Journal of Machine Learning Research 10 (2009) 935-976 Submitted 8/07; Revised 11/08; Published 4/09

Perturbation Corrections in Approximate Inference: Mixture Modelling Applications

Ulrich Paquet ULRICH@CANTAB.NET
Computer Laboratory
University of Cambridge
Cambridge CB3 0FD, United Kingdom

Ole Winther OWI@IMM.DTU.DK
Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark

Manfred Opper OPPERM@CS.TU-BERLIN.DE
Computer Science
TU Berlin
D - 10587 Berlin, Germany

Editor: Zoubin Ghahramani

Abstract

Bayesian inference is intractable for many interesting models, making deterministic algorithms for approximate inference highly desirable. Unlike stochastic methods, which are exact in the limit, the accuracy of these approaches cannot be reasonably judged. In this paper we show how low order perturbation corrections to an expectation-consistent (EC) approximation can provide the necessary tools to ameliorate inference accuracy, and to give an indication of the quality of approximation without having to resort to Monte Carlo methods. Further comparisons are given with variational Bayes and parallel tempering (PT) combined with thermodynamic integration on a Gaussian mixture model. To obtain practical results we further generalize PT to temper from arbitrary distributions rather than a prior in Bayesian inference.

Keywords: Bayesian inference, mixture models, expectation propagation, expectation consistent, perturbation correction, variational Bayes, parallel tempering, thermodynamic integration

1. Introduction

Approximate methods for Bayesian inference have recently enjoyed a limelight of attention. These methods can be either deterministic or stochastic. Deterministic methods, which typically turn integration and summation problems of Bayesian marginalization into optimization problems, include the Laplace approximation, mean field (or variational) methods like variational Bayes (VB), expectation propagation (EP), and expectation consistent (EC) and Bethe/Kikuchi approximations (also known as loopy belief propagation or generalized belief propagation). Their attraction lies in the precise but tractable inferences that they typically provide, but their drawback is the lack of a built-in sanity check, as we cannot assess the approximation error. Stochastic methods like Markov chain Monte Carlo (MCMC)

algorithms, which give exact estimates in a large enough sample limit, lie orthogonal to deterministic methods. They are normally much slower than their deterministic counterparts, but given a skilled user and enough computational resources stochastic methods are capable of giving more precise answers. Whether inference errors (of unknown size) are acceptable of course depends on the application in question. In statistical applications one might prefer simple models which allow for exact inferences, whereas in communication systems intractability is an inherent property of communication channels and to counter this, one instead designs fault tolerant error-correcting protocols.

The problem under consideration can be stated in general terms: We are presented with a data set of N independent and identically distributed (i.i.d.) examples $\mathcal{D} = \{x_n\}_{n=1}^N$, which we model by a generative model specified by the distribution $p(x|\theta)$, such that $p(\mathcal{D}|\theta) = \prod_n p(x_n|\theta)$. In Bayesian inference we introduce a prior distribution $p(\theta)$ over model parameters θ, and to infer unobserved random variables we compute different averages over the posterior distribution

$$p(\theta|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}|\theta)p(\theta) \quad \text{with} \quad Z = \int d\theta \, p(\mathcal{D}|\theta)p(\theta) \; . \tag{1}$$

In model selection or model averaging the normalizer (marginal likelihood) $Z = p(\mathcal{D})$ needs to be computed for different models under consideration, that is, $p(\mathcal{D}|\mathcal{M}_m)$, $m = 1, \ldots, |\mathcal{M}|$. Another central inference is about the density at a new (test) example, the so-called predictive density (or distribution):

$$p(x|\mathcal{D}) = \int d\theta \, p(x|\theta)p(\theta|\mathcal{D}) \; . \tag{2}$$

This paper mostly specializes to modelling the density with a mixture model

$$p(x|\theta) = \sum_{k=1}^{K} p(k)p(x|\theta_k)$$

such that mixing proportions $p(k)$ sum to one, and $\theta = \{p(k), \theta_k\}_{k=1}^K$. A mixture of Gaussians (MoG) corresponds to $p(x|\theta_k)$ being Gaussian. The prior distribution and the likelihood term for each component term $p(k)p(x|\theta_k)$ are chosen to be conjugate, such that their product is in the same distribution family as the prior and thus tractable. Intractability for the mixture model arises not because integration is intractable, but because the number of terms in the marginal likelihood is K^N.

This paper starts from the vantage point of an expectation consistent (EC) approximation (Opper and Winther, 2005) (and its algorithmic realisation by expectation propagation (EP) Minka 2001a) and substantiates these main contributions and findings:

1. We express the exact posterior distribution by an approximating distribution which is given by EC plus a series of error terms with increasing complexity. When low order corrections are small, one might hope that the remaining contributions will also decrease with the order, suggesting that the approximation can be improved by retaining only the lowest orders in the series. One can thus expect corrections to improve an already good approximation, but not a poor one. On the other hand, large lower order terms may indicate a poor approximation, providing an error check on the approximation without having to resort to Monte Carlo methods.

2. We derive corrections both for the marginal likelihood and the predictive distribution in the form of an expansion in terms of "clusters" of likelihood terms of the posterior. This expansion resembles the loop series expansions which were derived for correcting loopy belief propagation (LBP) (Chertkov and Chernyak, 2006, Gómez et al., 2007, Sudderth et al., 2008).[1] All these methods hold in common that the correction terms are expressed as averages over the approximating solution and can thus be calculated *after* the convergence of the EP or LBP iterative scheme.

3. We show that our first order correction to the posterior can be simply expressed by quantities already computed by the EP algorithm. No further averages are needed. In contrast, the lowest non-trivial correction to the marginal likelihood is of second order, with the number of terms growing as $\mathcal{O}(N^2)$. Corrections to the marginal likelihood can be tractably computed for example, for models where the likelihood is a mixture distribution. Each of error terms contain the original K-component mixture, such that a correction up to order j requires the computation of $\mathcal{O}((NK)^j)$ terms.

4. When the true distribution is multi-modal, EP will in most cases provide a local (single) mode approximation, with lower-order corrections also being local. One such example is the $K!$-fold labelling symmetry of the latent space of mixture models, which may cause $\mathcal{O}(K!)$ separated modes in the posterior distribution. While the predictive distribution is invariant to this symmetry, the log marginal likelihood usually has to be further corrected by a factor of $\mathcal{O}(\log K!)$, a correction that is typically much larger than a low-order perturbation correction.

5. Thorough empirical tests of EP validate its precision, and show errors that do not scale with N. The perturbation corrected predictions are almost uniformly more precise than EP. As a tool for improving inference accuracy, we show in a practical example that the first nontrivial correction term to the marginal likelihood approximation can make a clear difference in predicting which K maximizes the marginal likelihood, compared to when the correction was not used.

In this paper EC or EP and its resulting corrections are compared with variational Bayes (VB), Minka's α-divergence message passing scheme, and a gold standard benchmark of parallel tempering (PT) and thermodynamic integration (TI). PT is a Markov chain Monte Carlo (MCMC) method whose Markov chain operates on a "tempered posterior" and has very good convergence properties. Contrary to more standard Monte Carlo methods (for example Metropolis-Hastings or Gibbs sampling) it can also provide estimates of the marginal likelihood by TI, which interpolates the expected value of the log likelihood between the prior and the posterior. To increase the stability of estimates obtained by TI, we give a novel generalization of PT, which allows interpolation of the value of the log likelihood between *any* choice of distribution and the posterior. A good choice may also improve sampling when the tempered posterior exhibits phase transition-like properties. This choice might be obtained by some deterministic approximation, and although not in-

1. An information geometrical expansion for LBP is given by Ikeda et al. (2004), and for EP by Matsui and Tanaka (2008). LBP can also be improved with a message passing algorithm that corrects for the influence of loops (Mooij et al., 2007).

vestigated in this paper, provides a springboard for combining deterministic and stochastic inference algorithms.

As a further example it is also shown how the "cluster" perturbation expansion can be applied to Gaussian Process classification models, where the evaluation of integrals for Bayesian marginalization are not analytically tractable.

The rest of the paper follows with a description of EC and EP in Section 2. Section 3 shows an example of corrections for a marginal distribution in a Gaussian Process classification model. In Section 4 an inference algorithm is presented for mixture weights, that is, a mixture model with fixed component densities, while Appendix D treats the fully multivariate MoG. Section 5 contains short descriptions of PT with TI and a generalization suitable for statistical inference. Results are presented for real world examples in Section 6, and we conclude in Section 7.

2. Expectation Consistent Inference

The *expectation consistent* approximation provides a framework for finding a surrogate distribution $q(\theta)$ for $p(\theta|\mathcal{D})$ in Bayesian inference (Opper and Winther, 2005).[2] The message passing scheme of *expectation propagation* gives rise to an identical marginal likelihood approximation, and the following interpretation sheds light on both methods by looking at them as a set of self-consistent approximations to marginal or predictive distributions. The outline presented here allows for further *perturbation corrections* to be derived.

For the purpose of this paper the EC approximation rests on the observation that the predictive density $p(x|\mathcal{D})$ in (2) can be fairly precisely approximated without averaging over the actual posterior. The entire posterior can be replaced with a simpler distribution $q(\theta)$ *if* it produces the correct statistics for this average, that is,

$$p(x|\mathcal{D}) = \int d\theta\, p(x|\theta)p(\theta|\mathcal{D}) \approx \int d\theta\, p(x|\theta)q(\theta)\ .$$

It is sufficient for $q(\theta)$ to share some key properties, namely low order statistics, with $p(\theta|\mathcal{D})$. This is an ambitious demand that is generally not realizable, but we can transfer the principle of moment matching to the "cavity" posteriors $p(\theta|\mathcal{D}_{\backslash n})$, which correspond to reduced training sets $\mathcal{D}_{\backslash n}$ where the n^{th} example has been left out. By introducing a similar approximation to the "cavity" predictive distributions

$$p(x_n|\mathcal{D}_{\backslash n}) = \int d\theta\, p(x_n|\theta)p(\theta|\mathcal{D}_{\backslash n}) \approx \int d\theta\, p(x_n|\theta)q_{\backslash n}(\theta)$$

for each x_n in the training set, a computationally efficient approximation can be derived. We shall now *rather* require $q(\theta)$ to share key properties, namely lower order statistics, with *each* of the distributions $q_n(\theta) \propto p(x_n|\theta)q_{\backslash n}(\theta)$; this is explored in the next section.

2.1 EC and EP with Exponential Families

EC defines a tractable approximation $q(\theta)$ through expectation consistency with each $q_n(\theta)$. Our view of EC shall be narrowed to models factorizing in likelihood terms $p(x_n|\theta)$, and an

2. A more general interpretation is possible, but for clarity we show the approximation for the generative model in (1).

exponential family prior

$$p(\theta) = \frac{1}{Z_0} \exp\Big(\Lambda_0^T \phi(\theta)\Big) h(\theta) \ ,$$

where Z_0 is the normalizing constant, $\phi(\theta)$ is a fixed vector of the corresponding sufficient statistics—for example for a univariate Gaussian we can choose $\phi(\theta) = \big(\theta, -\theta^2/2\big)$, Λ_0 is the associated parameter vector and the fixed function $h(\theta)$ encodes additional constraints (positivity, normalizations, etc.). The desired quality of approximation, and the possible convenience of obtaining tractable moments, typically guide the choice of $\phi(\theta)$.

The posterior will be approximated with a tractable density of the same exponential family as the prior,

$$q(\theta) = \frac{1}{Z(\Lambda, 0)} \exp\Big(\Lambda^T \phi(\theta)\Big)\, p(\theta) \ . \tag{3}$$

By adding the condition $\Lambda = \sum_n \Lambda_n$, we allow each likelihood factor $p(x_n|\theta)$ of the posterior in (1) to correspond to simpler factor proportional to $\Lambda_{\backslash n} = \Lambda - \Lambda_n$ in (3): the Λ_n's therefore parameterize the likelihood term contributions to the approximation.[3] We here introduced a definition for normalization as

$$Z(\Lambda, a) = \int d\theta \prod_n \Big[p(x_n|\theta)\Big]^{a_n} \exp\Big(\Lambda^T \phi(\theta)\Big)\, p(\theta) \ ,$$

with a being a vector with elements a_n. The cavity posterior $p(\theta|\mathcal{D}_{\backslash n})$ should then be approximated by a member of the same exponential family

$$q_{\backslash n}(\theta) \propto \exp\Big(\Lambda_{\backslash n}^T \phi(\theta)\Big) p(\theta) \ ,$$

where $\Lambda_{\backslash n} = \Lambda - \Lambda_n$. This is obtained from (3) by removing a single likelihood approximation and renormalizing.

Let 1_n be a unit-vector in the n^{th} direction. We can now formalize our concluding remark: $q(\theta)$ is required to share lower order statistics with the *tilted* distributions

$$q_n(\theta) = \frac{1}{Z(\Lambda - \Lambda_n, 1_n)} p(x_n|\theta) \exp\Big((\Lambda - \Lambda_n)^T \phi(\theta)\Big)\, p(\theta) \ , \tag{4}$$

each of which are obtained from the posterior $p(\theta|\mathcal{D})$ by replacing the cavity posterior by its approximation. We therefore require consistency of the generalized moments, that is,

$$\big\langle \phi(\theta) \big\rangle_q = \big\langle \phi(\theta) \big\rangle_{q_n} \ , \qquad n = 1, \ldots, N \ .$$

One can also show that the corresponding marginal likelihood approximation is given by Minka (2005) and Opper and Winther (2005)

$$Z_{\text{EC}} = Z(\Lambda, 0) \prod_n \frac{Z(\Lambda - \Lambda_n, 1_n)}{Z(\Lambda, 0)} \ . \tag{5}$$

In Appendix A we relate this approximation to variational bounds on the marginal likelihood.

3. In this context the likelihood terms (factors) are sometimes referred to as *sites*, and hence the Λ_n's as *site parameters* of *site functions* that are proportional to $\exp(\Lambda_n^T \phi(\theta))$ (Seeger, 2003).

2.1.1 Expectation Propagation

The final expression for the EC partition function in (5) depends upon the partition functions for two distributions q and q_n in (3) and (4), and consistency on the statistics $\phi(\theta)$ determines the Λ_n parameters. This moment consistency can be achieved via a message passing framework called EP, which appear, together with VB,[4] as special cases of a more generic message passing framework recently proposed by Minka (2005). EP defines a specific message algorithm which iteratively refines each Λ_n by minimising local Kullback-Leibler divergences $\mathrm{KL}(q_n(\theta)\|q(\theta))$; in other words it iteratively performs the required moment matching $\langle\phi(\theta)\rangle_q = \langle\phi(\theta)\rangle_{q_n}$. EP is presented in Algorithm 1 for our choice of q and q_n, and we shall henceforth use the terms EP and EC interchangeably.

If EP converges we will have expectation consistency $\langle\phi(\theta)\rangle_{q_n(\theta)} = \langle\phi(\theta)\rangle_{q(\theta)} = \mu$ because of the moment matching in lines 4 and 5 of Algorithm 1. Line 6 ensures that q and q_n follow the forms in (3) and (4). Solving for q in line 5 is analytical for most of the parameters as long as q is in the exponential family. (In the mixture of Gaussian examples in this paper, one has to solve two independent scalar non-linear equations for Dirichlet and Wishart densities. All other vector and matrix parameters can be found analytically.)

EP is not guaranteed to converge, in which case double-loop algorithms may be used. It has been observed by Heskes and Zoeter (2002) that when EP does not converge to a stable fixed point, even when considerable damping (choosing γ small in Algorithm 1) is used, the corresponding double-loop algorithm has a Hessian with a significantly negative eigenvalue(s). It has been suggested that the failure of convergence of canonical EP usually implies an inaccurate solution, with the choice of approximating family not being rich enough (Minka, 2001a).

