomial (Bernoulli in this case) label is $l \in \mathbb{E}_2 = \{e_1, e_2\}$, and the symbol alphabet is $\mathbf{u} = [1, 0]'$ (*i.e.* $M = 2$), giving $u = l'\mathbf{u}$. Hence, $f(l|p) \equiv \mathcal{M}u_l(1, p)$, where $p = (p_1, p_2)'$, and $\Pr[u = 1] = p_1$, with, of course, $\Pr[u = 0] = p_2 = 1 - p_1$. Note that $\mathcal{M}u_l(1, p)$ is an EF member (7). Its conjugate prior is readily shown to be $p \sim \mathcal{D}ir(\nu_0)$, the Dirichlet distribution [2, 9, 22] (specialized as the beta distribution in this Bernoulli case). Hence, Dirichlet is the learning invariant under ciid multinomial observations, l_i:

$$f(p|\mathbf{u}_n) = f(p|\mathbf{l}_n) = \mathcal{D}ir(\nu_n),$$
$$\nu_n = \nu_{n-1} + l_n.$$

The sufficient statistics, ν_n, simply count the number of realizations of each bit, conforming with intuition. It is the mapping from ν_n to a quantifier of belief, $\mathcal{D}ir(\cdot)$, that marks this treatment out as a Bayesian one.

Next, consider case (ii) in the last paragraph. u are now hidden states, observed via binary channel outputs, x. The inference of u, via x, constitutes an epistemic inverse problem, as discussed in Section 1. We define the label of the output as m, with the same symbol alphabet, $\mathbf{u}$, as that of the input, so that $x = m'\mathbf{u}$. For a single observation, the joint model is

$$\begin{aligned} f(l, m, p) &= f(m|l) f(l|p) f(p) \\ &\equiv \mathcal{M}u_m(1, \epsilon l) \mathcal{M}u_l(1, p) \mathcal{D}ir_p(\nu_0), \end{aligned} \quad (10)$$

where ϵ is the $M \times M$ transition probability matrix of the M-ary channel (binary in this case). Marginalizing over the hidden variable, l, and noting that $\mathcal{M}u_m(1, \epsilon l) = m'\epsilon l$:

$$f(p|m) \propto \alpha_1^{(1)} \mathcal{D}ir_p(\nu_1^{(1)}) + \alpha_1^{(2)} \mathcal{D}ir_p(\nu_1^{(2)}), \quad (11)$$

where $\alpha_1^{(j)} = m'\epsilon_j \frac{B(\nu_1^{(j)})}{B(\nu_0)}$, with ϵ_j the jth column of ϵ, and

$$\nu_1^{(j)} = \nu_0 + e_j.$$

Here, we have used the fact that $\mathcal{D}ir \equiv \mathcal{B}eta$, the beta distribution in the $M = 2$ (binary) case, and is normalized via $B(\cdot)$, the beta-function. Finally, marginalizing over the unknown source probability, p, the predictor is $f(m) = \alpha_1^{(1)} + \alpha_1^{(2)}$, and dividing this into (11), we obtain

$$f(p|m) = w_1^{(1)} \mathcal{D}ir_p(\nu_1^{(1)}) + w_1^{(2)} \mathcal{D}ir_p(\nu_1^{(2)}), \quad (12)$$

where $w_1^{(j)} = \alpha_1^{(j)} / f(m)$. This binary mixture model involves two statistics, namely the (normalized) component weights, $w_1^{(j)}$. In general, given a sequence of n ciid input bits, $u_1, \ldots, u_n$, the number of components in the mixture, $f(p|\mathbf{x}_n) = f(p|\mathbf{m}_n)$, is $n + 1$, requiring evaluation of $n + 1$ statistics. Hence, $\mathbf{x}_n$ is not finitely encoded as $n \to \infty$, and the Bayesian inference procedure is non-recursive (6). Technically, the loss of recursiveness is due to the fact that the (marginal) observation model, $f(m|p) = p_1 \mathcal{M}u_m(1, \epsilon_1) + p_2 \mathcal{M}u_m(1, \epsilon_2)$ is–like all finite mixture models—not an EF member. Owing to the *exchangeability* of the input symbols [5], u_i, *i.e.* they are ciid, the number of components increases *linearly* with n in this case. More generally, for non-exchangeable data, the number of components grows exponentially, as M^n.

For completeness, the *decoding* or *filtering* distribution, $f(l|m)$ is found by marginalizing over p in (10):

$$f(l|m) = m'\epsilon l \frac{B(\nu_0 + l)}{B(\nu_0) f(m)}.$$

After n samples, this marginalization is with respect to $f(p|\mathbf{x}_n)$, and becomes intractable, as noted above.

2.4 Conjugate Inference for Augmented Models

It may be possible to express a non-EF observation model, $f(x|\theta)$, via an augmented model with auxiliary variable, u, such that the latter is an EF member. Recalling (2), then

$$f(\theta|x) \propto \int_u f(x|\theta, u) f(u) f(\theta|u) du. \quad (13)$$

We are free to design the conditional prior, $f(\theta|u)$, and do so as

$$f(\theta|u) = \frac{f(\theta, u|s_0)}{f(u) K(u)},$$

where $f(\theta, u|s_0)$ is the conjugate distribution for $f(x|\theta, u)$, and $K(u)$ is the required normalizing term. Inserting into (13), the following recursive Bayesian inference procedure is achieved:

$$f(\theta|\mathbf{x}_n) \propto \int_u \frac{1}{K(u)} f(\theta, u|s_n) du, \quad (14)$$

where s_n is the finite encoding of $\mathbf{x}_n$ into sufficient statistics for augmented parameters, $\{\theta, u\}$, and conjugate learning has been achieved (9). Hence, θ is inferred at any time via the projection above. An important example is

Bayesian learning with Student-t ($\mathcal{S}t$) data. While not an EF member, $\mathcal{S}t$ can be recognized as a scale mixture of normals [2]. The conjugate prior for normal learning is the normal-inverse-gamma ($\mathcal{N}i\mathcal{G}$) distribution [22]. Hence, recursive inference is achieved for $\mathcal{S}t$ using $\mathcal{N}i\mathcal{G}$ as the learning invariant, with projection via (14) at any time, n.