2.2 Perturbation Corrections

The goal of this section is to derive formal expressions for the errors of the EC approximation to the marginal likelihood and the predictive distribution and to discuss ways of how this error can be computed using a formal perturbation expansion. In order to expand the EC approximation we use (4) to express each likelihood term by the approximating densities as

$$p(x_n|\theta) = \frac{Z(\Lambda - \Lambda_n, 1_n)}{Z(\Lambda, 0)} \frac{q_n(\theta)}{q(\theta)} \exp\left(\Lambda_n^T \phi(\theta)\right) ,$$

to find that

$$p(\theta) \prod_n p(x_n|\theta) = Z_{\mathrm{EC}}\ q(\theta) \prod_n \left(\frac{q_n(\theta)}{q(\theta)}\right) . \tag{6}$$

If we define

$$\varepsilon_n(\theta) = \frac{q_n(\theta) - q(\theta)}{q(\theta)}$$

4. VB finds its approximation $q(\theta)$ by lower-bounding the log marginal likelihood with Jensen's inequality (Jordan et al., 1999), giving $\log Z_{\mathrm{VB}} \leq \log p(\mathcal{D})$. By writing

$$\log Z_{\mathrm{VB}} = -\mathrm{KL}(q(\theta)\|p(\theta|\mathcal{D})) + \log p(\mathcal{D})$$

the bound can be made as tight as possible by adjusting $q(\theta)$; this is achieved by minimizing the KL-divergence between $q(\theta)$ and $p(\theta|\mathcal{D})$.

Algorithm 1 EP message passing (Minka, 2001a)

1: **initialize:** Set all Λ_n to zero, $\Lambda_n \leftarrow 0$, $n = 1, \ldots, N$. This choice corresponds to initializing in the prior, setting the sufficient statistics to $\mu \leftarrow \langle \phi(\theta) \rangle_{p(\theta)}$.
2: **repeat**
3: Randomly choose example n, and make the following update steps:
4: Update sufficient statistics

$$\mu \leftarrow \left\langle \phi(\theta) \right\rangle_{q_n(\theta;\Lambda_n)} .$$

5: Determine $q(\theta;\Lambda)$ from μ, that is, solve

$$\left\langle \phi(\theta) \right\rangle_{q(\theta;\Lambda')} = \mu$$

with respect to Λ' and update

$$\Delta\Lambda \leftarrow \Lambda' - \Lambda \quad \text{followed by} \quad \Lambda \leftarrow \Lambda + \Delta\Lambda .$$

The EP updates can also be damped by $\gamma \in [0,1]$ *through* $\Delta\Lambda \leftarrow \gamma(\Lambda' - \Lambda)$.
6: Update $q_n(\theta;\Lambda_n)$:

$$\Lambda_n \leftarrow \Lambda_n + \Delta\Lambda .$$

This update ensures that $\Lambda = \sum_n \Lambda_n$; q and q_n are therefore in the forms of (3) and (4). We have no guarantee in this step that q_n stays a proper distribution. A robust heuristic is to skip any update that makes q_n improper.
7: **until** expectation consistency $\langle \phi(\theta) \rangle_{q_n(\theta;\Lambda_n)} = \langle \phi(\theta) \rangle_{q(\theta;\Lambda)} = \mu$ holds for $n = 1, \ldots, N$.

such that $\frac{q_n(\theta)}{q(\theta)} = 1 + \varepsilon_n(\theta)$, we should expect $\varepsilon_n(\theta)$ to be on average small over a suitable measure when the EC approximation works well. Bearing this definition in mind, the exact posterior and the exact marginal likelihood can be written as

$$p(\theta|\mathcal{D}) = \frac{1}{R}\, q(\theta) \prod_n (1 + \varepsilon_n(\theta)) \qquad \text{and} \qquad Z = Z_{\text{EC}}\, R\, , \tag{7}$$

with

$$R = \int d\theta\, q(\theta) \prod_n (1 + \varepsilon_n(\theta)) \ .$$

We expect that an expansion of posterior and Z in terms of $\varepsilon_n(\theta)$ truncated at low orders might give the dominant corrections to EC. Hence, we get the (2^N term finite) expansion

$$R = 1 + \sum_{n_1<n_2} \left\langle \varepsilon_{n_1}(\theta)\varepsilon_{n_2}(\theta) \right\rangle_q + \sum_{n_1<n_2<n_3} \left\langle \varepsilon_{n_1}(\theta)\varepsilon_{n_2}(\theta)\varepsilon_{n_3}(\theta) \right\rangle_q + \ldots \ , \tag{8}$$

showing that EC is correct to the first order as the term $\sum_n \langle \varepsilon_n(\theta) \rangle_q = 0$ vanishes. The posterior in (7) can be similarly expanded with

$$p(\theta|\mathcal{D}) = \frac{q(\theta)\ \left(1 + \sum_n \varepsilon_n(\theta) + \sum_{n_1<n_2} \varepsilon_{n_1}(\theta)\varepsilon_{n_2}(\theta) + \ldots\right)}{1 + \sum_{n_1<n_2} \langle \varepsilon_{n_1}(\theta)\varepsilon_{n_2}(\theta) \rangle_q + \ldots} \ , \tag{9}$$

where we should keep as many terms in the numerator as in the denominator in order to keep the resulting density normalized to one.

The corresponding predictive distribution is

$$p(x|\mathcal{D}) = \int d\theta\, p(x|\theta)\, p(\theta|\mathcal{D}) \\ = \frac{\int d\theta\, q(\theta)\, p(x|\theta)\, \left(1 + \sum_n \varepsilon_n(\theta) + \sum_{n_1<n_2} \varepsilon_{n_1}(\theta)\varepsilon_{n_2}(\theta) + \ldots\right)}{1 + \sum_{n_1<n_2} \langle \varepsilon_{n_1}(\theta)\varepsilon_{n_2}(\theta)\rangle_q + \ldots}\,, \tag{10}$$

where again as many terms in the numerator as in the denominator should be kept to ensure proper normalization.

If the expansions in (9) and (10) are truncated, the approximations are not guaranteed to be valid probability distributions, since as functional approximations they may be negative. Nevertheless, the quality of EC approximation is still improved, as is illustrated in Figures 1, 8, 12, and Table 1.

2.3 Tractability of Corrections

For the case where q_n is just a finite *mixture* of K simpler densities from the exponential family to which q belongs, then the number of mixture components in the j-th term of the expansion of R is just of the order $\mathcal{O}(K^j)$ and an evaluation of low order terms is tractable and can be computed in $\mathcal{O}((KN)^j)$ after q has been found.

In other cases, an exact computation of even the low order terms may be analytically intractable. If the dimensionality of necessary integrations is proportional to the order of the correction one may still resort to numerical quadratures. A different approach would be to re-expand each term ε_n in a different "measure of closeness" of densities which takes into account the moments $\phi(\theta)$ of the densities. This can be for example achieved in the case where $q(\theta)$ is Gaussian and the statistics $\phi(\theta)$ denote just the set of all first and a subset of second moments (or cumulants) of the random variable θ. Then we could resort to the use of characteristic functions $\chi(\kappa)$ and $\chi_n(\kappa)$ defined through

$$q(\theta) = \int d\kappa\; e^{i\kappa^T\theta}\chi(\kappa), \qquad q_n(\theta) = \int d\kappa\; e^{i\kappa^T\theta}\chi_n(\kappa)$$

for all n. The coefficients in a formal multivariate Taylor expansion of $\log\chi_n(\kappa)$ in powers of the vector κ define (up to a factor) the *cumulants* of q_n. Hence, the multivariate Taylor expansion of $r_n(\kappa) \equiv \log\chi_n(\kappa) - \log\chi(\kappa)$ in powers of κ contains only those cumulants in which q_n and q *differ*. Thus, we may write

$$q_n(\theta) - q(\theta) = \int d\kappa\; e^{i\kappa^T\theta}\chi(\kappa)\left(1 - e^{\log\left(\frac{\chi_n(\kappa)}{\chi(\kappa)}\right)}\right) = \int d\kappa\; e^{i\kappa^T\theta}\chi(\kappa)\left(1 - e^{r_n(\kappa)}\right) \tag{11} \\ = -\int d\kappa\; e^{i\kappa^T\theta}\chi(\kappa)\left(r_n(\kappa) + \frac{1}{2}r_n^2(\kappa) + \ldots\right)\,.$$

Hence, when the statistics $\phi(\theta)$ contain *all* first and *all* second moments of θ, the integral is expressed through cumulants of order 3 and higher. In this way the error of the EC approximation can be expressed in terms of higher order cumulants.

If we expand r_n in powers of κ, it is possible to express the integral (11) explicitly in a series containing derivatives of increasing order of the Gaussian $q(\theta) = \int d\kappa \; e^{i\kappa^T\theta}\chi(\kappa)$ with respect to θ. This is because each such derivative creates a factor κ in the Fourier integral via differentiations of the exponential $e^{i\kappa^T\theta}$. Finally, each term $\varepsilon_n(\theta) = \frac{q_n(\theta)-q(\theta)}{q(\theta)}$ can then be expressed by a series of Hermite polynomials in a standard way. This alternative expansion is introduced by Opper et al. (2008); its details and applications will be presented in a future paper.

2.4 First Order Correction

We have seen that in general, higher order correction terms require the computation of extra expectations. Remarkably, in contrast, the first order correction to the EC posterior (9) is obtained as simple sum of terms which where already computed in the EC approximation. Hence, it provides a simple and efficiently computable quantity to improve on EC/EP or judge its validity. A straightforward calculation gives

$$p(\theta|\mathcal{D}) \approx \sum_n q_n(\theta) - (N-1)q(\theta) \; . \tag{12}$$

The first order correction does *not* change the moments which are consistent in EC, but provides an approximation to nontrivial higher cumulants, which, for example, in the case of a Gaussian $q(\theta)$ would be *zero* in EC.

3. Gaussian Process Classification

The cluster expansion can be applied in a limited setting to non-parametric models with a Gaussian process prior. This section provides as an introductory case a correction to the marginal distribution, illustrating that a lower-order correction can be very accurate. For this family of models corrections to other quantities of interest, for example the log marginal likelihood and predictive distribution, have to rely on cumulant expansions (Opper et al., 2008), and will be treated in detail a companion paper.

A Gaussian process prior forms the cornerstone of many popular non-parametric Bayesian methods. It has been used to great effect on various regression and classification problems. A Gaussian prior is placed on an N-dimensional unobserved variable f, for example

$$p(f) = \mathcal{N}(f\,;\,0,\,K) \; ,$$

where each f_n is associated with an input vector x_n, and K is a kernel matrix with entries $k(x_n, x_{n'})$ (Rasmussen and Williams, 2005). A binary classification task attaches a class label $y_n \in \{-1, +1\}$ to each input x_n, and a typical prediction would be the class of a new input x_* given the data $\mathcal{D} = \{x_n, y_n\}_{n=1}^N$. It is common to use the cumulative Normal distribution function $\Phi(\cdot)$ as a likelihood for correctly classifying a data point (Opper and Winther, 2000). The likelihood is dependent on the unobserved f_n associated with x_n, and hence

$$p(y_n|f_n) = \Phi(y_n f_n) \; .$$

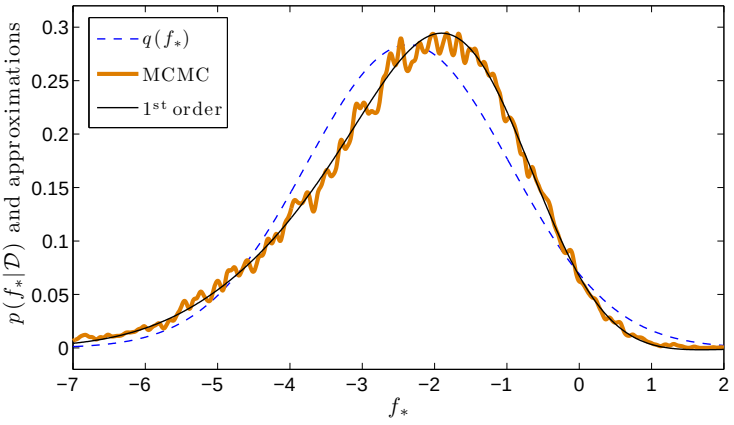


Figure 1: The first-order correction (13) is shown in black, with $q(f_*)$ in blue. Full details about the data set in question, as well as the individual terms in (13), are illustrated in Figure 13 in Appendix B. An MCMC estimate for the true marginal is overlayed in black, and comes from averaging $p(f_*|f)$ over 20,000 MCMC samples from the posterior of $p(f|\mathcal{D})$. The "spikiness" is a result of the variance of $p(f_*|f)$ being very narrow: if the "noise-free" latent f is given, then f_* is highly correlated with f and well determined for this example. The first-order correction gives an excellent approximation.

The posterior distribution of f is therefore

$$p(f|\mathcal{D}) = \frac{1}{Z}\prod_{n=1}^{N} p(y_n|f_n)\,\mathcal{N}(f\,;\,0,\,K)\ .$$

With this factorization the site functions are chosen to depend on only f_n such that the posterior is approximated by the same exponential family distribution (Gaussian) as the prior,

$$q(f) \propto \prod_{n=1}^{N} \exp\left(\tilde{\nu}_n f_n - \frac{1}{2}\tilde{s}_n f_n^2\right)\mathcal{N}(f\,;\,0,\,K)\ .$$

The notation in this section is deliberately chosen to be consistent with that of Rasmussen and Williams (2005, chapter 3), and we refer the reader to the reference for an example EP algorithm. We assume that a fixed point of EP has been reached. Let $\tilde{S}$ be a diagonal matrix containing $\tilde{s}_n$, and $\tilde{\nu}$ be a vector containing $\tilde{\nu}_n$. The posterior approximation is therefore $q(f) = \mathcal{N}(f;\mu,\Sigma)$, with $\Sigma = (K^{-1}+\tilde{S})^{-1}$ and $\mu = \Sigma\tilde{\nu}$.

The cavity posterior approximations $q_{\backslash n}(f) = \mathcal{N}(f;\mu_{\backslash n},\Sigma_{\backslash n})$ arise from setting $\tilde{\nu}_n = \tilde{s}_n = 0$ (giving diagonal matrix $\tilde{S}_{\backslash n}$ and vector $\tilde{\nu}_{\backslash n}$), where $\Sigma_{\backslash n}$ can be determined with a rank one update of Σ. The tilted distributions are therefore $q_n(f) \propto q_{\backslash n}(f)\,\Phi(y_n f_n)$.

The first order correction (12) can be applied to compute a correction to the marginal distribution of f_*, the latent function associated with a novel input x_*. Integrating $p(f_*|f)$

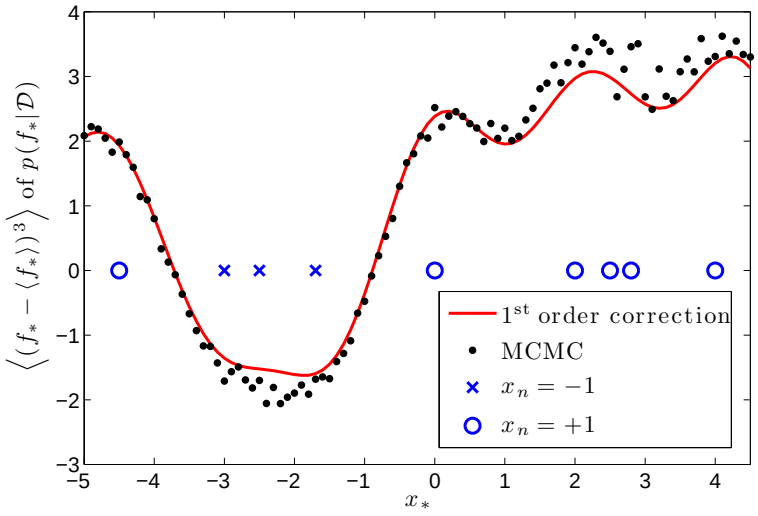


Figure 2: For different inputs x_*—and hence latent function f_*—the *third* cumulant of the first-order correction (13) is shown in red. It closely matches the true third cumulant of $p(f_*|f)$, which is plotted in black. The EC approximation $q(f_*)$'s higher cumulants are all zero. Figure 1 shows the particular approximations at $x_* = -2$, with further details appearing in Figure 13 and Appendix B.

with (12) yields

$$p(f_*|\mathcal{D}) \approx \sum_{n=1}^{N} q_n(f_*) - (N-1)q(f_*) \ . \tag{13}$$

Notice that corrections for the predictive distribution and log marginal likelihood cannot be expressed analytically in this way. Hence numerical quadrature or an expansion in terms of cumulants (Opper et al., 2008) is required. Higher-order terms of the above correction are also analytically intractable.

The detailed derivation of the correction is presented in Appendix B. Figure 1 provides a summary comparison of a first-order correction, $q(f_*)$, and a MCMC estimate of $p(f_*|\mathcal{D})$. The correction is very accurate and provides a much better fit than EC or EP at a negligible additional computational cost. Figure 13 in Appendix B gives further illustrations to accompany Figure 1. In Section 2.4 it was noted that the first order correction provides an approximation to nontrivial higher cumulants which would otherwise be *zero* in EC, even though the moments which are consistent in EC are not changed. Figure 2 illustrates this observation, showing an accurate approximation of the third cumulant for various distributions $p(f_*|\mathcal{D})$.

4. Mixture of Gaussians

We shall empirically examine the corrections to EP approximations through a multivariate mixture of Gaussians (MoG). Mixture models provide a more challenging testbed for EP

than the Gaussian Process model illustrated in Section 3, as the posterior is multi-modal with many symmetries, and the site distributions are not log-concave. For clarity we relegate the MoG derivations to Appendix D, favouring a simpler but similar model here. As an outline to deriving an algorithm for a MoG we consider the task of inferring the mixing proportions $\pi_k = p(k)$ in a model of the form

$$p(x|\theta) = \sum_k \pi_k p(x|k) \ ,$$

with $p(x|k)$ being fixed (Minka, 2001b). Since the mixing proportions should sum to one a Dirichlet prior for π is is a natural choice, and Appendix C gives a detailed description of all its properties needed in this context. We give the explicit EP message passing updates for the mixing proportions with fixed component densities in Algorithm 2 (this scheme is generalized to adaptive components in a straightforward way in Appendix D). Details for the required computations in Algorithm 2 are given below.

4.1 Variational and Predictive Distributions

The prior—and thus also the q-distribution in (3)—are Dirichlet,

$$q(\pi) = \mathcal{D}(\pi; \lambda) \ ,$$

with $\mathcal{D}(\pi; \lambda)$ given in Appendix C by (23). The parameters of q are $\lambda_k = \lambda_{k,0} + \sum_n \lambda_{k,n}$ (here λ_0 are the parameters of the prior, which we include into λ for simplicity).

We can also get the EC approximation to the predictive distribution both for new datum x, $p(x|\mathcal{D})$ and the cavity predictive distribution: $p(x_n|\mathcal{D}_{\backslash n})$. For the new datum x the approximation is straightforward using $q(\pi)$ as an approximate posterior:

$$p(x|\mathcal{D}) \approx \int d\pi \, p(x|\pi) \, q(\pi) = \sum_k \left\langle \pi_k \right\rangle_q p(x|k) \ , \tag{14}$$

with the mean value being

$$\left\langle \pi_k \right\rangle_q = \frac{\lambda_k}{\sum_{k'} \lambda_{k'}} \ .$$

For the "within data set" version we introduce the cavity distribution $q_{\backslash n}(\pi) = \mathcal{D}(\pi; \lambda_{\backslash n})$, using $\lambda_{k\backslash n} = \lambda_k - \lambda_{k,n}$, and derive a result that is very similar to the one above:

$$p(x_n|\mathcal{D}_{\backslash n}) \approx \sum_k \left\langle \pi_k \right\rangle_{q_{\backslash n}} p(x_n|k) \ . \tag{15}$$

For message passing we also need expectations of $q_n(\pi)$ from (4):

$$q_n(\pi) = \frac{1}{Z_n(\lambda_{\backslash n}, 1_n)} e^{\sum_{k'} (\lambda_{k'\backslash n} - 1) \log \pi_{k'}} \, \delta \left(\sum_{k'} \pi_{k'} - 1 \right) \sum_k \pi_k p(x_n|k) \ .$$

The above normalizer can easily be found by noting that

$$q_n(\pi) = \frac{Z(\lambda_{\backslash n}, 0)}{Z_n(\lambda_{\backslash n}, 1_n)} \, q_{\backslash n}(\pi) \sum_k \pi_k p(x_n|k) \ ,$$

such that

$$Z_n(\lambda_{\backslash n}, 1_n) = Z(\lambda_{\backslash n}, 0) \sum_k \left\langle \pi_k \right\rangle_{q_{\backslash n}} p(x_n|k) \ .$$

In this simple case we have $Z(\lambda_{\backslash n}, 0) = Z_{\mathcal{D}}(\lambda_{\backslash n})$, with the normalization $Z_{\mathcal{D}}$ of the Dirichlet being given by (24) in Appendix C.

4.2 Expectations

When updating $\mu \leftarrow \langle \phi(\theta) \rangle_{q_n(\theta;\Lambda_n)}$ in Algorithm 2 the sufficient statistics can be computed using $\log Z_n(\lambda_{\backslash n})$ as a generating function:

$$\left\langle \log \pi_k \right\rangle_{q_n} = \frac{d \log Z_n(\lambda_{\backslash n})}{d\lambda_{k\backslash n}} = \left\langle \log \pi_k \right\rangle_{q_{\backslash n}} + \frac{r_{nk}}{\lambda_{k\backslash n}} - \frac{1}{\sum_{k'} \lambda_{k'\backslash n}} \ , \tag{16}$$

where the expression for $\langle \log \pi_k \rangle_{q_{\backslash n}}$ is given by (25) in Appendix C with $\lambda \to \lambda_{\backslash n}$, and the "responsibility" r_{nk} was introduced as

$$r_{nk} = \frac{\lambda_{k\backslash n} p(x_n|k)}{\sum_{k'} \lambda_{k'\backslash n} p(x_n|k')} \ .$$

5. Parallel Tempering and Thermodynamic Integration

Having considered deterministic inference algorithms, the last bit of machinery that we shall need is a stochastic method to provide exact estimates in a large enough sample limit. Parallel tempering (PT) and thermodynamic integration (TI) are ideal for our purposes: PT is an efficient method of combining separate Monte Carlo simulations to sample across different modes of a target distribution and, as a by-product, TI can be used to estimate the normalizing constant or log marginal likelihood.

We conclude this section with a new practical generalization of PT and TI, which can in principle be used to combine stochastic and approximate methods. A further novel extension to the generalization is given in Appendix E.2.

5.1 Parallel Tempering (Replica Exchange)

A single MCMC simulation may run into difficulties if the target distribution is multimodal. The chain may get stuck in a local mode, and fail to fully explore other areas of the parameter space that have significant probability. A conceptual solution to this problem is to create a series of progressively flatter distributions through an inverse temperature parameter β, which ranges from zero to one. This gives a "tempered" posterior

$$p(\theta|\mathcal{D}, \beta) = \frac{1}{Z(\beta)} p(\mathcal{D}|\theta)^\beta p(\theta) \ , \tag{17}$$

where the normalizing constant (partition function) is $Z(\beta) = \int d\theta \, p(\mathcal{D}|\theta)^\beta p(\theta)$. The prior is recaptured with $\beta = 0$, and the posterior with $\beta = 1$. We now simulate N_β replicas of (17) in parallel, each using a $\beta \in \{\beta_i\}_{i=1}^{N_\beta}$. Let the set $\{\beta_i\}$ be ordered as a ladder with $\beta_i < \beta_{i+1}$.

Algorithm 2 Message Passing for Mixing Proportions

1: **initialize:** Set $\lambda_{k,n} \leftarrow 0$, for $n = 1, \ldots, N$ and $k = 1, \ldots, K$, initializing $q(\pi)$ to the prior. Set $\mu_k \leftarrow \langle \log \pi_k \rangle_{p(\pi)} = \psi(\pi_{k,0}) - \psi(\sum_k \pi_{k,0})$, where the digamma function $\psi(x)$ is defined as $\log \Gamma(x)/dx$.

2: **repeat**

3: Randomly choose example n, and make the following update steps:

4: Update the sufficient statistics

$$\mu_k \leftarrow \left\langle \log \pi_k \right\rangle_{q_n(\pi;\lambda_n)} = \psi(\lambda_k - \lambda_{k,n}) - \psi\left(\sum_k (\lambda_k - \lambda_{k,n})\right) \ .$$

5: Determine $q(\pi; \lambda')$ from μ, that is, by solving $\langle \log \pi_k \rangle_{q(\pi;\lambda')} = \mu_k$ with respect to λ'. As shown in Appendix C, this involves solving for $\alpha \equiv \psi(\sum_k \psi^{-1}(\mu_k + \alpha))$, followed by with $\lambda'_k = \psi^{-1}(\mu_k + \alpha)$. Update

$$\Delta\lambda_k \leftarrow \lambda'_k - \lambda_k \quad \text{and} \quad \lambda_k \leftarrow \lambda'_k \ .$$

6: Update $q_n(\pi; \lambda_n)$ with

$$\lambda_{k,n} \leftarrow \lambda_{k,n} + \Delta\lambda_k \ ,$$

ensuring that $\lambda_k = \lambda_{k,0} + \sum_n \lambda_{k,n}$.

7: **until** expectation consistency $\langle \log \pi_k \rangle_{q_n(\pi;\lambda_n)} = \langle \log \pi_k \rangle_{q(\pi;\lambda)} = \mu$ holds $\forall n, k$.

8: Compute $\log Z_{\mathrm{EC}}$ from (5) with

$$\begin{aligned} \log Z_{\mathrm{EC}} &= \sum_n \log Z(\lambda - \lambda_0 - \lambda_n, 1_n) - (N-1) \log Z(\lambda - \lambda_0, 0) \\ &= \sum_n \log \left[\frac{Z_{\mathcal{D}}(\lambda_{\backslash n})}{Z_{\mathcal{D}}(\lambda)} p(x_n | \mathcal{D}_{\backslash n}) \right] + \log Z_{\mathcal{D}}(\lambda_0) + \log Z_{\mathcal{D}}(\lambda) \ , \end{aligned}$$

where $p(x_n|\mathcal{D}_{\backslash n})$ signifies the "cavity" predictive distribution from (15).

The parameter space is replicated N_β times to $\{\theta_i\}_{i=1}^{N_\beta}$, and the full target distribution that is being sampled from is

$$p(\{\theta_i\}) = \prod_{i=1}^{N_\beta} \frac{1}{Z(\beta_i)} \exp\Big(\beta_i \log p(\mathcal{D}|\theta_i)\Big) p(\theta_i) \ .$$

We run the N_β chains independently to sample from distributions $p(\theta|\mathcal{D}, \beta_i)$, and add an additional *replica-exchange* Metropolis-Hastings move to swap two β's, or equivalently two parameters, between chains. Let $\{\theta_i\}^{\mathrm{new}}$ be a parameter set with θ_i and θ_j swapped. The acceptance probability of the move is $p(\mathrm{accept}) = \min(1, p(\{\theta_i\}^{\mathrm{new}})/p(\{\theta_i\}))$, and the acceptance ratio simplifies to

$$\frac{p(\{\theta_i\}^{\mathrm{new}})}{p(\{\theta_i\})} = \exp\Big((\beta_i - \beta_j)\big(\log p(\mathcal{D}|\theta_j) - \log p(\mathcal{D}|\theta_i)\big)\Big) \ . \tag{18}$$

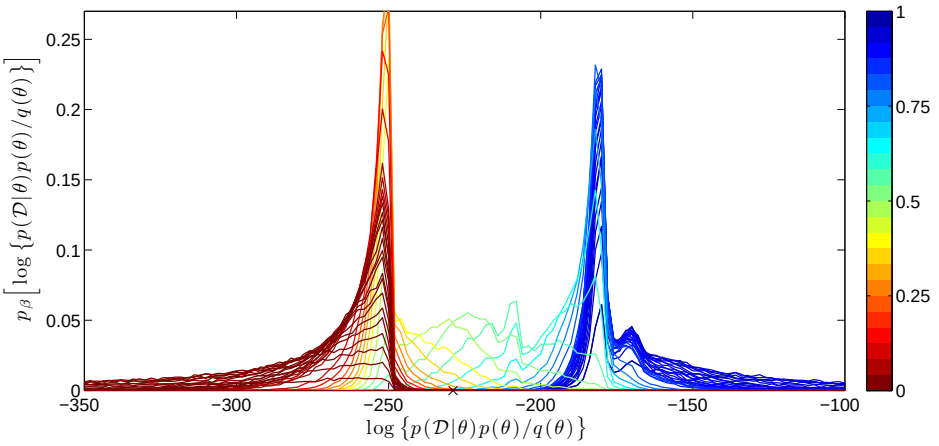


Figure 3: The density of $\log\{p(\mathcal{D}|\theta)p(\theta)/q(\theta)\}$ under replicas at different temperatures, $p(\theta|\mathcal{D},\beta)$, defined in (20). These densities correspond to "energy histograms," and following (18) there should be an overlap between adjacent replicas at different temperatures, so that acceptance of configuration or parameter swaps is allowed for. For interest, the log marginal likelihood $\log p(\mathcal{D})$ is indicated with a $\times$. The color bar indicates the inverse temperature β. This illustration comes from the **galaxy** data set with $K = 3$ components.

The temperatures of the two replica i and j have to be close to ensure non-negligible acceptance rates; neighboring pairs are typically taken as candidates. To fully satisfy detailed balance, pairs $\{i, i+1\}$ can be uniformly chosen, for example. With this formulation the states of the replicas are effectively propagated between chains, and the mixing of the Markov chain is facilitated by the fast relaxation at small β's.

From (18), the acceptance probability depends on the difference between $\log p(\mathcal{D}|\theta_i)$ and $\log p(\mathcal{D}|\theta_{i+1})$, and for some swaps to be accepted this difference should not be "too big"; there should be an *overlap* of some of the log likelihood evaluations of adjacent chains, as illustrated in Figure 3. For a simulation at inverse temperature β, define the mean evaluation of the log likelihood as

$$\left\langle \log p(\mathcal{D}|\theta) \right\rangle_\beta = \int d\theta \, \log p(\mathcal{D}|\theta)\, p(\theta|\mathcal{D},\beta) \; .$$

If we knew the variance in chain β, $\sigma_\beta^2 = \langle [\log p(\mathcal{D}|\theta)]^2 \rangle_\beta - \langle \log p(\mathcal{D}|\theta) \rangle_\beta^2$, then it can be shown that temperatures should be chosen according to the density $Q(\beta) \propto \sigma_\beta$ (Iba, 2001). This is obviously difficult, as σ_β^2 is not known in advance, and has to be estimated. Good results can be achieved under the assumption $\sigma_\beta^2 \propto 1/\beta^2$ (the equivalent of assuming a *constant* heat capacity in a physical system), giving a *geometric* progression, hence choosing β_i/β_{i+1} constant (Kofke, 2002).

5.2 Thermodynamic Integration

The samples from parallel tempering can be used for model comparison (Gregory, 2005, Skilling, 1998), as the marginal likelihood can be obtained from tempering. Firstly, notice that the integral

$$\int_0^1 d\log Z(\beta) = \int_0^1 d\beta \, \frac{d\log Z(\beta)}{d\beta} = \log Z(1) - \log Z(0) = \log Z(1) = \log p(\mathcal{D})$$

is equal to the log marginal likelihood, as $\beta = 0$ gives the prior, which integrates to one. We therefore have to determine the derivative $\frac{d}{d\beta}\log Z(\beta)$. By taking the derivative of the log normalizer (log partition function), we see that it evaluates as an average over the posterior

$$\frac{d\log Z(\beta)}{d\beta} = \frac{1}{Z(\beta)} \int d\theta \, \log p(\mathcal{D}|\theta) \times p(\mathcal{D}|\theta)^\beta p(\theta) = \left\langle \log p(\mathcal{D}|\theta) \right\rangle_\beta .$$

The log marginal likelihood equals

$$\log p(\mathcal{D}) = \int_0^1 d\beta \left\langle \log p(\mathcal{D}|\theta) \right\rangle_\beta \tag{19}$$

and can be numerically estimated from the Markov chain samples. If $\{\theta_i^{(t)}\}$ represents the samples for tempering parameter β_i, then the expectation is approximated with

$$\left\langle \log p(\mathcal{D}|\theta) \right\rangle_{\beta_i} \approx \frac{1}{T} \sum_{t=1}^{T} \log p(\mathcal{D}|\theta_i^{(t)}) .$$

We assume that a burn-in sample is discarded in the sum over t. As a set of chains are run in parallel at different inverse temperatures $0 = \beta_1 < \cdots < \beta_{N_\beta} = 1$, the integral can be evaluated numerically by interpolating the N_β expectations between zero and one (say with a piecewise cubic Hermite interpolation, available as part of `Matlab` and other standard software packages), and using for example the trapesium rule to obtain the desired result. Figure 4 illustrates how $\log p(\mathcal{D})$ is estimated.

Parallel tempering can be done *complementary* to any Monte Carlo method at a single temperature. Appendix E presents Gibbs sampling to sample from $p(\theta|\mathcal{D}, \beta)$ for the MoG problem.