3. RECURSIVE STRUCTURES VIA THE VARIATIONAL BAYES APPROXIMATION

In cases where the number of state variables (hidden field) increases as n—such as in a system context with time-variant input $u \equiv \mathbf{u}_n$ (Section 1)—the posterior inference will typically be intractable, in the sense that integrals required for important inferences such *Bayesian filtering*,

$$f(\mathbf{u}_i|\mathbf{x}_n) \propto \int_{\mathbf{u}_{n\setminus i}} f(\mathbf{u}_n|\mathbf{x}_n) d\mathbf{u}_{n\setminus i}$$

(where $\mathbf{u}_{n\setminus i}$ is the complement of $\mathbf{u}_i$ in $\mathbf{u}_n$), and system identification (14), cannot, in most cases, be expressed in closed form as $n \to \infty$, even in the case of an EF augmented model (Section 2.4). The restrictive conditions for tractability in Bayesian filtering are given in [4]. In an on-line interpretation of the same problem, an integration over each new u_n is required after each data update, x_n, disrupting the additive recursive structure (9) in many cases. This is immediately evident for finite-state u_n, such as the digital learning example in Section 2.3, where the (marginal) observation model is a finite mixture model (FMM) (5). The augmented model in binary learning is $f(m|l,p) = f(m|l) = \mathcal{M}u_m(1,\epsilon l) \in \text{EF}$, but we found, nevertheless, that $\mathbf{x}_n$ could not be finitely encoded in the posterior inference of p as $n \to \infty$ (12). For FMMs, in general, the statistic accumulated for each possible state trajectory is, from (9):

$$v_n(\mathbf{l}_n) = v_0 + \sum_{i=1}^{n} h_{l_i}(x_i). \tag{15}$$

In the case where this is distinguishable under re-orderings of $\mathbf{x}_n$, M^n distinct statistics must be maintained, via this number of components in $f(\theta|\mathbf{x}_n)$. In the case where (15) is indistinguishable under data re-orderings (exchangeable), there are $\begin{pmatrix} n+M-1 \\ M-1 \end{pmatrix}$ statistics/components, implying a polynomial increase with n, as was the case for binary learning above.

In all such cases, an approximation is required to bound the computations associated with the epistemic inverse problem. The golden standard exploits the invariance (sifting property) of $\delta(\cdot)$, and the strong law of large numbers, by *stochastically* sampling from the space of trajectories, $\mathbf{u}_n$, using various strategies collectively referred to as *particle filtering* [6, 22]. While convergence criteria can be satisfied for the particle filter, the approach is often computationally expensive.

Deterministic approximation of the posterior distribution is an important alternative. A naïve certainty equivalence (CE) approach imposes the projection $f(u_n) \to \delta(u_n - \widehat{u_n})$, for some deterministic choice, $\widehat{u_n}$. Inserting into (2):

$$f(x_n|\theta) \approx f(x_n|\theta_{\widehat{u_n}}, \widehat{u_n}). \tag{16}$$

If (i) $\cup_n \theta_{u_n} = \theta$ is finite-dimensional, and (ii) the augmented model (Section 2.4), $f(x_n|\theta_{u_n}, u_n)$, is an EF member $\forall u_n$, then the additive recursive structure in (9) is restored for each (of the finite number of choices of) θ_{u_n}. For example, the FMM (5) with EF components, $f_j(x_n|\theta_j)$, satisfies these conditions. Adopting a CE approximation, $\widehat{u_n} \in \mathbf{u}$, such that the estimated pointer is $\widehat{l_n} \in \mathbb{E}_M$ (*i.e.* $\widehat{u_n} = \widehat{l'_n}\mathbf{u}$ (Section 1.1)), then the recursive Bayesian inference algorithm under this CE approximation is

$$\left.\begin{aligned} v_{j,n} &= v_{j,n-1} + \delta_{l_j,\widehat{l_j}} h_j(x_n), \\ \nu_n &= \nu_{n-1} + \widehat{l_n}, \end{aligned}\right\} \quad j = 1,\ldots,M, \tag{17}$$

where $\delta_{\cdot,\cdot}$ is the Kronecker δ-function. Only the sufficient statistics of the actively inferred component, $\widehat{u_n}$, are updated at each step, with ν_n now an array of M counters, one of whose elements is incremented at each step. For simplicity, we have made the assumption here that $\widehat{u_n} \in \mathbf{u}$, the state-space of the hidden field, u. This assumption is satisfied, for example, when $\widehat{u_n} = \arg\max_{u_n} f(u_n|\mathbf{x}_n)$, or the median, and where $f(u_n|\mathbf{x}_n)$ is computed via the approximate inference (17) at $n-1$. In contrast, the *Quasi-Bayes (QB)* approximation [19, 22] assigns $\widehat{u_n} = \mathsf{E}_{f(u_n|\mathbf{x}_n)}[u_n]$, in which case $\widehat{u_n} \notin \mathbf{u}$ in general. QB will be treated as a special case of the the variational Bayes approximation, which follows.