5.3 A Practical Generalization of Parallel Tempering

The success of the interpolation obtaining $\langle \log p(\mathcal{D}|\theta) \rangle_\beta$, illustrated in Figure 4, is dependent on the slope

$$\frac{d\langle \log p(\mathcal{D}|\theta) \rangle_\beta}{d\beta} = \frac{d^2 \log Z(\beta)}{d\beta^2} = \sigma_\beta^2$$

at $\beta \approx 0$. Consider the following thought exercise: Imagine a non-informative (infinitely wide) prior at $\beta = 0$. Samples from this prior will strictly speaking have an infinite variance σ_0^2. With $\beta \approx 0$ we introduce the likelihood, practically infinitely decreasing the variance of our samples, causing $\langle \log p(\mathcal{D}|\theta) \rangle_\beta$ to asymptotically diverge at zero. As we narrow our

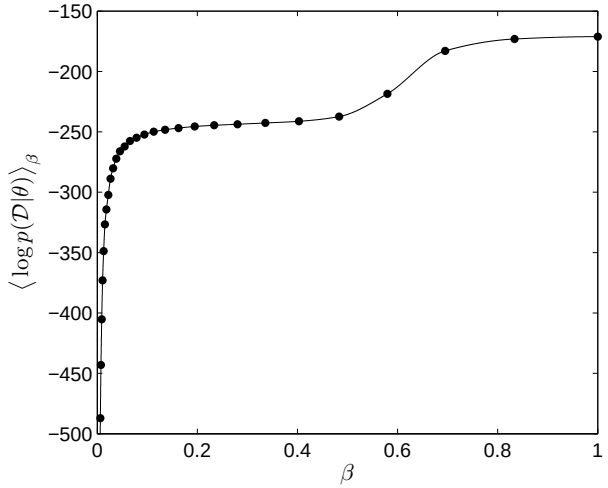


Figure 4: The log likelihood averages $\langle \log p(\mathcal{D}|\theta) \rangle_\beta$ are estimated from each of the MCMC simulations at temperatures $\{\beta_i\}$, and interpolated, so that Equation (19)'s integral can be evaluated numerically. This illustration comes from the **galaxy** data set with $K = 3$ components.

prior the change in this mean should be less rapid, and this motivates a generalization of PT and TI such that we get a more stable interpolation.

We introduce a new distribution $q(\theta)$, which might be a narrower version of the prior, and modify (17) to

$$p(\theta|\mathcal{D}, \beta) = \frac{1}{Z(\beta)} \left[p(\mathcal{D}|\theta) \frac{p(\theta)}{q(\theta)} \right]^\beta q(\theta) \ . \tag{20}$$

The log marginal likelihood can, as before, be determined with

$$\log p(\mathcal{D}) = \int_0^1 d\beta \left\langle \log p(\mathcal{D}|\theta) + \log \frac{p(\theta)}{q(\theta)} \right\rangle_\beta \ .$$

It is evident that setting $q(\theta) = p(\theta|\mathcal{D})$ gives an integral over a constant function, $\log p(\mathcal{D}) = \int_0^1 d\beta \, \langle \log p(\mathcal{D}) \rangle_\beta$. This suggests a wealth of possibilities of approximating $p(\theta|\mathcal{D})$ with $q(\theta)$ to effectively combine deterministic methods of inference with Markov chains. This comes with a cautionary note as VB, for example, may give a $q(\theta)$ that captures (lower-bounds) a mode of a possibly multimodal posterior, causing PT to lose its pleasing property of fast relaxation at high temperatures. In our results presented in Section 6, we have found it completely adequate to use a narrower version of the prior where necessary. Appendix E concludes with a short generalization to sample from (20) for the MoG problem.

6. Results

Low order corrections provide the tools to both improve inference accuracy, and to give an indication of the quality of approximate solutions. We illustrate and elaborate on these

claims, with comparisons between various deterministic and stochastic methods, through this practical discussion. Data is viewed as being observed from a mixture model $p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Gamma_k^{-1})$, as is discussed in Appendices D, E, and F. A Dirichlet prior is placed on π, and Normal-Wishart priors on μ_k and Γ_k; the approximating distribution $q(\theta) = q(\pi) \prod_k q(\mu_k, \Gamma_k)$ follows the same distribution as the prior.

6.1 Modes and Symmetries

Mixture models are invariant under component relabelling, with a $K!$ growth in the number of permutations also manifesting itself in symmetries in the posterior density. In aid of interpreting later results, we present some basic understanding of VB, EP, and low order corrections under this property. Our aim in this section is to use simple toy posteriors to facilitate discussion on the behavior of q under various scenarios and discuss how that might affect the estimation of the marginal likelihood and predictive distribution.

The labelling of hidden units of a two-layer neural network gives rise to symmetries similar to those observed in mixture models. For neural networks a statistical mechanics analysis shows that for small N the posterior is uni-modal and "star-like," as convex combinations of parameters with high posterior value which are equivalent under permutations will also have high density (Engel et al., 1992). The symmetry is broken into equivalent disconnected modes for large N.

For mixture models we can analyze the situation where q is restricted to approximate the posterior in one of the symmetric modes, as what will typically be the solution for both VB and EP/C when N is large. Minimizing the KL-divergence $\mathrm{KL}(q\|p)$ leads to a solution where q is proportional to p within the mode (and by construction zero otherwise). If there are $K!$ modes contributing equally to the normalizer and q is restricted to one of them, then q's normalizer is a factor of $K!$ smaller than p's and consequently $\mathrm{KL}(q\|p) = \log K!$ at the minimum (Bishop, 2006, page 484). However, *groups* of equivalent modes are often present. A simple example is a 3-component mixture with three "clusters" of data. If each component is associated with a cluster, there are 3! labelling symmetries. Another VB or EP fixed point may prune one mixture component (see MacKay 2001 for VB and Figure 11 for EP), leaving one component to cover two clusters of data, and one component the other; this solution has *yet another* 3! labelling symmetries, albeit possibly with a lesser contribution to the normalizer. In effect the correction is rather $\mathcal{O}(\log K!)$, as illustrated in Figure 9. A useful approximation would be to correct the marginal likelihood estimate by a factor of $K!$ when N is large. The predictive distribution is invariant under the symmetry and will thus not be greatly affected by q approximating only one mode, as is shown in Figure 5. For small to intermediate values of N the situation is less clear, as the following example illustrates.

In Table 1 we illustrate a number of posterior distributions, with the VB and EC approximations overlaid. We also overlay the first order correction to $q(\theta)$, given from (12) by $p(\theta|\mathcal{D}) \approx \sum_n q_n(\theta) - (N-1)q(\theta)$. For Table 1 all parameters but the means were kept fixed, such that with $K = 2$ the approximation $q(\theta) = q(\mu_1)q(\mu_2)$ is a factorised Gaussian. Both component variances were equal, and we used $(\pi_1, \pi_2) = (0.4, 0.6)$. The modes are thus not completely symmetric but this set-up still illustrates the points made above well. We chose the component variances (set to one) such that the posterior modes overlap when

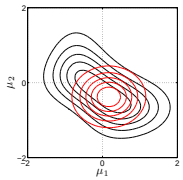
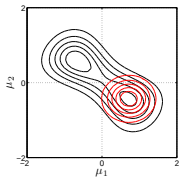
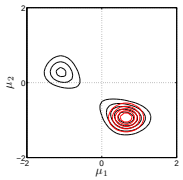
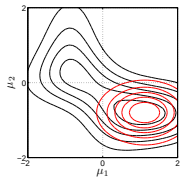
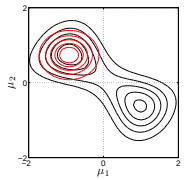
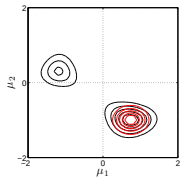
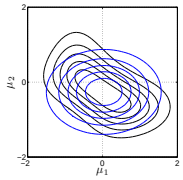
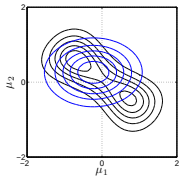
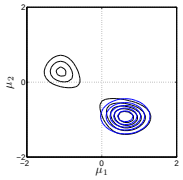
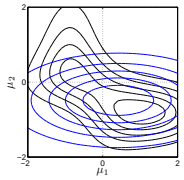
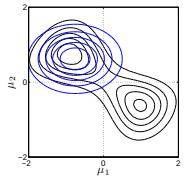
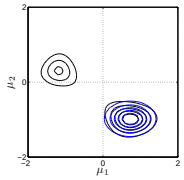
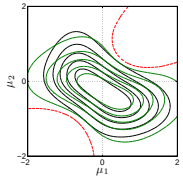
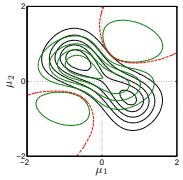
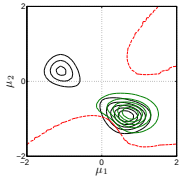
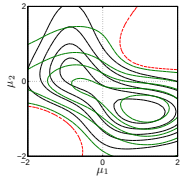
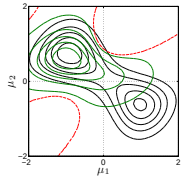
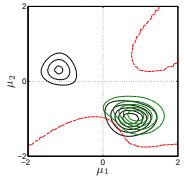
$N = 7$	$N = 14$	$N = 40$	$N = 7$	$N = 14$	$N = 40$
narrower prior:			broader prior:		
					
					
					

Table 1: A comparison between the VB (top row), and EC (middle row) approximations $q(\theta)$, and a first order correction to the EC approximation (bottom row). Data is assumed to come from a two-component mixture with *only* the means $\theta = \{\mu_1, \mu_2\}$ unknown. Under various priors and data set sizes we show the posterior $p(\theta|\mathcal{D})$ in thin black lines, with the VB, EC, and first-order corrected approximations overlaid in thicker lines. The first order correction integrates to one but is not guaranteed to be nonnegative (bottom row); dashed red lines are used to demarcate the regions of parameter space where the correction dips below zero.

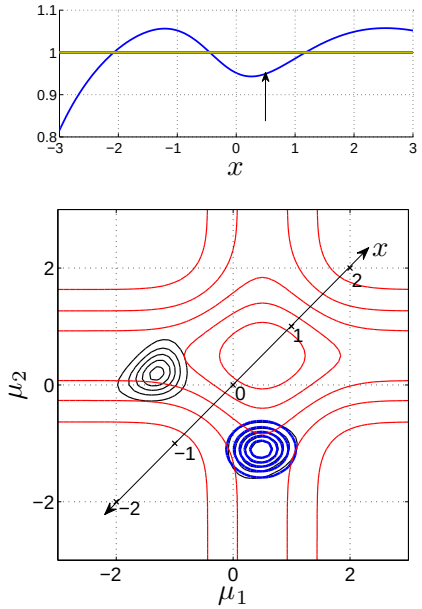
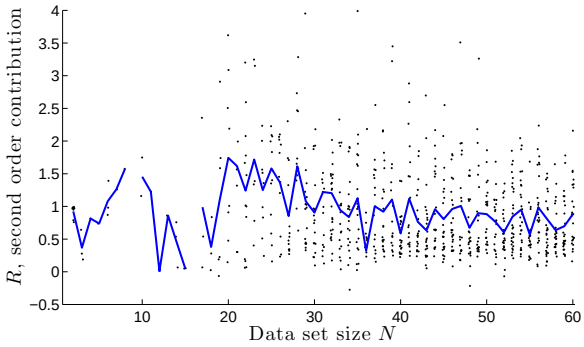


Figure 5: Symmetries and averages: The likelihood $p(x|\{\mu_1, \mu_2\}) = 0.4\mathcal{N}(x; \mu_1, 1) + 0.6\mathcal{N}(x; \mu_2, 1)$ is plotted as red contours for the novel observation $x = \frac{1}{2}$ at the arrow in the top figure. Observing x centres the likelihood function at (x, x) along the μ_1–μ_2 axis in the bottom figure. The *predictive density* $p(x|\mathcal{D})$ is average of the likelihood over the *bi-modal* posterior (black contours), while the *approximate* predictive density is the average of the likelihood function over the *uni-modal* EC approximation (overlaid in blue). The near-symmetry of the posterior implies that each mode contributes approximately *half* its mass. When the EC approximation puts *all* its mass on one mode, and the modes are well separated, the two predictive densities are therefore similar. The top figure shows the ratio between the true and approximate $p(x|\mathcal{D})$; the discrepancy at negative x is due to the fact that the posterior is not perfectly symmetrical (e.g., when $x = -2$ its likelihood is centred at (-2,-2) and overlaps less with the EC mode).

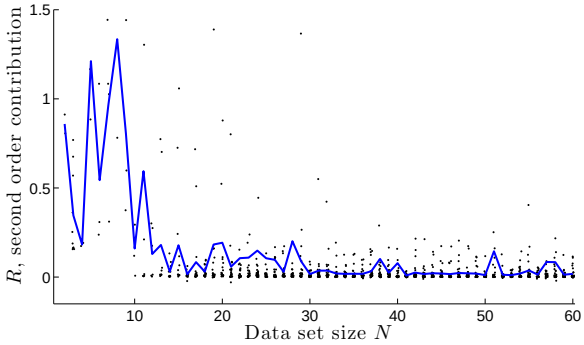
N is small, with bimodality arising as N increases. We will see in the following sections that even though q is a rather crude approximation to the posterior the predictions for the predictive distribution are fairly precise.

The correction given by (12) integrates to one but is not guaranteed to be nonnegative, as it follows from discarding the higher order terms in (7). The first order correction to the predictive density, however, usually remains nonnegative because it is an *average* of $p(x|\theta)$ over (12). This underlines the fact that average properties will not be strongly affected by imprecision in approximating distributions.

Other local minima that we did not show in Table 1 was for N small, where $q(\mu_1)$ remains as broad as the prior and the component is effectively pruned, while $q(\mu_2)$, on the other hand, caters for both mixture components (MacKay, 2001).



(a) With overlapping mixture components, damped EP does not necessarily converge for small N (e.g., EP failed, in the sense that the 2^{nd} order correction cannot be computed, on all 30 random data sets of size 8). The corrections are on average (blue line) large for broadly overlapping posterior modes, as EP does not necessarily lock onto one of them.



(b) With well separated clusters the 2^{nd} order corrections indicate for which N, on average, EP prefers a modal solution. Note the better convergence of EP for larger N, with on average stabler fixed points than Figure 6(a). This is reflected in the corrections being close to zero.

Figure 6: The second order term of (8) on 3600 random data sets.

6.2 Corrections

The illustrations in Table 1 suggest that the lowest nontrivial corrections can provide insight into the quality of approximation, as we expect corrections to be small for good approximations. To illustrate this claim, 30 random data sets $\mathcal{D}_N$ were drawn for each size $N = 1, \ldots, 60$ according to $\mathcal{D}_N \sim p(x)$, with $p(x)$ being a three-component mixture with $\pi = (0.2, 0.3, 0.5)$ and $\mu = (-2, 0, 2)$. Two cases were used for the variance: firstly, $\Gamma_k^{-1} = 0.5$ provides a model with overlapping mixture components; secondly, $\Gamma_k^{-1} = 0.1$ gives a model with components that are further separated. EP was run with damping $\gamma = 0.5$, and the second order corrections to the log marginal likelihood approximations were computed where possible (i.e., EP converged, etc.); see Appendices D and F.

Figures 6(a) and 6(b) illustrate the "overlapping" and "separate" examples, showing that when N is small compared to $\log K!$, and the posterior is "starlike" and comparatively

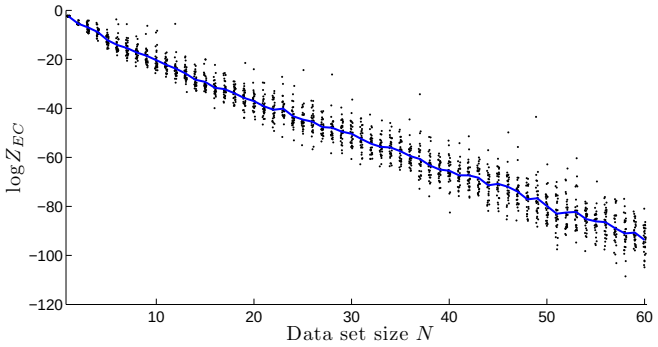


Figure 7: The growth of $\log Z_{\mathrm{EC}}$ for the random data sets $\mathcal{D}_N$ used to obtain Figure 6(b).

unimodal, EP often fails, and the nature of the problem is reflected in the large corrections. When N becomes large EP often converges (to one of $K!$ equivalent modes); small corrections immediately tell us that solution is close to exact, apart from here a $\log K!$ correction to the marginal likelihood.

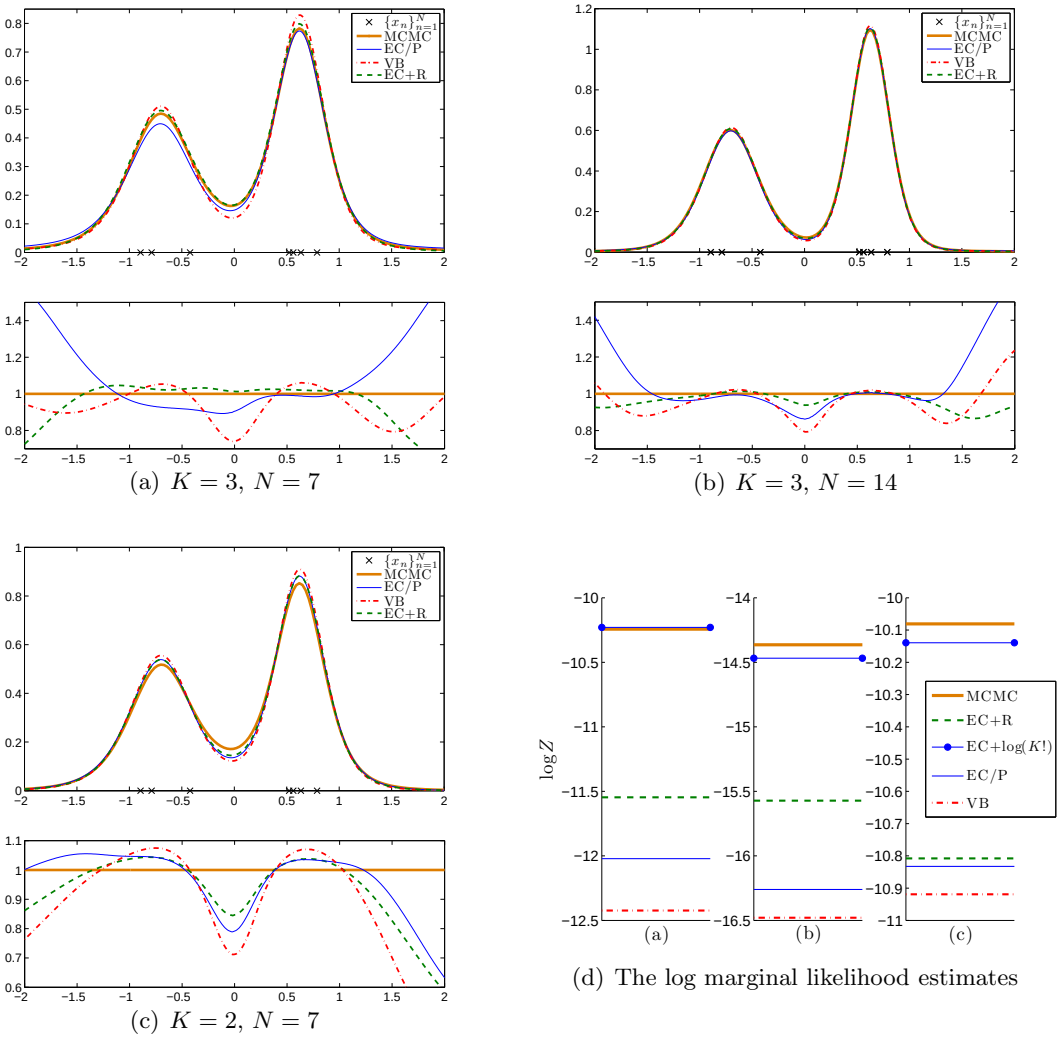
We also observe that the corrections do not scale with N, whereas the free energy $\log Z_{\mathrm{EC}}$ does, as shown in Figure 7. This is an important property, as it means that the quality of approximation does not deteriorate with increasing N.

When the observations $\mathcal{D}_N \sim p(x)$ are i.i.d. we expect that $\log Z_{\mathrm{EC}}/N$, by its form as an empirical average over N terms, should converge to a non-random c_Z as $N \to \infty$. In fact a linear scaling $\log Z_{\mathrm{EC}} \to c_Z N$ is observed in Figure 7. When and whether the expected correction $\langle \log R \rangle_{\mathcal{D}_N} = c_R(N) \to 0$ as $N \to \infty$ (and hence EP becomes exact) is an open question. This does not seem true for Figure 6(a): If the the posterior modes were well-separated then for large N, a change in one mean parameter in a factorized approximation will not greatly affect the other. If, in this case, the means are close compared to the standard deviations of the normal densities, the mean parameters will stay correlated also for large data sets, and the corrections will persistently stay bigger for large N.

6.3 Toy Example

To illustrate the difference between EC and VB, and show additional gains from perturbation corrections, we generated a small data set ($N = 7$) from a mixture of two Gaussians. The hyperparameters followed that of Section 6.4.

Under two model assumptions we show in Figure 8 that EC or EP (labeled "EC/P") gives a predictive density that is generally closer to the truth than that given by VB. Each example in the toy data set was duplicated (see Figure 8(b)) to show that this gain decreases under larger data sets; this decrease is due to the predictive density being an average of $p(x|\theta)$ over now more concentrated VB and EC posterior approximations. Secondly, meaningful improvements can be achieved through perturbation corrections. Figure 8(d) shows a second order correction to the log marginal likelihoods of the examples in question, labeled "EC+R" (see Appendix F). A lower bound to the log marginal likelihood is provided by VB. The improvement is also visible when we are concerned with the predictive density, for which we show a first order correction.



(a) $K = 3$, $N = 7$

(b) $K = 3$, $N = 14$

(c) $K = 2$, $N = 7$

(d) The log marginal likelihood estimates

Figure 8: Predictive densities $p(x|\mathcal{D}, \mathcal{M}_K)$ given by VB, EC, and a perturbation correction (EC+R), with accompanying log marginal likelihood estimates and MCMC "truth" baselines. Note that if we "correct" with a factor $\log K!$ we get very close to the "truth" for VB and to a even higher degree for EC. EC+R overshoots in two cases but that might be because the perturbation corrected posterior is actually multi-modal. The lower figures in 8(a) to 8(c) show the *ratio* between each of the approximate predictive densities and the "truth."

6.4 A Practical Comparison

In this section we draw a comparison between the approximate log marginal likelihoods and predictive distributions given by VB, EP, and various corrections, and use estimations given by PT and TI as a benchmark. For interest we also include results from an implementation of $\alpha = \frac{1}{2}$ in Minka's general α-divergence message passing scheme for this problem, but

refer the interested reader to Minka (2005) for further details. Finding a VB approximation follows directly from the expectation maximisation algorithm given by Attias (2000), with a slightly different parameterization of the Wishart distribution.

From the results that follow, we observe that the growth of $\log Z$, as a function of model size, gives a characteristic "Ockham hill" (defined in more detail later in this section), where the "peak" of the hill indicates the model with highest approximate $\log p(\mathcal{D})$. This graph can be used for model comparison or selection, as its form closely matches the MCMC evaluation of $\log p(\mathcal{D})$. We will also see that, following Section 6.1's discussion, the discrepancy between a $\log Z$ estimate and the true $\log p(\mathcal{D})$ grows as the model size is increased. Furthermore, the EC approximation gives a predictive distribution that is closer to the truth than VB, with the gain decreasing with increasing N. We will show that a principled algorithm initialization can circumvent many spurious local minima in the log marginal likelihood estimate. If completely arbitrary initialization schemes are implemented, one may note that the number of local solutions is influenced by the width of the prior distribution, with *more* local minima arising under broader prior distributions.

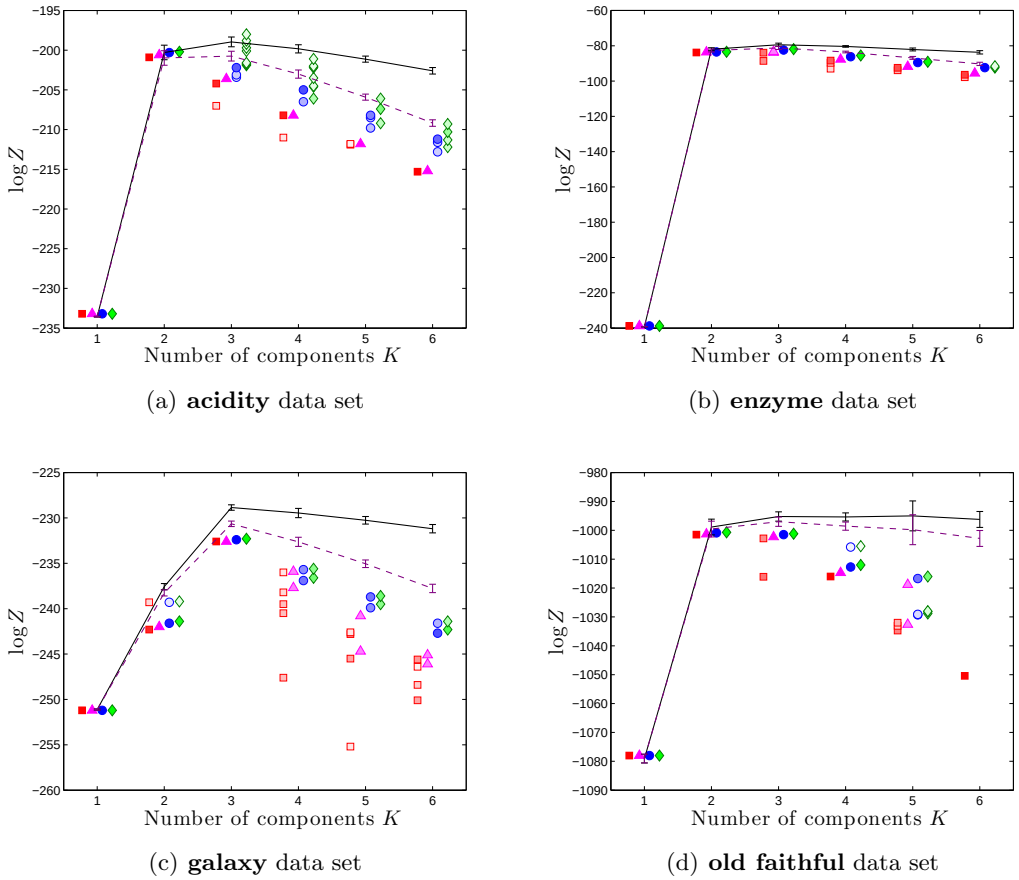
The data sets under investigation have been well studied, for example, by Richardson and Green (1997) for a reversible jump MCMC, and by Corduneanu and Bishop (2001) for variational Bayesian model selection: the **galaxy** data set contains the velocities (in 1000s of km/second) of 82 galaxies, diverging from our own, in the Corona Borealis region; the **acidity** data set contains the log measured acid neutralizing capacity indices for 155 lakes in North-central Wisconsin (USA); the **enzyme** data set contains enzymatic activity measurements, for an enzyme involved in the metabolism of carcinogenic substances, taken from 245 unrelated individuals; the **old faithful** data set contains 222 observation pairs consisting of eruption time and waiting time to the next eruption, from the Old Faithful Geyser in the Yellowstone National Park.

6.4.1 The Approximate Log Marginal Likelihood

Ockham hills are useful for visualizing log marginal likelihood estimates for a set of plausible models with increasing explanatory power, for example, mixture models with increasing K. The largest estimates of $\log Z$ for the various models typically form a hill, peaking at the "optimal" model. As models become *less* complex, the hill falls steeply due to a poorer explanation of the data. For *more* complex models the plots show a slower downward trend, as an improvement in data fit is counterbalanced by a penalty from a larger parameter space in Bayesian marginalization. For mixture models this downward trend is even slower when the true log marginal is considered; this is mainly due to the number of modes in the true posterior increasing with the number of components, with an approximation possibly only capturing one of them.[5]

In the case of VB, the $\log Z$ approximation provides a lower bound to the marginal likelihood $p(\mathcal{D}|\mathcal{M})$, and this quantity is often used for model selection (Beal and Ghahramani, 2003, Bishop and Svensén, 2003, Corduneanu and Bishop, 2001). The model with the largest bound is typically kept, although the bound can also be used for model averaging. Regardless of our method of approximation, poor local minima in the objective function have to be avoided in order to obtain meaningful results.

5. Rasmussen and Ghahramani (2001) present an account which includes "Ockham plateaus."



(a) **acidity** data set

(b) **enzyme** data set

(c) **galaxy** data set

(d) **old faithful** data set

Figure 9: Ockham hills for various data sets. VB is shown as red squares, $\alpha = \frac{1}{2}$ as magenta triangles, EC/P as blue circles, and EC+R as green diamonds. An estimate of $\log p(\mathcal{D}|\mathcal{M}_K)$ found by PT and TI is shown as a line. The effect of an $\mathcal{O}(\log K!)$ correction on any of the approximate solutions can be seen by comparing them against the dashed-line plot of $\log p(\mathcal{D}|\mathcal{M}_K) - \log K!$. (For Figure 9(d)'s $K = 6$ the EP and $\alpha = \frac{1}{2}$ schemes did not converge.)

Figure 9 shows such hills for the marginal likelihood approximations for different data sets for VB, $\alpha = \frac{1}{2}$ message passing, EP, and a second-order perturbation correction. The prior hyperparameters were $\lambda_{k,0} = 1$, $m_{k,0} = 0$, $\nu_{k,0} = 10^{-2}$, $a_{k,0} = 1$ and $B_{k,0} = 0.11$. For Figure 9(d) we took $B_{k,0} = [0.11, 0.01; 0.01, 0.11]$. For each of the models $\mathcal{M}_K$, with K mixture components, the figures show twenty approximations for each method, with the colour intensity of each plot corresponding to the frequency of reaching different approximations for $\log Z$. Each plot is complemented with estimates of $\log p(\mathcal{D}|\mathcal{M}_K)$. The estimates—shown as lines—were obtained form an average over ten PT and TI simulations, with two standard deviation error bars also being shown.

Finally, it is evident that the "true peak" in Figure 9(a) does not match the peak obtained by approximate inference. Without having to resort to MCMC and TI, the second order correction for $K = 3$ already confirms that the approximation might be inadequate.

6.4.2 The Effect of a Good Initialization

Finding the best VB/EP solution is strongly seed-dependent in the problem considered here. In this council of despair an educated guess may take us a long way: many inferior local minima in the VB/EP objective functions can be suppressed with a good algorithm initialization.

We base our factor initializations around a scaled version of the solution obtained by the VB expectation maximisation algorithm,

$$\exp(\Lambda_n^T \phi(\theta)) \propto \exp\left(\int dz_n \, q(z_n) \log p(x_n, z_n|\theta)\right) ,$$

which was seeded with a data clustering based on the k-means algorithm.[6] This is illustrated in Figure 9.

When using an "out of the box" EP scheme, starting with a slight asymmetric prior that is later corrected for, many lower minima are also found. The same behavior arises when the VB parameters are randomly initialized. Figure 10 shows more local minima than Figure 9(c), and the results in Bishop (2006, chapter 10), where the same principled initial guess for VB was used.[7]

The canonical EP scheme (and indeed $\alpha = \frac{1}{2}$) sometimes did not converge to a fixed point. This is evident in Figure 10 and has been observed in practice (Minka, 2001a): when EP does not converge, the reason can be traced back to the approximating family being a poor match to the exact posterior distribution.

6.4.3 The Predictive Distribution

Given a specific model $\mathcal{M}$, the predictive distribution can be approximated by using $p(x|\mathcal{D}) \approx \int d\theta \, p(x|\theta) q(\theta)$, as is shown for example in Figure 11. The final predictive distribution strongly depends on whether or not a global minimum in the objective function in (5) has been found, as is clear from Figure 11. To illustrate how much the approximate predictive distribution differs from the true predictive distribution, the figures show $p(x|\mathcal{D})$ obtained from an average over ten thousand $\beta = 1$ samples from a parallel tempered Markov chain. Figure 12 shows the gain achieved by EC/P over VB, and in turn the further improvement from a perturbation correction to the EC approximation (see Appendix F).