3.1 The Variational Bayes Approximation

Formally, the projection, $u_n \to \widehat{u_n}$ is a *mean-field approximation* [13]. A more principled design of the mean-field approximation involves approximating $f(\theta, u_n|\mathbf{x}_n)$ *variationally* within a constrained space, $\mathbb{F}_c$, of approximating distributions. If (i) the space of independent distributions, $f(\theta|\mathbf{x}_n)f(u_n|\mathbf{x}_n)$, is chosen as $\mathbb{F}_c$, and (ii) Kullback-Leibler divergence (KLD) from members of $\mathbb{F}_c$ to the exact distribution, $f(\cdot)$, is chosen as the objective function, then the variational minimizer, $\tilde{f}(\theta|\mathbf{x}_n)\tilde{f}(u_n|\mathbf{x}_n)$, is known as the *variational Bayes (VB)* approximation [22]. Under non-restrictive regularity conditions on $f(\cdot)$, the VB-approximation, $\tilde{f}(\cdot)$, satisfies the following implicit equations:

$$\begin{aligned} \ln \tilde{f}(\theta|\mathbf{x}_n) &\propto \mathsf{E}_{\tilde{f}(u_n|\mathbf{x}_n)}\left[f(\theta, u_n, \mathbf{x}_n)\right], \\ \ln \tilde{f}(u_n|\mathbf{x}_n) &\propto \mathsf{E}_{\tilde{f}(\theta|\mathbf{x}_n)}\left[f(\theta, u_n, \mathbf{x}_n)\right]. \end{aligned} \tag{18}$$

The iterative solution of (18) implements a gradient descent algorithm, guaranteeing convergence only to a local minimum. In practice, initialization issues (sensitivity to initial values) are rarely encountered. The systematic and tractable solution of the VB approximation via the iterative VB (*i.e.* IVB) algorithm is known as the *VB method*, and is elaborated for offline and online inference of time-invariant parameters, θ, in [22], and for the Bayesian filtering problem (inference of $\mathbf{u}_n$) in [23]. In the machine learning literature, it is common to characterize the VB approximation, $\tilde{f}(\cdot)$ (where it is known as the naïve mean-field approximation due to the independence constraint on $\mathbb{F}_c$ in (i) above), as that element of $\mathbb{F}_c$ providing the greatest lower bound on the (intractable) negative free energy of $f(\cdot)$. Note that the *VB-marginals* in (18) are normalizable without requiring access to the (typically intractable) normalizing constant of $f(\theta, u_n|\mathbf{x}_n)$, hence the use of $f(\theta, u_n, \mathbf{x}_n)$—and not the posterior, $f(\theta, u_n|\mathbf{x}_n)$—in the integrand.

The VB approximation is a local approximation, and therefore no convergence criterion can be posited in the filtering context above[1] (18), where the approximation must be applied at each time, n. However, the computational cost of VB is typically far less than that for particle filtering, with the prospect of out-performing the latter in computationally constrained environments [23].

3.2 VB-Conjugacy and Recursive VB Inference

An important role for the VB approximation is in restoring recursive computations in the epistemic inverse problem for non-EF models, $f(x|\theta,u)$. Consider the recursive application of Bayes' rule (sometimes called the 'data step' [23]) at time n in the Bayesian filtering context. For clarity of presentation, we will assume an independent (static) hidden field, $\mathbf{u}_n$, but the principles developed below can be applied more widely:

$$f(\theta,u_n|\mathbf{x}_n) \propto f(x_n|\theta,u_n)f(\theta|\mathbf{x}_{n-1})f(u_n). \tag{19}$$

Applying VB (18) as a local approximation after this nth data step, the *VB-marginal* of θ is

$$\tilde{f}(\theta|\mathbf{x}_n) \propto \tilde{f}(x_n|\mathbf{x}_{n-1},\theta)f(\theta|\mathbf{x}_{n-1}). \tag{20}$$

Here, we define the *VB-observation model* [22, 23] as

$$\ln \tilde{f}(x_n|\mathbf{x}_{n-1},\theta) \propto \mathsf{E}_{\tilde{f}(u_n|\mathbf{x}_n)}\left[\ln f(x_n|\theta,u_n)\right], \tag{21}$$

which is implicitly defined via the *VB-filtering distribution*:

$$\ln \tilde{f}(u_n|\mathbf{x}_n) \propto \mathsf{E}_{\tilde{f}(\theta|\mathbf{x}_n)}\left[\ln f(x_n|\theta,u_n)\right] f(u_n). \tag{22}$$

If (21) is a member of the dynamic exponential family (DEF) [9, 22], then it has a conjugate distribution. In this case, we can *design* our prior conjugate to the VB-observation model (8), *i.e.* $f(\theta|\mathbf{x}_0) \equiv f(\theta) \equiv f(\theta|s_0)$. The VB-marginal of θ therefore maintains the same functional form $\forall n$; *i.e.* $\tilde{f}(\theta|\mathbf{x}_n) = f(\theta|s_n)$ in (20), with recursive computation of the the sufficient statistics restored:

$$\begin{aligned} v_n &= v_{n-1} + h(x_n,\sigma(\mathbf{x}_n)) \\ \nu_n &= \nu_{n-1} + 1. \end{aligned} \tag{23}$$

Note that the hidden field (input) is inferred statically, $\forall n$, via the VB-filtering distribution, $\tilde{f}(u_n|\mathbf{x}_n) \equiv \tilde{f}(u_n|s_n)$ (22), which is therefore not required to have conjugate properties. The extension of these results to the case of a dynamic (dependent) hidden field, $\mathbf{u}_n$, is presented in [22].

The following special cases of the VB approximation provide further simplifications:

RVB: Restricted VB. If the marginal of u_n is fixed at $\overline{f}(u_n|\mathbf{x}_n)$ in $\mathbb{F}_c$ (Section 3.1), then the KLD may be minimized variationally in θ alone. The VB-observation model in (21) becomes

$$\ln \tilde{f}(x_n|\mathbf{x}_{n-1},\theta) \propto \mathsf{E}_{\overline{f}(u_n|\mathbf{x}_n)}\left[\ln f(x_n|\theta,u_n)\right].$$

This is an explicit assignment. Hence, the iterative VB (IVB) solution is obviated. If the true marginal, $\overline{f} \equiv f(u_n|\mathbf{x}_n)$, is available, then the approximation is known as the Quasi-Bayes approximation (Section 3), though, originally, this name was reserved for the further restriction, $\overline{f} \equiv \delta(u_n - \widehat{u_n})$, where $\widehat{u_n} \equiv \mathsf{E}[u_n]$, the posterior mean. In the latter case, $\tilde{f}(x_n|\mathbf{x}_{n-1},\theta) = f(x_n|\theta,\widehat{u_n})$, by the sifting property of the Dirac δ-function.

FCVB: Functionally Constrained VB. If the projection, $\tilde{f}(u_n|\mathbf{x}_n) \to \delta(u_n - \widehat{u_n})$, is performed in *each* IVB cycle, for some certainty equivalent (mode, mean, *etc*), $\widehat{u_n}$, drawn from $\tilde{f}(u_n|\mathbf{x}_n)$ (22), then the VB-observation model is again $f(x_n|\theta,\widehat{u_n})$, but now defined implicitly, and therefore requiring IVB cycles until convergence. FCVB specializes to the (Bayesian) *EM algorithm* in the case where $\widehat{u_n}$ is chosen as the mode [8, 22].

These VB variants ease the computational load of the VB approximation at each step, since they eliminate IVB cycles (RVB), and/or the need to evaluate moments of an intractable VB-marginal, $\tilde{f}(u_n|\mathbf{x}_n)$ (FCVB). The cost is an increase in the KLD associated with the variational optimizer in these constrained cases.

The functional invariance of the VB-filtering procedure, and the recursive nature of the resulting algorithm (23), is called *VB-conjugacy* [22]. The significance of the procedure springs from the fact that these recursions can be achieved for a far wider class of filtering problems than those satisfying the conditions [4] for recursiveness in exact Bayesian filtering, as we now show. Again, we will confine the presentation to the static filtering case (19), but the theory can be extended to wider classes of dynamic hidden fields, $\mathbf{u}_n$.

Definition 1 (Separable-in-Parameters EF (SIPEF)). *An observation model belongs to the separable-in-parameters exponential family (SIPEF) if*

$$f(x|\theta_1,\theta_2) \equiv a_1(\theta_1)^{a_2(\theta_2)} b(x) \exp\left[\langle c_1(\theta_1) \circ c_2(\theta_2), h(x)\rangle\right], \tag{24}$$

where $\circ$ denotes the Hadamard (element-wise) product.

Comment: An EF observation model (7) is a special case of a SIPEF model, under the assignments

$$\theta_1 = \theta; \quad a_2(\theta_2) = 1; \quad c_2(\theta_2) = 1. \tag{25}$$

Theorem 1. *A sufficient condition for VB-conjugacy is that the augmented observation model have the form:*

$$f(x_n|\theta,u_n) \equiv \prod_{j=1}^{M} \left[f_j(x_n|\theta,u_n)\right]^{l_j(u_n)}, \tag{26}$$

with (i) $f_j(x_n|\theta,u_n) \in$ SIPEF, $j = 1,\ldots,M$, and (ii) $l(u_n) \in \mathbb{E}_M$, the elementary basis in $\mathbb{R}^M$; i.e. *the hidden field, u_n, classifies x_n into one of a finite number of SIPEF classes. The following regularity condition must hold:*

$$\begin{aligned} \int_{x_n} \prod_{j=1}^{M} \left\{f_j(x_n|\theta,u_n)\right\}^{w_j} dx_n &= a^{-1}(\theta) q^{-1}(u_n,w), \\ \forall w &\in \Delta_{M-1}, \end{aligned} \tag{27}$$

where Δ_{M-1} is the $(M-1)$-dimensional probability simplex.

[1] Convergence results *are* available for the time-invariant case [18].

PROOF. Using (24) in (26), with $\theta_1 \equiv \theta$ and $\theta_2 \equiv u_n$, then

$$\begin{aligned}\ln f(x_n|\theta,u_n) = \sum_{j=1}^{M} \{ & a_{2,j}(u_n) l_j(u_n) \ln a_{1,j}(\theta) + \cdots \\ & + l_j(u_n) \langle (c_{1,j}(\theta) \circ c_{2,j}(u_n)), h_j(x_n) \rangle + \cdots \\ & + l_j(u_n) \ln b_j(x_n) \}.\end{aligned}$$

Substituting into (21):

$$\tilde{f}(x_n|\mathbf{x}_{n-1},\theta) \propto \prod_{j=1}^{M} \left\{ \tilde{f}_j(x_n|\mathbf{x}_{n-1},\theta) \right\}^{w_{j,n}}, \qquad (28)$$

where

$$\begin{aligned}\tilde{f}_j(x_n|\mathbf{x}_{n-1},\theta) = a_{1,j}(\theta)^{\mathsf{E}\left[a_{2,j}(u_n)|l(u_n)=e_j\right]} b_j(x_n). \\ . \exp\left[\langle c_{1,j}(\theta) \circ \mathsf{E}\left[c_{2,j}(u_n)|l(u_n)=e_j\right], h_j(x_n)\rangle\right].\end{aligned} \qquad (29)$$