7. Conclusion and Outlook

In this paper we presented a method for computing systematic corrections to EC approximations in Bayesian inference. These corrections are useful not only in improving estimates like log marginal likelihood and predictive density approximations, but can also provide

6. Similar to Appendix E, z indicates latent variables, with $p(\theta, z|\mathcal{D})$ approximated by $q(\theta)q(z)$. We point the interested reader to Attias (2000).
7. Markus Svensén, personal communication.

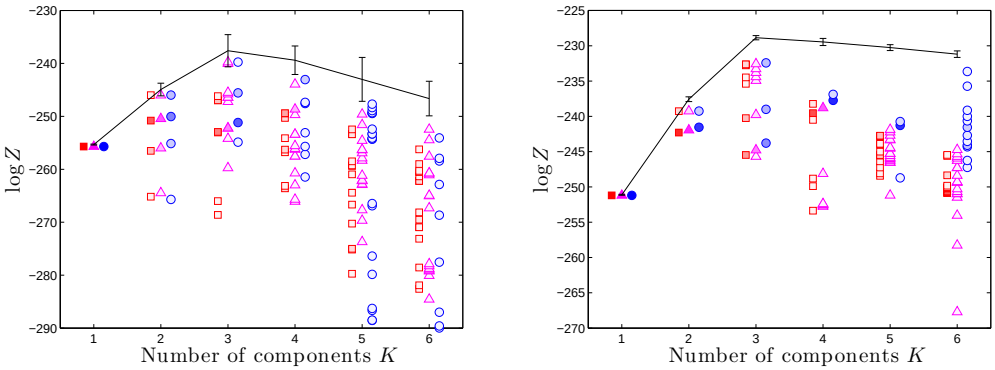


Figure 10: The effect of random algorithm initializations using the **galaxy** data set: For the *left* figure a broader prior with $\nu_{k,0} = 10^{-6}$, and for the *right* figure a much narrower prior with $\nu_{k,0} = 10^{-2}$, was used. Compared to Figure 9(c), note for example the additional local maxima at $K = 3$, and the greater number of local minima under a broader prior. (For $K = 6$ the EP scheme failed to converge without a sensible initialization.)

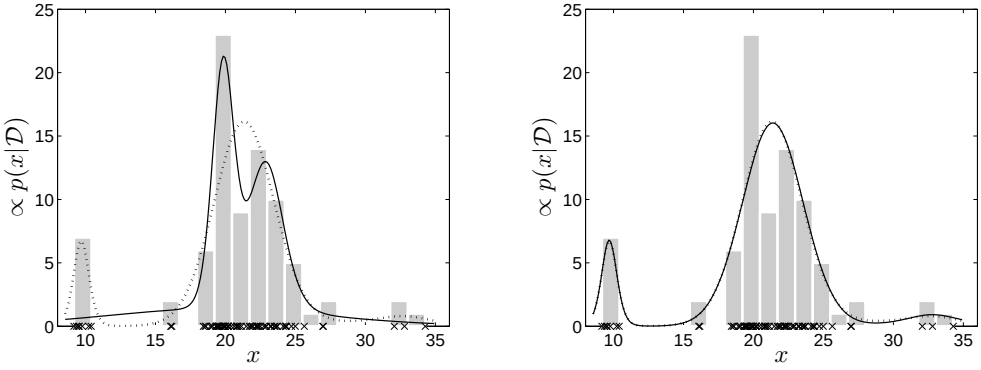


Figure 11: EP can have more than one stable fixed point: The predictive distribution $p(x|\mathcal{D}, \mathcal{M}_3)$, from two *different* approximations for the **galaxy** data set. For $K = 3$ under narrower prior in Figure 10, we see three local maxima of the EC objective function in (5): the predictive distribution shown on the *left* coincided with $\log Z_{\mathrm{EC}} = -243.8$, whereas the approximation on the *right* coincided with a much higher $\log Z_{\mathrm{EC}} = -232.4$. The true predictive distribution, obtained from an average over a PT MCMC sample, is shown with a dotted line.

insight into the quality of an approximation in polynomial time. When the corrections are large the EC approximation may be questioned or discarded, and we hope to address the question of how it is done in practice in future work.

A juxtaposition of VB, EC and EP, PT with TI, and EC with corrections, was given in the context of Gaussian mixture models. We argued that EC can give improvements over

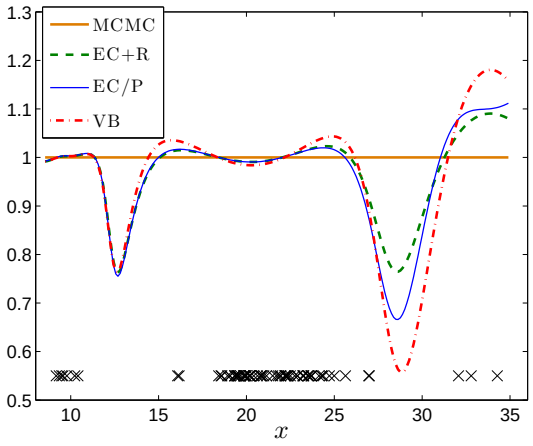


Figure 12: The ratio between each of the approximate predictive densities and the MCMC "truth" of $p(x|\mathcal{D}, \mathcal{M}_3)$ for the **galaxy** data set. This figure corresponds to Figure 11 (right).

VB, and can in turn be improved through a perturbation expansion. Throughout the paper our "gold standard" was given by PT, and we presented possible ways of improving it. We would like to include better MCMC algorithms in this rich tapestry of methods: PT is not the best choice for near first-order phase transitions. In Figure 3 the high probability regions are very different above and below the transition at $\beta \approx 0.5$, suggesting multicanonical sampling as a viable alternative, since it aims at sampling from a distribution that is flat in the log likelihood and will therefore not have this "bottle-neck."

The choice of a unimodal $q(\theta)$ to capture the characteristics of a typically multimodal $p(\theta|\mathcal{D})$ also leads to various questions. When there are symmetries in the parameter space, with overlapping modes, we may ask whether or not we would achieve a better predictive density with EC, say, if the approximation is restricted to one mode. In the case where $p(\theta|\mathcal{D})$ is multimodal (large N) then fairly general arguments suggest that we should correct the marginal likelihood estimate by a factor of $K!$—higher order corrections may clarify for which N and under which conditions this transition will typically take place.

One way be improve the approximation is to generalize $q(\theta)$ for example by including a small fraction of the data points (similar to the proposed generalization of PT). However, that poses an additional problem an the matching of moments step in EP message passing gets much more complicated.

Finally, we focused on a MoG where the lower order terms in the correction R are tractable. For models where this is not the case, R can be expanded in terms of the higher order cumulants of $q_n(\theta)$ and $q(\theta)$. This approach will be presented in a companion paper.

Acknowledgments

We thank Zoubin Ghahramani for suggesting the generalization presented in Appendix E.2. We thank Tom Minka, Markus Svensén, and Neil Lawrence for helpful discussions and comments, and the anonymous reviewers for their comments.

Appendix A. Bounds on the Marginal Likelihood

It is interesting to compare the marginal likelihood approximation (5) with the one given by a variational approximation. Here one would use the fact that the relative entropy $\left\langle \log \frac{q(\theta)}{p(\theta|\mathcal{D})} \right\rangle_q \geq 0$ to approximate the free energy $-\log Z$ by the upper bound

$$-\log Z \leq \left\langle \log \left(\frac{q(\theta)}{p(\theta)\ \prod_n p(x_n|\theta)} \right) \right\rangle_q .$$

To compare with (5) we use the definition (3) to get

$$\frac{Z(\Lambda - \Lambda_n, 1_n)}{Z(\Lambda, 0)} = \left\langle p(x_n|\theta) \exp\Big(-\Lambda_n^T \phi(\theta) \Big) \right\rangle_q .$$

After inserting this expression into (5), taking logs and applying Jensen's inequality we arrive at

$$-\log Z_{\mathrm{EC}} \leq \left\langle \log \left(\frac{\exp\Big(\sum_n \Lambda_n^T \phi(\theta) \Big)}{\prod_n p(x_n|\theta)} \right) \right\rangle_q - \log Z(\Lambda, 0) = \left\langle \log \left(\frac{q(\theta)}{p(\theta)\ \prod_n p(x_n|\theta)} \right) \right\rangle_q ,$$

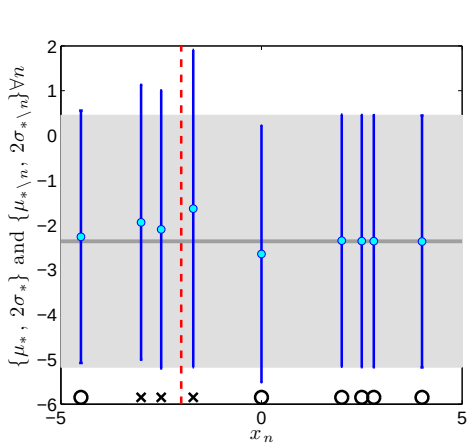
where in the last step we have used $\sum_n \Lambda_n = \Lambda$. This shows that if one would use the distribution $q(\theta)$ derived from EC within the variational approximation, the EC approximation achieves a lower free energy. Since approximating densities in the variational approximation will usually differ from the EC result (by the way they are optimised) this does not imply that variational free energies are always higher than the EC counterpart. Also we cannot draw any conclusion about the relation to the true free energy.

Appendix B. Gaussian Process Classification

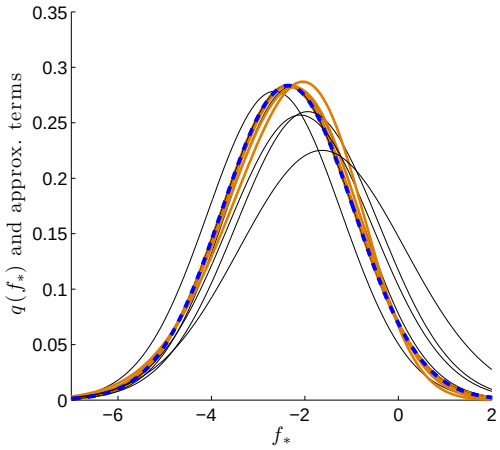
This appendix provides the details of the derivation of first order correction to marginal distribution $p(f_*|\mathcal{D})$ for Gaussian process classification, as introduced in Section 3. Let k_* be the kernel vector with entries $k(x_*, x_n)$ for all n, and $\kappa_* = k(x_*, x_*)$. It is well-established that

$$\begin{aligned} q(f_*) &= \mathcal{N}(f_*\,;\, \mu_*,\, \sigma_*^2), \\ \mu_* &= k_*^T K^{-1} \mu, \\ \sigma_*^2 &= \kappa_* - k_*^T (K + \tilde{S}^{-1})^{-1} k_* \end{aligned} \tag{21}$$

where $p(f_*|f) = \mathcal{N}(f_*\,;\, k_*^\top K^{-1} f,\, \kappa_* - k_*^\top K^{-1} k_*)$ was averaged over $q(f)$. To determine $q_n(f_*)$, we have to average $p(f_*|f)$ over $q_n(f)$: a lengthy derivation shows that the required



(a) The mean μ_* and 2 standard deviations $2\sigma_*$ of the marginal $q(f_*)$ is shown as a gray error bar for a *novel* data point at $x_* = -2$ (dashed red line). Positive examples x_n are indicated with a $\circ$, and negative examples with a $\times$. At each x_n an error bar from $\mathcal{N}(f_*\,;\,\mu_{*\backslash n},\,\sigma^2_{*\backslash n})$ in (22) is shown for comparison. An exclusion of a negative example brings the density of the mean function f_* closer to zero; an exclusion of the the positive example at $x_n = 0$ increases the certainty about the class of x_*.



(b) The distributions $q(f_*)$ and each of the $q_n(f_*)$'s (which are not Gaussian) used in (13) are respectively shown in dashed blue and (thick solid) orange. These distributions are accumulated into the correction shown in Figure 1. To observe the influence of the non-linear "weight-function" of f_* in (22), the distributions $\mathcal{N}(f_*\,;\,\mu_{*\backslash n},\,\sigma^2_{*\backslash n})$ which are shown in Figure 13(a) are plotted as (thin) black lines.

Figure 13: An illustration of all the terms occurring in the first-order correction in (13) for an example data set.

integral $q_n(f_*) = \int df\, p(f_*|f)\, q_n(f)$ simplifies to

$$q_n(f_*) = \Phi\left(\frac{y_n\, m_n(f_*)}{\sqrt{1+V_n}}\right) \Big/ \Phi\left(\frac{y_n\, \mu_{\backslash n;n}}{\sqrt{1+\Sigma_{\backslash n;n,n}}}\right) \times \mathcal{N}(f_*\,;\,\mu_{*\backslash n},\,\sigma^2_{*\backslash n}), \tag{22}$$
$$\mu_{*\backslash n} = k_*^T K^{-1} \mu_{\backslash n},$$
$$\sigma^2_{*\backslash n} = \kappa_* - k_*^T (K + \tilde{S}^{-1}_{\backslash n})^{-1} k_* \ .$$

This "tilted" predictive marginal in (22) has exactly the same form as $q(f_*)$ in (21), except[8] for its use of $\mu_{\backslash n}$ and $\tilde{S}_{\backslash n}$, and the nonlinear "weight" that is still a function of f_*, so that $q_n(f_*)$ is ultimately non-Gaussian. $\Sigma_{\backslash n;n,n}$ and $\mu_{\backslash n;n}$ denote elements (n,n) and n in $\Sigma_{\backslash n}$ and $\mu_{\backslash n}$.

In the ratio of cumulative Normals in (22) we have

$$V_n = \Sigma_{\backslash n;n,n} - \frac{\eta^2}{\sigma^2_{*\backslash n}},$$

8. This representation is chosen for simplicity, although $\tilde{S}_{\backslash n}$ contains a zero on its diagonal.

$$m_n(f_*) = \mu_{\backslash n;n} + \frac{\eta(f_* - \mu_{*\backslash n})}{\sigma^2_{*\backslash n}} \ ,$$

where we define $\eta = k_*^T K^{-1} c_n$, with c_n being column n of $\Sigma_{\backslash n}$.

When comparing V_n and $\Sigma_{\backslash n;n,n}$ in the numerator and denominator in (22), we see that V_n is close to $\Sigma_{\backslash n;n,n}$ whenever η is small compared to $\sigma_{*\backslash n}$. Function $m_n(f_*)$ adjusts $\mu_{\backslash n;n}$ with a term *linear* in how far f_* differs from the Gaussian mean in (22), and is similarly close to $\mu_{\backslash n;n}$ when η is small compared to $\sigma^2_{*\backslash n}$.

Figure 13 provides an illustration of how the correction in (13) works. A squared exponential kernel $k(x_n, x_{n'}) = a \exp(-\frac{1}{2}\|x_n - x_{n'}\|^2/\ell^2)$ was used, with a being the (positive) amplitude, and ℓ the characteristic length-scale of the latent function. In Figure 13 the marginals $q(f_*)$ and $q_n(f_*)$ are shown, leading to the first-order correction originally shown in Figure 1.

Appendix C. Useful Distributions in the Exponential Family

In this paper we use the Dirichlet, Normal-Gamma and Normal-Wishart distributions for the MoG problem. For these distribution have to 1) compute their sufficient statistics, 2) for message passing solve the inverse problem: given the sufficient statistics we must solve for the parameters of the distribution and 3) for the predictions with the mixture models compute their predictive distribution.

C.1 Dirichlet

The Dirichlet distribution over the probability simplex π, $\sum_k \pi_k = 1$, is commonly used in two contexts: here as a prior over mixing proportions in the mixture model, and as a prior/posterior for the parameters of multinomial distribution. We denote the Dirichlet with parameters λ_k by

$$\mathcal{D}(\pi; \lambda) = \frac{1}{Z_{\mathcal{D}}(\lambda)} e^{\sum_k (\lambda_k - 1) \log \pi_k} \delta\left(\sum_k \pi_k - 1\right) \ , \tag{23}$$

$$Z_{\mathcal{D}}(\lambda) = \frac{\prod_k \Gamma(\lambda_k)}{\Gamma(\sum_k \lambda_k)} \ . \tag{24}$$

C.1.1 Sufficient Statistics

The sufficient statistic of the Dirichlet is

$$\langle \log \pi_k \rangle = \frac{\partial \log Z_{\mathcal{D}}(\lambda)}{\partial \lambda_k} = \psi(\lambda_k) - \psi\left(\sum_k \lambda_k\right) \ , \tag{25}$$

where ψ is the digamma-function $d \log \Gamma(x)/dx$.

C.1.2 Inverse

In line 5 of the Algorithms 1 and 2 we have to solve the inverse problem: given the statistics $m_k = \langle \log \pi_k \rangle$ find the parameters λ. This can be done effectively by first solving for

$$\alpha \equiv \psi\left(\sum_k \lambda_k\right) = \psi\left(\sum_k \psi^{-1}(m_k + \alpha)\right)$$

by Newton's method, and then setting $\lambda_k := \psi^{-1}(m_k + \alpha)$.

C.1.3 Predictive Distribution

The Dirichlet can also be used as a prior for the parameters of the multinomial distribution. This distribution is used multi-category either counts or sequence data. For counts, $x = (x_1, \ldots, x_d)$ is a vector of counts for each of the possible d outcomes. For sequence data x is an indicator variable being one for the outcome and zero in all other entries. The multinomial distribution is:

$$p(x|\pi) = \frac{(\sum_k x_k)!}{x_1! \ldots x_d!} \prod_{k=1}^{d} \pi_k^{x_k} \ ,$$

where the combinatorial prefactor goes away in the sequence case. The predictive distribution for multinomial data and Dirichlet distributed posterior is straightforward to calculate using the result for the normalizer of the Dirichlet:

$$p(x|\lambda) = \int d\pi p(x|\pi)\mathcal{D}(\pi;\lambda) = \frac{(\sum_l x_l)!}{x_1! \ldots x_d!} \frac{Z_{\mathcal{D}}(x+\lambda)}{Z_{\mathcal{D}}(\lambda)} \ .$$

C.2 Normal-Gamma

The Normal-Gamma (or Gauss-Gamma) model is use for a joint distribution of one dimensional mean and precision (inverse variance) variables:

$$\mathcal{NG}(\mu,\gamma;m,\nu,a,b) = \frac{1}{Z_{\mathcal{NG}}(m,\nu,a,b)} \exp\left[\begin{pmatrix} m\nu \\ -\frac{1}{2}\nu \\ -b - \frac{1}{2}\nu m^2 \\ a - \frac{1}{2} \end{pmatrix}^T \begin{pmatrix} \mu\gamma \\ \mu^2\gamma \\ \gamma \\ \log\gamma \end{pmatrix}\right],$$

$$Z_{\mathcal{NG}}(m,\nu,a,b) = \sqrt{\frac{2\pi}{\nu}} \frac{\Gamma(a)}{b^a} \ ,$$

where this distribution is obtained by multiplying the Normal and Gamma distributions:

$$\mathcal{N}(\mu;m,(\nu\gamma)^{-1}) = \sqrt{\frac{\nu\gamma}{2\pi}} \exp\left(-\frac{1}{2}(\mu-m)^2\nu\gamma\right),$$

$$\mathcal{G}(\gamma;a,b) = \frac{b^a}{\Gamma(a)} \exp((a-1)\log\gamma - b\gamma) \ ,$$

where a and b must be positive.

C.2.1 Sufficient Statistics

The sufficient statistics are obtained by using $\log Z_{\mathcal{NG}}(m,\nu,a,b)$ as a generating function for the sufficient statistic. By taking derivatives of $\log Z_{\mathcal{NG}}$ with respect to the parameters $\{m,\nu,a,b\}$,

$$\begin{aligned}
\nu\langle\mu\gamma\rangle - \nu m\langle\gamma\rangle &= 0,\\
m\langle\mu\gamma\rangle - \frac{1}{2}\langle\mu^2\gamma\rangle - m^2\langle\gamma\rangle &= -\frac{1}{2\nu},\\
\langle\log\gamma\rangle &= \psi(a) - \log b,\\
-\langle\gamma\rangle &= -a/b\ ,
\end{aligned}$$

we can solve for $\langle\mu\gamma\rangle$, $\langle\mu^2\gamma\rangle$, $\langle\gamma\rangle$ and $\langle\log\gamma\rangle$.

C.2.2 Inverse

We can use these expressions to solve for the parameters in terms of the sufficient statistics in the same fashion as above. We get closed form expressions for three of parameters

$$m = \frac{\langle\mu\gamma\rangle}{\langle\gamma\rangle}\ , \quad \nu = \frac{1}{\langle\mu^2\gamma\rangle - m^2\langle\gamma\rangle}\ , \quad b = \frac{a}{\langle\gamma\rangle}\ ,$$

and a should be found from

$$\psi(a) - \log a = \langle\log\gamma\rangle - \log\langle\gamma\rangle$$

by for example Newton's method.

C.2.3 Predictive Distribution

The predictive distribution can be calculated straightforwardly from the normalizer and is a univariate Student-t distribution:

$$\begin{aligned}
p(x|m,\nu,a,b) &= \frac{1}{\sqrt{2\pi}}\frac{Z_{\mathcal{NG}}\left(\frac{x+\nu m}{\nu+1}, \nu+1, a+\frac{1}{2}, b+\frac{\nu}{\nu+1}(x-m)^2\right)}{Z_{\mathcal{NG}}(m,\nu,a,b)}\\
&= \mathcal{T}\left(x;\ m,\ \frac{b}{a}\frac{\nu+1}{\nu},\ 2a\right)\ ,
\end{aligned}$$

where $\mathcal{T}\left(x;\mu,\sigma^2,d_{\mathrm{f}}\right)$ is a Student-t distribution with mean μ, variance σ^2 and d_{f} degrees of freedom:

$$\begin{aligned}
\mathcal{T}\left(x;\mu,\sigma^2,d_{\mathrm{f}}\right) &= \frac{1}{Z_{\mathcal{T}}(\mu,\sigma^2,d_{\mathrm{f}})}\exp\left[-\frac{d_{\mathrm{f}}+1}{2}\log\left(1+\frac{1}{d_{\mathrm{f}}}\left(\frac{x-\mu}{\sigma}\right)^2\right)\right],\\
Z_{\mathcal{T}}(\mu,\sigma^2,d_{\mathrm{f}}) &= \sqrt{2\pi\sigma^2}\sqrt{\frac{d_{\mathrm{f}}}{2}}\frac{\Gamma\left(\frac{d_{\mathrm{f}}}{2}\right)}{\Gamma\left(\frac{d_{\mathrm{f}}+1}{2}\right)}\ .
\end{aligned}$$

C.3 Normal-Wishart

The Normal-Wishart is the multidimensional generalization of the Normal-Gamma. We will write the Wishart distribution over positive definite symmetric matrices in the same form as the Gamma distribution:

$$\mathcal{W}(\Gamma; a, B) \propto \exp\left(\left(a - \frac{d+1}{2}\right)\log\det\Gamma - \operatorname{tr} B\Gamma\right) ,$$

where the degrees of freedom $2a$ should be greater than $d-1$ for the distribution to be normalizable. The Normal-Wishart is given by

$$\mathcal{NW}(\mu, \Gamma; m, \nu, a, B) = \frac{1}{Z_{\mathcal{NW}}(m, \nu, a, B)} \exp\left[\begin{pmatrix} \nu m \\ -\frac{1}{2}\nu \\ a - \frac{d}{2} \end{pmatrix}^T \begin{pmatrix} \Gamma\mu \\ \mu^T\Gamma\mu \\ \log\det\Gamma \end{pmatrix} - \operatorname{tr}(B + \frac{1}{2}\nu m m^T)\Gamma\right] ,$$

$$Z_{\mathcal{NW}}(m, \nu, a, B) = \pi^{d(d-1)/4}\left(\frac{2\pi}{\nu}\right)^{d/2}\prod_{l=1}^{d}\Gamma\left(a + \frac{1-l}{2}\right)e^{-a\log\det B} .$$

C.3.1 Sufficient Statistics

The sufficient statistics follow from a straightforward generalization of the results from the Normal-Gamma model:

$$\begin{aligned} \langle\Gamma\rangle &= aB^{-1}, \\ \langle\Gamma\mu\rangle &= \langle\Gamma\rangle m, \\ \langle\mu^T\Gamma\mu\rangle &= \frac{d}{\nu} + m^T\langle\Gamma\rangle m, \\ \langle\log\det\Gamma\rangle &= \sum_{l=1}^{d}\psi\left(a + \frac{1-l}{2}\right) - \log\det B . \end{aligned}$$

C.3.2 Inverse

From the sufficient statistics we can get closed form expressions for the parameters

$$m = \langle\Gamma\rangle^{-1}\langle\Gamma\mu\rangle , \quad \nu = \frac{d}{\langle\mu^T\Gamma\mu\rangle - m^T\langle\Gamma\rangle m} , \quad B = a\langle\Gamma\rangle^{-1} ,$$

whereas a should be found from

$$\sum_{l=1}^{d}\psi\left(a + \frac{1-l}{2}\right) - d\log a = \langle\log\det\Gamma\rangle - \log\det\langle\Gamma\rangle .$$

C.3.3 Predictive Distribution

A generalization of the result for the predictive distribution in one dimension to the multivariate case gives:

$$p(x|m, \nu, a, b) = \mathcal{T}\left(x;\ m,\ \frac{2B}{2a - d + 1}\frac{\nu+1}{\nu},\ 2a - d + 1\right) ,$$

where $\mathcal{T}(x;\mu,\Sigma,d_\mathrm{f})$ is the d-dimensional multivariate Student-t distribution with mean μ, covariance Σ and d_f degrees of freedom:

$$\mathcal{T}(x;\mu,\Sigma,d_\mathrm{f}) = \frac{1}{Z_\mathcal{T}(\mu,\Sigma,d_\mathrm{f})}\exp\left[-\frac{d_\mathrm{f}+d}{2}\log\left(1+\frac{1}{d_\mathrm{f}}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right],$$

$$Z_\mathcal{T}(\mu,\sigma^2,d_\mathrm{f}) = \sqrt{\det(2\pi\Sigma)}\left(\frac{d_\mathrm{f}}{2}\right)^{d/2}\frac{\Gamma\left(\frac{d_\mathrm{f}}{2}\right)}{\Gamma\left(\frac{d_\mathrm{f}+d}{2}\right)}\,.$$

Appendix D. Inference for a Mixture of Gaussians

When we have unconstrained d-dimensional data we can model this with a Normal (or Gaussian) distribution,

$$p(x|\mu,\Gamma) = \mathcal{N}(x;\mu,\Gamma^{-1}) = \sqrt{\frac{\det\Gamma}{(2\pi)^d}}\exp\left(-\frac{1}{2}(x-\mu)^T\Gamma(x-\mu)\right)\,,$$

where μ and Γ are respectively the mean vector and precision (or inverse covariance) matrix. The conjugate prior for the mean and precision is the Normal-Wishart distribution, which we describe in appendix C. In the following we choose Gaussians $p(x|\mu_k,\Gamma_k)$ as components densities $p(x|k)$ in the mixture.

D.1 Variational and Predictive Distributions

The q distribution follows the prior and is conveniently chosen to factorise over mixture components:

$$q(\theta) = q(\pi)\prod_k q(\mu_k,\Gamma_k)$$

with $q(\mu_k,\Gamma_k|m_k,\nu_k,a_k,B_k)$ being a Normal-Wishart distribution; see Appendix C. We can use the same machinery as Section 4 to derive the EC approximation to the predictive distribution and the statistics needed for message passing.

The predictive distribution $p(x|\mathcal{D})$ is again approximated by the form given in (14), with $p(x|k)$ replaced by

$$p(x|k) \equiv p(x|m_k,\nu_k,a_k,B_k) = \mathcal{T}\left(x;\ m_k,\ \frac{2B_k}{2a_k-d+1}\frac{\nu_k+1}{\nu_k},\ 2a_k-d+1\right)\,. \qquad (26)$$

Here $\mathcal{T}$ is a Student-t distribution, which we describe in greater detail in Appendix C.

Likewise, the within data set predictive distribution $p(x_n|\mathcal{D}_{\backslash n})$ follows (15); only now $p(x_n|k)$ is replaced with $p(x_n|k\backslash n)$, which takes the same form as (26) above, with Λ replaced by $\Lambda_{\backslash n}$. The Dirichlet cavity parameters are again $\lambda_{k\backslash n} = \lambda_{k,0} + \sum_{n'\neq n}\lambda_{k,n'}$. The other cavity parameters are similarly defined in terms of the appropriate parameter vector components, for instance $\nu_{k\backslash n}m_{k\backslash n} = \nu_{k,0}m_{k,0} + \sum_{n'\neq n}\nu_{k,n'}m_{k,n'}$.

We can again use the cavity parameters to write q_n in terms of $q_{\backslash n}$:

$$q_n(\theta) = \frac{Z(\Lambda_{\backslash n},0)}{Z_n(\Lambda_{\backslash n},1_n)}\,q_{\backslash n}(\theta)\,\sum_k \pi_k p(x_n|\mu_k,\Gamma_k)\,.$$

The normalizer is given by

$$Z_n(\Lambda_{\backslash n}, 1_n) = Z(\Lambda_{\backslash n}, 0) \sum_k \langle \pi_k \rangle_{q_{\backslash n}} p(x_n|k\backslash n) \ , \tag{27}$$

where the explicit form of $Z(\Lambda_{\backslash n}, 0)$ is

$$Z(\Lambda_{\backslash n}, 0) = Z_{\mathcal{D}}(\lambda_{\backslash n}) \prod_k Z_{\mathcal{NW}}(m_{k\backslash n}, \nu_{k\backslash n}, a_{k\backslash n}, B_{k\backslash n}) \ .$$

D.2 Expectations

The statistics of $q_n(\theta)$ for the mixture of Gaussians are computed by using $\log Z_n(\Lambda_{\backslash n})$ from (27) as a generating function. To simplify the derivative with respect to the predictive distribution $p(x_n|k\backslash n)$, we introduce another component-specific Normal-Wishart distribution

$$q_{k,n}(\mu_k, \Gamma_k) \propto p(x_n|\mu_k, \Gamma_k) q_{\backslash n}(\mu_k, \Gamma_k) \ ,$$

and write the predictive distribution as a ratio between the normalizers of $q_{k,n}$ and $q_{\backslash n}$:

$$p(x_n|k\backslash n) = \frac{(2\pi)^{-d/2} Z_{\mathcal{NW}}\big(\frac{\nu_{k\backslash n} m_{k\backslash n} + x_n}{\nu_{k\backslash n}+1}, \nu_{k\backslash n}+1, a_{k\backslash n}+\frac{1}{2}, B_{k\backslash n} + \frac{1}{2}\frac{\nu_{k\backslash n}}{\nu_{k\backslash n}+1}(x_n - m_{k\backslash n})(x_n - m_{k\backslash n})^T\big)}{Z_{\mathcal{NW}}(m_{k\backslash n}, \nu_{k\backslash n}, a_{k\backslash n}, B_{k\backslash n})} \ .$$

For example, for $\langle \Gamma_k \mu_k \rangle$ we get

$$\big\langle \Gamma_k \mu_k \big\rangle_{q_n} = \frac{1}{\nu_{k\backslash n}} \frac{d \log Z_n(\Lambda_{\backslash n})}{d m_{k\backslash n}} = (1 - r_{nk}) \big\langle \Gamma_k \mu_k \big\rangle_{q_{\backslash n}} + r_{nk} \big\langle \Gamma_k \mu_k \big\rangle_{q_{k,n}} \ , \tag{28}$$

where the "responsibility" (the probability of example n being generated by the kth mixture component) is defined as

$$r_{nk} = \frac{\lambda_{k\backslash n} p(x_n|k\backslash n)}{\sum_{k'} \lambda_{k'\backslash n} p(x_n|k'\backslash n)} \ . \tag{29}$$

The expectation in (28) is expressed as a weighed sum of a "prior" expectation over the cavity distribution $q_{\backslash n}(\mu_k, \Gamma_k)$, and a "posterior" expectation over $q_{k,n}(\mu_k, \Gamma_k)$. The other moments $\langle \Gamma_k \rangle$, $\langle \mu_k^T \Gamma_k \mu_k \rangle$ and $\langle \log \det \Gamma_k \rangle$ can be expressed as weighed sums similar to (28):

$$\begin{aligned}
\big\langle \Gamma_k \big\rangle_{q_n} &= (1 - r_{nk}) \big\langle \Gamma_k \big\rangle_{q_{\backslash n}} + r_{nk} \big\langle \Gamma_k \big\rangle_{q_{k,n}} , \\
\big\langle \mu_k^T \Gamma_k \mu_k \big\rangle_{q_n} &= (1 - r_{nk}) \big\langle \mu_k^T \Gamma_k \mu_k \big\rangle_{q_{\backslash n}} + r_{nk} \big\langle \mu_k^T \Gamma_k \mu_k \big\rangle_{q_{k,n}} , \\
\big\langle \log \det \Gamma_k \big\rangle_{q_n} &= (1 - r_{nk}) \big\langle \log \det \Gamma_k \big\rangle_{q_{\backslash n}} + r_{nk} \big\langle \log \det \Gamma_k \big\rangle_{q_{k,n}} \ .
\end{aligned}$$

The explicit expressions for the $q_{k,n}$ are given below, whereas those for $q_{\backslash n}$ can be obtained from Appendix C:

$$\big\langle \Gamma_k \big\rangle_{q_{k,n}} = \left(a_{k\backslash n} + \frac{1}{2}\right) \left[B_{k\backslash n} + \frac{1}{2} \frac{\nu_{k\backslash n}}{\nu_{k\backslash n}+1} (x_n - m_{k\backslash n})(x_n - m_{k\backslash n})^T\right]^{-1} ,$$

$$\begin{aligned}
\left\langle \Gamma_k \mu_k \right\rangle_{q_{k,n}} &= \left\langle \Gamma_k \right\rangle_{q_{k,n}} \frac{\nu_{k\backslash n} m_{k\backslash n} + x_n}{\nu_{k\backslash n} + 1}, \\
\left\langle \mu_k^T \Gamma_k \mu_k \right\rangle_{q_{k,n}} &= \frac{d}{v_{k\backslash n} + 1} + \left(\frac{\nu_{k\backslash n} m_{k\backslash n} + x_n}{\nu_{k\backslash n} + 1} \right)^T \left\langle \Gamma_k \right\rangle_{q_{k,n}} \left(\frac{\nu_{k\backslash n} m_{k\backslash n} + x_n}{v_{k\backslash n} + 1} \right), \\
\left\langle \log\det \Gamma_k \right\rangle_{q_{k,n}} &= \sum_{i=1}^{d} \psi\left(a_{k\backslash n} + \frac{1}{2} + \frac{1-i}{2} \right) \\
&\qquad - \log\det\left(B_{k\backslash n} + \frac{1}{2} \frac{\nu_{k\backslash n}}{\nu_{k\backslash n}+1} (x_n - m_{k\backslash n})(x_n - m_{k\backslash n})^T \right) .
\end{aligned}$$

We have already seen how to solve for $\langle \log \pi_k \rangle_{q_n}$, the only difference being r_{nk} in (16), which we now take from (29).