Here, $\mathsf{E}\left[a_{2,j}(u_n)|l(u_n)=e_j\right] \equiv \mathsf{E}_{\tilde{f}(u_n|\mathbf{x}_n,l(u_n)=e_j)}\left[a_{2,j}(u_n)\right]$, and $w_n = \mathsf{E}_{\tilde{f}(u_n|\mathbf{x}_n)}[l(u_n)] \in \Delta_{M-1}$, since $\mathsf{E}_{\tilde{f}(u_n|\mathbf{x}_n)}[l_j(u_n)] = \Pr[l(u_n)=e_j|\mathbf{x}_n]$. Now, since $\tilde{f}_j(x_n|\mathbf{x}_{n-1},\theta)$ in (29) has the same functional form in x_n as $f_j(x_n|\theta,u_n)$ (26), and given the regularity condition stated in the Theorem, therefore θ enters the normalizing constant of (28) only via $a^{-1}(\theta)$. Hence, the normalized VB-observation model is

$$\tilde{f}(x_n|\mathbf{x}_{n-1},\theta) = a(\theta) q(\mathbf{x}_{n-1}) \prod_{j=1}^{M} \left\{ \tilde{f}_j(x_n|\mathbf{x}_{n-1},\theta) \right\}^{w_{j,n}}. \qquad (30)$$

The conjugate prior is designed as follows:

$$f(\theta|s_0) \propto a(\theta)^{\kappa_0} \prod_{j=1}^{M} a_{1,j}(\theta)^{\nu_{j,0}} \exp\left[\langle c_{1,j}(\theta), v_{j,0}\rangle\right]. \qquad (31)$$

Multiplying (31) by (30), and noting (29), it follows that the posterior distribution of θ is invariant, $\forall n$, being

$$f(\theta|s_n) \propto a(\theta)^{\kappa_n} \prod_{j=1}^{M} a_{1,j}(\theta)^{\nu_{j,n}} \exp\left[\langle c_{1,j}(\theta), v_{j,n}\rangle\right], \qquad (32)$$

where the sufficient statistics, $s_n \equiv \{v_n, \nu_n, \kappa_n\}$, are updated recursively, as follows:

$$\begin{aligned} v_{j,n} &= v_{j,n-1} + w_{j,n} \mathsf{E}\left[c_{2,j}(u_n)|l(u_n)=e_j\right] \circ h_j(x_n), \\ \nu_{j,n} &= \nu_{j,n-1} + w_{j,n} \mathsf{E}\left[a_{2,j}(u_n)|l(u_n)=e_j\right], \\ & \qquad j = 1,\ldots,M, \end{aligned} \qquad (33)$$

and

$$\kappa_n = \kappa_{n-1} + 1.$$

□

From (33), the data, x_n, are finitely and recursively encoded via M statistics, $v_{j,n}$, one for identification of the parameters of each class of x_n, and these are weighted by the 'soft classifications', $w_{j,n}$. Meanwhile, ν_n acts as an array of 'soft counters', while κ_n is a (hard) counter of the number of data, n, offset by hyperparameter, κ_0. It is interesting to note the presence of this extra counter, compared to standard conjugate learning (9). It is required as n is uncoded by ν_n (33).

The main additivity of the VB approximation is to provide a principled assignment of the class probabilities, $w_{j,n}$, via the VB-filtering distribution, $\tilde{f}(u_n|\mathbf{x}_n)$. The latter is evaluated iteratively via (22). Hence, one set of sufficient statistic updates (33) is required for each IVB cycle.

The VB-observation model (30) is not an EF member (7), and yet it possesses a conjugate prior (31). This recalls the sufficiency of the EF assumption for conjugacy. Its relaxation via (30) is very general. While the necessity of the structure (26) for VB-conjugacy has not been proved, it appears to include all multiple model and mixture model cases for which VB-conjugacy can be secured. One important such case follows in Section 3.4.

3.3 The Signal-Independent System (SIS) Assumption

Let us return to the system interpretation of the state-space model (Section 1), with u_n the input to a system, and

$$f(x_n|\theta,u_n) \equiv f(x_n|\theta_{u_n},u_n)$$

modelling our belief about the output, x_n, given the input. The stochastic system model is, in general, parameterized by unknown, input-dependent parameter θ_{u_n}. For example, in the M-ary channel case (Section 2.3), $\theta_{l_n} = \theta_n = \epsilon l_n$. In many engineering contexts, however, *our belief about the system is independent of the input*, in which case $\theta_{u_n} \equiv \theta$, and

$$f(x_n|\theta_{u_n},u_n) \equiv f_0(x_n|\theta,u_n), \qquad (34)$$

where $f_0(\cdot)$, is a fixed parametric function (typically we suppress the subscript in our notation). Let us adopt this assumption in Theorem 1. Then $M=1$ in (26), and $f(x_n|\theta,u_n) \in$ SIPEF. In this case, regularity condition (27) always holds, with $a(\theta)=1$. (32) then specializes to

$$f(\theta|s_n) \propto a_1(\theta)^{\nu_n} \exp\left[\langle c_1(\theta), v_n\rangle\right], \qquad (35)$$

with the sufficient statistics now updated recursively, as follows:

$$\begin{aligned} v_n &= v_{n-1} + \mathsf{E}\left[c_2(u_n)\right] \circ h(x_n), \\ \nu_n &= \nu_{n-1} + \mathsf{E}\left[a_2(u_n)\right]. \end{aligned} \qquad (36)$$

Once again, $\mathsf{E}[\cdot]$ is evaluated via $\tilde{f}(u_n|\mathbf{x}_n)$ (22), in each IVB cycle. Only statistics, v_n (along with the 'soft counter', ν_n), need to be maintained under the SIS assumption, in order to learn the parameters, θ, of the single learning mode. The recursive scheme (35) and (36) confirms that the SIPEF observation model (24) is an extension of the EF class of models (7), for which Bayesian conjugate learning remains feasible (see the Comment after Definition 1).

3.4 VB-Conjugate Inference for Finite Mixture Models (FMMs)

The certainty equivalence (CE) approach was given in Section 3 as an example of a mean field approximation for recursive learning with FMMs, assuming the mixture components are EF members. The relaxation of CE via the VB approximation is now outlined. It follows directly from Theorem 1, by recognizing that a FMM with EF components is a special case of (26).

COROLLARY 1 (OF THEOREM 1). *VB-conjugate inference is feasible for a FMM with EF components.*

PROOF. The augmented observation model in the FMM case with EF components is

$$f(x_n|\theta, u_n) = \prod_{j=1}^{M} [f_j(x_n|\theta)]^{l_{j,n}} . \tag{37}$$

Hence, $f(x_n|\theta, u_n) = f(x_n|\theta, l_n)$, and the input to the system is effectively digital (Section 1.1):

$$l(u_n) \equiv l_n \in \mathbb{E}_M.$$

We may interpret $l(u_n)$ as a *deterministic quantizer* between the input, u_n, and the system, and our epistemic modelling is then between l_n and x_n. Since each component is an EF member and, therefore, a SIPEF component (see the comment after Definition 1), (37) satisfies the conditions for VB-conjugacy given in Theorem 1, but again requiring the regularity condition (27) to be active. Then, adopting the conjugate prior (31), the sufficient statistics, $s_n \equiv \{v_n, \nu_n, \kappa_n\}$, are updated recursively, as follows:

$$\begin{aligned} v_{j,n} &= v_{j,n-1} + w_{j,n} h_j(x_n), \quad j = 1, \ldots, M, \\ \nu_n &= \nu_{n-1} + w_n, \\ \kappa_n &= \kappa_{n-1} + 1, \end{aligned} \tag{38}$$

where $w_n = \mathsf{E}_{\tilde{f}(l_n|\mathbf{x}_n)}[l_n] \in \Delta_{M-1}$, as before. □

Once again, statistics, $v_{j,n}$, are accumulated in parallel for the parameters of each component of (37), along with an accumulator of the soft classifications, ν_n, and a data counter, κ_n. In this FMM context, we typically model at the level of the pointer, l_n, as stated above. Its multinomial filtering distribution,

$$\tilde{f}(l_n|\mathbf{x}_n) = \mathcal{M}u(w_n),$$

is updated 'statically' in each IVB cycle, via (22).

Recursive algorithm (38) allows for distinct parameterization in each component, $f_j(x_n|\theta) = f_j(x_n|\theta_j)$, in which case the SIS assumption (34) does not hold. If we now impose the SIS restriction in the FMM, then we revert to (36), with, in this case, $a_2(u_n) \equiv a_2(l_n)$ and $c_2(u_n) \equiv c_2(l_n)$, since the effective input is the digital sequence, l_n, as explained above.

4. CASE STUDY: BAYESIAN INFERENCE IN THE DIGITAL RECEIVER

We consider the task of communicating a bitstream—involving unspecified coding—across a noisy channel via appropriate modulation onto a (quadrature) carrier, $e^{j\Omega t}$, where Ω is the angular frequency in rads/sec [8, 12, 15]. The bitstream is partitioned into K binary words, $\mathbf{u}_K = (u_1, \ldots, u_K)$, each of length $\log_2(M)$, so that $\mathbf{u}_K$ is an M-ary digital sequence. This must be mapped to an analogue signal via modulation of the parameters of the carrier. In quadrature amplitude modulation (QAM), each state of u_k is bijectively mapped to an appropriately-designed set of M points in the Argand plane, known as the QAM constellation, $\mathcal{S}$. We will order these into a vector of complex numbers, $\mathbf{a} = [a_1, \ldots, a_M]'$. The modulator maps each of these states to an analogue signal by keying a shift of amplitude and phase into the quadrature carrier, via the magnitude and phase of the associated point in the Argand plane: $a_j \to a_j e^{j\Omega t} w_T(t)$, where $w_T(t)$ is a windowing function, typically of length, T, equal to the signalling period of the M-ary codewords, u_k. At the receiver, the concatenated signals are (uniformly) sampled, period $T_s = T/N$. Denoting the digital frequency of the sampled carrier by $\omega = \Omega T_s$ (rads/sample), an appropriate Wold representation of the data block associated with the kth transmitted symbol, $k = 1, \ldots, K$, is therefore

$$x_{k,i} = (l_k' \mathbf{a})\, e^{j\omega(i+(k-1)N)+\theta} + e_i, \quad i = i, \ldots, N. \tag{39}$$

$x_{k,i}$ is the ith element of the kth length-N data vector, x_k, concatenated into the complete length-NK data vector, $\mathbf{x}$. In (39), we have made the following simplifying assumptions for the purpose of this presentation: (i) $e_i \overset{\text{ciid}}{\sim} \mathcal{C}N(0, r)$, the additive white complex Gaussian channel noise assumption, with variance (power) r, assumed known; and (ii) uncertainty in the carrier parameters is confined to its phase, θ. The task of the decoder in the receiver is to infer the pointer sequence, $\mathbf{l}$, of the transmitted sequence (the *decoding* task), accounting for uncertainty in the carrier phase (the *phase synchronization* task). The epistemic inference problem is therefore to compute the decoding distribution, $f(\mathbf{l}|\mathbf{x})$, and the synchronization distribution, $f(\theta|\mathbf{x})$.

The joint parametric model is

$$f(\mathbf{x}, \mathbf{l}, \theta|r, p) = \prod_{k=1}^{K} f(x_k|l_k, \theta) f(l_k|p) f(\theta),$$

where we have made the simplifying assumption (whose relaxation we discuss in Section 5) of independent u_k, with invariant state probabilities, $p = (p_1, \ldots, p_M)'$. Hence:

$$f(l_k|p) \equiv \mathcal{M}u(1, p) = l_k' p.$$

The (augmented) observation model is, from (39):

$$f(x_k|l_k, \theta) = \mathcal{C}N(g_k l_k' \mathbf{a} e^{j\theta}, rI_N), \tag{40}$$

where $g_k = [e^{j\omega(1+(k-1)N)}, \ldots, e^{j\omega kN}]'$, the kth synchronized carrier block. I_N denotes the $N \times N$ identity matrix, and we will suppress conditioning on constants in the notation in the sequel. It follows that the (marginal) observation model, $f(x_k|\theta)$, is a finite mixture of M Gaussian components, of the type in (37). Furthermore, it satisfies the signal-independent system (SIS) assumption of Section 3.3. We will treat the decoding problem recursively as a *Bayesian filtering* problem, attempting to construct $f(l_k|\mathbf{x}_k)$, where $\mathbf{x}_k \equiv [x_1', \ldots, x_k']'$.

4.1 The Receiver VB-Observation Model

From (40), the observation model for the kth data block may be expressed as

$$\ln f(x_k|l_k, \theta) \propto \frac{2}{r} \Re\left\{ x_k^H g_k l_k' \mathbf{a} e^{j\theta} \right\} - \frac{1}{r} l_k' (\mathbf{a} \circ \mathbf{a}) g_k^H g_k - \frac{1}{r} x_k^H x_k. \tag{41}$$

Inserting into (21), the VB-observation model is easily found to be

$$\tilde{f}(x_k|\theta) = \mathcal{N}(g_k w_k' \mathbf{a} e^{j\theta}, rI_N), \tag{42}$$

exhibiting the implicit dependence on moments of the VB-marginal of l_k, via $w_k = \mathsf{E}_{\tilde{f}(l_k|x_k)}[l_k]$. Multiplying (41) by

$f(l_k) \equiv l_k p$ yields the augmented observation model, and using this in (21), we obtain

$$\tilde{f}(l_k|x_k) \propto l'_k w_k \equiv \mathcal{M}u(w_k), \tag{43}$$

where

$$w_k \propto \exp\left\{\frac{1}{r}\left[2\Re\{x_k^H g_k \widehat{e^{j\theta}}\mathbf{a}\} - g_k^H g_k(\mathbf{a}\circ\mathbf{a})\right]\right\} \circ p, \tag{44}$$

normalized such that $w_k \in \Delta_{M-1}$. This depends implicitly on $\tilde{f}(\theta|\mathbf{x}_k)$—yet to be determined—via circular moment, $\widehat{e^{j\theta}}$.

Since (40) is a SIPEF member (Theorem 1), the VB-observation model is certainly an EF member. Comparing (41) with (7), the conjugate prior is now *designed* as

$$f(\theta|\mathbf{x}_{k-1}) \propto \exp\left[\Re\{\kappa_{k-1}e^{j\theta}\}\right]. \tag{45}$$

This is the *von Mises* distribution [10, 15], with support on the unit circle, assuming θ is the principal value of the carrier phase. A key parametric model in directional statistics [10], it was derived in [15] as the conjugate prior for learning with phase-uncertain data. This is the first time that it has been derived as a VB-vonjugate prior. Its normalizing constant is $2\pi I_0(\kappa_{k-1})$, the modified Bessel function of the first kind, order 0. Multiplying (45) by (42), the designed recursive block update of sufficient statistics is revealed:

$$\kappa_k = \kappa_{k-1} + \frac{2}{r}x_k^H g_k w'_k \mathbf{a}. \tag{46}$$

This single-statistic update is in agreement with the general recursive structure derived for a FMM under the SIS constraint (36).

- The update is in respect of the DTFT, $x_k^H g_k$, of the data block for the kth symbol, evaluated at the known carrier frequency, ω (39). It is evaluated efficiently via the Goertzl algorithm [15].
- The assumption of a signal-independent channel leads to conflation of the sufficient statistics into a single scalar accumulator, κ_k, as noted in Section 3.3. The posterior weights, w_k, (*i.e.* the filtering/decoding distribution (43) are deduced in a principled way, rather than using informal classification concepts.
- The requirement for w_k to be evaluated *before* the data recursion can be performed in (46) obviates a recursive evaluation of κ_k within each block.
- The circular moment of $f(\theta|\mathbf{x}_k)$ required in evaluation of $\tilde{f}(l_k|x_k)$ (44) is available tractably as

$$\widehat{e^{j\theta}} = \frac{I_1(\kappa_k)}{I_0(\kappa_k)}e^{j\angle\kappa_k}.$$

- The computational overhead of iterating the IVB algorithm for each block, x_k, is sustainable, as convergence occurs after very few cycles in practice. The net cost is far less than full evaluation of $f(\theta|\mathbf{x}_K)$, which requires maintenance of M^K statistics (Section 3). Experimental evidence [23] also suggests that a particle filter solution has a far higher computational requirement in order to achieve the same accuracy as the VB solution outlined here.

5. DISCUSSION

The state-of-the-art in the iterative (turbo) receiver is to synchronize via 'soft decisions' regarding the transmitted symbols (soft-decision-directed (SDD) synchronization). It has been shown to be an application of the EM algorithm [8]. Its relaxation via VB yields a symmetric algorithm, since 'soft information' about phase—*i.e.* $\tilde{f}(\theta|\mathbf{x}_k)$—is available at each step in order to decode the transmitted symbols (via $\tilde{f}(l_k|\mathbf{x}_k)$). This can be referred to as soft-synchronization-directed (SSD) decoding [14]. The discovery, here, of the von Mises distribution as the VB-learning invariant for phase synchronization is important for several reasons. Our main emphasis has been on its role in assuring a recursive computational structure. Another advantage is that it provides a key degree-of-freedom for the Bayesian designer, being the hyperparameter of the prior, κ_0. Since a structured prior for phase has not been used in the literature, the default [12] has been to use a uniform prior, equivalent to the von Mises prior, with $\kappa_0 = 0$. The QAM decoding problem in the phase-uncertain channel is susceptible to gross errors due to rotational invariance effects, and many practical proposals for countering this problem have been advanced. The use of a non-uniform von Mises prior overcomes this problem elegantly [3]. From (46), κ_0 may be interpreted as the DTFT of externally-available or fictitious data.

In other recent work [3], the assumption of known channel noise, via r, has been relaxed, improving the performance of the receiver when compared to extant solutions which condition on a certainty equivalent. Once again, a recursive structure is achieved, but at the cost of evaluating expensive hypergeometric functions in each IVB cycle. Further work is required to implement a full VB-based synchronization for other possible modulator uncertainties, such as timing offset in the symbol transitions, and drift in the oscillator frequency.

The Viterbi algorithm is the standard computational technology for point estimation of hidden Markov chains. The work reported in this paper differs fundamentally in motivation in that it is a Bayesian strategy, solving the epistemic inverse problem by computing posterior state probabilities. Nevertheless, the FCVB technique (Section 3.2) can also be used to design an iterative procedure for evaluation of a posterior certainty equivalent, $\hat{\mathbf{u}}_k$, with performance-computation trade-offs that compare well with Viterbi [21].

6. CONCLUSION

The VB approximation has an important role to play in restoring tractability in the epistemic inverse problem. In this paper, it has been shown that elegant recursive computational procedures result from application of VB as a local approximation at each time step.The class of augmented observation models for which this is possible was defined, and shown to be a potentially rich class. A core principle was to identify the VB-observation model and design the prior to be conjugate to it. While the theory, and the reported case study in digital communication, assumed a static (ciid) input/state variable, the same principle readily applies under a Markov assumption, as reported in [22] for the discrete case and [23] for the continuous case. Being a local approximation, the main concern relates to asymptotic performance of the algorithm, as compared to stochastic techniques such as

particle filters. Nonetheless, the VB approach offers attractive trade-offs in computationally constrained environments such as mobile and embedded platforms. Efforts to quantify these trade-offs are ongoing. Another area of active investigation is the design of parameter transformations to reduce the KLD associated with the approximation [20].

7. REFERENCES

[1] J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition, 1985.

[2] J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 2 edition, 1997.

[3] A. Das and A. Quinn. Variational Bayes extension to turbo-synchronization and phase ambiguity resolution. In *Proc. Irish Sig. and Sys. Conf.*, Dublin, 2011.

[4] E. Daum. New exact nonlinear filters. In J. Spall, editor, *Bayesian Analysis of Time Series and Dynamic Models*. Marcel Dekker, New York, 1988.

[5] B. de Finetti. *Theory of Probability*, volume 2. Wiley, 1975.

[6] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[7] R. Gray. *Probability, Random Processes, and Ergodic Properties*. Springer, 2nd edition, 2009.

[8] C. Herzet, N. Noels, V. Lottici, H. Wymeersch, M. Luise, M. Moeneclaey, and L. Vandendorpe. Code-Aided Turbo-Synchronization. *Proc. IEEE*, 95(6):1–17, June 2007.

[9] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, London, 2005.

[10] C. G. Khatri and K. V. Mardia. The von Mises-Fisher distribution in orientation statistics. *Journal of Royal Statistical Society B*, 39:95–106, 1977.

[11] A. Mohammad-Djafari, editor. *Inverse Problems in Vision and 3D Tomography*. Wiley, 2010.

[12] N. Nissilä and S. Pasupathy. Adaptive iterative detectors for phase-uncertain channels via variational bounding. *IEEE Trans. Comms*, 57(3):716–725, 2009.

[13] M. Opper and D. Saad. *Advanced Mean Field Methods: Theory and Practice*. The MIT Press, Cambridge, Massachusetts, 2001.

[14] A. Quinn. On-line statistical inference for telecommunications using the Variational Bayes approximation. Technical report, Trinity College Dublin, 2008.

[15] A. Quinn, J.-P. Barbot, and P. Larzabal. The Bayesian inference of phase. In *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Process.*, Prague, 2011.

[16] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

[17] E. Ranguelova and A. Quinn. Analysis and synthesis of three-dimensional Gaussian Markov random fields. In *Proc. IEEE Int. Conf. on Image Process. (ICIP)*, Kobe, Japan, 1999.

[18] M. Sato. Online model selection based on the variational Bayes. *Neural Computation*, 13:1649–1681, 2001.

[19] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixtures*. John Wiley, New York, 1985.

[20] V.-H. Tran and A. Quinn. The transformed variational Bayes approximation. In *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Process.*, Prague, 2011.

[21] V.-H. Tran and A. Quinn. Variational Bayes variants of the Viterbi algorithm. In *Proc. Irish Sig. and Sys. Conf.*, Dublin, 2011.

[22] V. Šmídl and A. Quinn. *The Variational Bayes Method in Signal Processing*. Springer, 2006.

[23] V. Šmídl and A. Quinn. Variational Bayesian filtering. *IEEE Transactions On Signal Processing*, (10):5020–5030, 2008.