Appendix E. Gibbs Sampling for Parallel Tempering

Parallel tempering of a mixture of Gaussian distributions $p(x_n|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Gamma_k^{-1})$ requires a Monte Carlo simulation at inverse temperature β. We can either sample from $p(\theta|\mathcal{D}, \beta)$ using a Metropolis-Hastings (MH) method, or augment the parameter space with latent allocation variables z so that we can sample from $p(\theta, z|\mathcal{D}, \beta)$ with Gibbs sampling. We may also define $p(z|\mathcal{D}, \beta) = \int d\theta\, p(\theta, z|\mathcal{D}, \beta)$, and devise a MH scheme on this distribution by making random assignment changes to z.

The road of Gibbs sampling is pathed with tractable conditional distributions of $p(\theta, z|\mathcal{D}, \beta)$, and it is the one we choose. We extend the parameter space to include a binary latent allocation vector z_n for each data point n to indicate which mixture component was responsible for generating it (Diebolt and Robert, 1994); consequently $z_{nk} \in \{0, 1\}$, and $\sum_k z_{nk} = 1$. The complete joint distribution is therefore

$$p(\mathcal{D}, z|\theta)p(\theta) = \prod_n \prod_k \left[\pi_k \mathcal{N}(x_n; \mu_k, \Gamma_k^{-1}) \right]^{z_{nk}} p(\theta) .$$

We can write the complete data likelihood as $p(\mathcal{D}, z|\theta) = p(\mathcal{D}|z, \theta)p(z|\theta)$, and in this form the likelihood, to the power β, multiplied by the prior over z and θ, is

$$p(\mathcal{D}|z, \theta)^\beta \times p(z, \theta) = \prod_n \prod_k \mathcal{N}(x_n; \mu_j, \Gamma_k^{-1})^{\beta z_{nk}} \times \prod_n \prod_k \pi_k^{z_{nk}} p(\theta) .$$

(Note that the introduction of z moves π to the prior.) With inverse temperature parameter β the tempered posterior distribution is $p(\theta, z|\mathcal{D}, \beta) \propto p(\mathcal{D}|z, \theta)^\beta p(z, \theta)$, and can be treated as any missing-value Gibbs sampling problem. The allocation variables are sampled with

$$z_{nk}|\pi, \mu, \Gamma \sim \frac{\pi_k \mathcal{N}(x_n; \mu_k, \Gamma_k^{-1})^\beta}{\sum_{k'} \pi_{k'} \mathcal{N}(x_n; \mu_{k'}, \Gamma_{k'}^{-1})^\beta} .$$

Given the allocation variables, we define

$$r_{nk} = \beta z_{nk}, \qquad \bar{x}_k = \frac{1}{N_k} \sum_n r_{nk} x_n,$$

$$N_k = \sum_n r_{nk} \qquad\qquad \Sigma_k, = \frac{1}{N_k}\sum_n r_{nk}(x_n - \bar{x}_k)(x_n - \bar{x}_k)^T \ ,$$

to give the conditional distributions needed for sampling the mixture parameters as

$$\begin{aligned} \pi|z &\sim \mathcal{D}\big(\pi; \lambda_{1,0} + \tfrac{1}{\beta}N_1, \ldots, \lambda_{K,0} + \tfrac{1}{\beta}N_K\big), \\ \mu_k, \Gamma_k|z &\sim \mathcal{NW}\big(\mu_k, \Gamma_k; m, \nu, a, B\big) \ , \end{aligned} \tag{30}$$

with

$$\begin{aligned} m &= \frac{\nu_{k,0}m_{k,0} + N_k\bar{x}_k}{\nu_{k,0} + N_k}, \\ \nu &= \nu_{k,0} + N_k, \\ a &= a_{k,0} + N_k/2, \\ B &= B_{k,0} + \frac{1}{2}N_k\Sigma_k + \frac{1}{2}\frac{N_k v_{k,0}(\bar{x}_k - m_{k,0})(\bar{x}_k - m_{k,0})^T}{\nu_{k,0} + N_k} \ . \end{aligned} \tag{31}$$

As $p(\mathcal{D}) = \int d\theta dz\, p(\mathcal{D}, \theta, z)$, we use the samples over θ and z to estimate the average log likelihood. If $\{\{\pi_{k,i}^{(t)}, \mu_{k,i}^{(t)}, \Gamma_{k,i}^{(t)}\}_{k=1}^K, \{z_{n,i}^{(t)}\}_{n=1}^N\}_{t=1}^T$ indicates the samples of chain i (after a burn-in period), then

$$\langle \log p(\mathcal{D}|\theta, z)\rangle_{\beta_i} \approx \frac{1}{T}\sum_t \sum_n \sum_k z_{nk,i}^{(t)} \log \mathcal{N}\big(x_n; \mu_{k,i}^{(t)}, \Gamma_{k,i}^{(t)^{-1}}\big) \ . \tag{32}$$

Notice that the samples of the mixing weights $\pi_{k,i}^{(t)}$ are not used in estimating the log likelihood average over the posterior, but occur in the prior.

E.1 A Practical Generalization (I)

The generalization of PT and TI from Section 5.3 can be made by writing the tempered posterior distribution as

$$p(\theta, z|\mathcal{D}, \beta) \propto \prod_n \prod_k \mathcal{N}(x_n; \mu_k, \Gamma_k^{-1})^{\beta z_{nk}}\, \pi_k^{z_{nk}}\, p(\theta)^\beta\, q(\theta)^{1-\beta} \ ,$$

where $\prod_{n,k} \pi_k^{z_{nk}} = p(z|\theta)^\beta q(z|\theta)^{1-\beta}$ follows from $p(z|\theta) = q(z|\theta)$. To do Gibbs sampling as before, we have to determine the parameters of the "effective" prior in the above scenario. Here we let $q(\theta)$ be in the same family—for example a narrower version—of the prior. If superscripts p and q now differentiate between the parameters of $p(\theta)$ and $q(\theta)$, we use

$$\begin{aligned} \lambda_0 &= \beta\lambda_0^p + (1-\beta)\lambda_0^q, \\ m_{k,0} &= \frac{\beta\nu_{k,0}^p m_{k,0}^p + (1-\beta)\nu_{k,0}^q m_{k,0}^q}{\beta\nu_{k,0}^p + (1-\beta)\nu_{k,0}^q}, \\ \nu_{k,0} &= \beta\nu_{k,0}^p + (1-\beta)\nu_{k,0}^q, \\ a_{k,0} &= \beta a_{k,0}^p + (1-\beta)a_{k,0}^q, \end{aligned}$$

$$B_{k,0} = \beta B^p_{k,0} + (1-\beta) B^q_{k,0} + \frac{1}{2} \frac{\beta\nu^p_{k,0}(1-\beta)\nu^q_{k,0}}{\beta\nu^p_{k,0} + (1-\beta)\nu^q_{k,0}} (m^p_{k,0} - m^q_{k,0})(m^p_{k,0} - m^q_{k,0})^T$$

as substitute for the usual prior in (30) and (31). The empirical expectation given in (32) should be generalized—simplifying the left hand side below with $p(z,\theta)/q(z,\theta) = p(\theta)/q(\theta)$—to

$$\Big\langle \log p(\mathcal{D}|\theta, z) + \log \frac{p(\theta)}{q(\theta)} \Big\rangle_{\beta_i} \approx \frac{1}{T} \sum_t \Bigg[\log p(\pi^{(t)}_i) - \log q(\pi^{(t)}_i) + \sum_k \Big[\log p(\mu^{(t)}_{k,i}, \Gamma^{(t)}_{k,i}) - \log q(\mu^{(t)}_{k,i}, \Gamma^{(t)}_{k,i}) + \sum_n z^{(t)}_{nk,i} \log \mathcal{N}\big(x_n; \mu^{(t)}_{k,i}, {\Gamma^{(t)}_{k,i}}^{-1}\big) \Big] \Bigg] .$$

E.2 A Practical Generalization (II)

We implemented a further possible generalization, which arises from choosing $q(\theta) = p(\theta|\mathcal{D}')$, where $\mathcal{D}'$ contains a small subset of data points from $\mathcal{D}$. This sensibly restricts q to parameter space closer to the posterior, with the benefit that the normalizer of q needs to be calculated only once. With $|\mathcal{D}'|$ being small, q can be evaluated analytically without feeling the effect of the exponentially expanding number of terms. This brings an interesting tradeoff, as setting $\mathcal{D}' \leftarrow \mathcal{D}$ solves our original (difficult) problem. (Figure 3 used this choice of surrogate prior, with $\mathcal{D}'$ containing 3 out of a possible 82 data points.)

We can construct $q(\theta)$ as follows: Let $N' = |\mathcal{D}'|$ be the number of data points in $\mathcal{D}'$, so that $q(\theta)$ expands as a sum over $(N')^K$ terms. Allow $1, \ldots, K$ to be the digit set of a number system in base K. Make a list $\mathcal{S}$ of the first $(N')^K$ numbers in base K, such that each number s consists of N' digits, and each $x_{n'} \in \mathcal{D}'$ can be associated with a corresponding digit *position*. Each $s \in \mathcal{S}$ therefore defines a unique allocation for all $x_{n'} \in \mathcal{D}'$ to clusters $1, \ldots, K$ ($x_{n'}$'s digit *value*). We shall use the shorthand $\mathcal{D}'_s$ to indicate a data set with a data point to cluster allocation given by s.

The surrogate prior is a weighted sum of various posteriors

$$q(\theta) = p(\theta|\mathcal{D}') = \sum_{s\in\mathcal{S}} w_s p(\pi|\mathcal{D}'_s) \prod_k p(\mu_k, \Gamma_k|\mathcal{D}'_s) .$$

If $Z_s = \int d\theta\, p(\theta, \mathcal{D}'_s)$, the weights are determined with $w_s = Z_s / \sum_{s'} Z_{s'}$.

Instead of merely raising q to the power $1-\beta$, we first turn q into a product amenable to Gibbs sampling by augmenting it with binary indicator variables $y = \{y_s\}_{s\in\mathcal{S}}$ that pick a particular component of q; $\sum_s y_s = 1$. With $q(y) = p(y) = \prod_s w_s^{y_s}$, the surrogate prior now takes the form $q(z|\theta)q(\theta|y)q(y)$; the prior becomes $p(z|\theta)p(\theta)p(y)$. The empirical expectation in (32) generalizes to

$$\Big\langle \log p(\mathcal{D}|\theta, z) + \log \frac{p(\theta)}{q(\theta|y)} \Big\rangle_{\beta_i} \approx \frac{1}{T} \sum_t \Bigg[\log p(\pi^{(t)}_i) - \sum_s y^{(t)}_{s,i} \log p(\pi^{(t)}_i|\mathcal{D}'_s) + \sum_k \Big[\log p(\mu^{(t)}_{k,i}, \Gamma^{(t)}_{k,i}) - \sum_s y^{(t)}_{s,i} \log p(\mu^{(t)}_{k,i}, \Gamma^{(t)}_{k,i}|\mathcal{D}'_s) + \sum_n z^{(t)}_{nk,i} \log \mathcal{N}\big(x_n; \mu^{(t)}_{k,i}, {\Gamma^{(t)}_{k,i}}^{-1}\big) \Big] \Bigg] .$$

Appendix F. Perturbation Corrections for Mixture of Gaussians

In this appendix we show how to compute the second-order terms of (8),

$$R = 1 + \sum_{n_1<n_2} \Big\langle \varepsilon_{n_1}(\theta)\varepsilon_{n_2}(\theta) \Big\rangle_q + \sum_{n_1<n_2<n_3} \Big\langle \varepsilon_{n_1}(\theta)\varepsilon_{n_2}(\theta)\varepsilon_{n_3}(\theta) \Big\rangle_q + \dots \ , \tag{33}$$

and the first-order term in the numerator of (10),

$$p(x|\mathcal{D}) = \frac{\int d\theta\, q(\theta)\, p(x|\theta)\ \big(1 + \sum_n \varepsilon_n(\theta) + \sum_{n_1<n_2} \varepsilon_{n_1}(\theta)\varepsilon_{n_2}(\theta) + \dots\big)}{1 + \sum_{n_1<n_2} \langle \varepsilon_{n_1}(\theta)\varepsilon_{n_2}(\theta)\rangle_q + \dots} \ . \tag{34}$$

The sum in the second order term runs over all distinct pairs and the complexity thus grows as $\mathcal{O}(N^2)$. However, one would expect that the largest contribution comes from nearby points, or more precisely points that belong to the same component, as indicated by a large responsibility for the same component. Although not done here, it is plausible to restrict the summation to only include these pairs without sacrificing much precision.

Let $\Lambda = \{\lambda, \{m_k, v_k, a_k, B_k\}_{k=1}^K\}$ be the parameters that solve the EC equations. We also have access to the parameters of each of the cavity distributions $\Lambda_{\backslash n}$.

For each n the parameters of $q_n(\theta)$ is given by the parameters of $p(x_n|\theta)q_{\backslash n}(\theta)$, which expands as a sum over the K mixture components. Each element k in the sum contains a product of a Dirichlet density and K Normal-Wishart densities. The Dirichlet parameter vector will have element k incremented by one, and as x_n is associated with component k, it will affect only the parameters of the k^{th} Normal-Wishart. Therefore, apart from the cavity parameters $\Lambda_{\backslash n}$, we will also need for each $k = 1, \dots, K$:

$$\begin{aligned}
\lambda^*_{k\backslash n} &= \lambda_{k\backslash n} + 1 \quad \text{and} \quad \lambda^*_{k'\backslash n} = \lambda_{k'\backslash n} \text{ for } k' \neq k, \\
v^*_{k\backslash n} &= v_{k\backslash n} + 1, \\
m^*_{k\backslash n} &= \frac{v_{k\backslash n} m_{k\backslash n} + x_n}{v_{k\backslash n} + 1}, \\
a^*_{k\backslash n} &= a_{k\backslash n} + \tfrac{1}{2}, \\
B^*_{k\backslash n} &= B_{k\backslash n} + \frac{1}{2}\frac{v_{k\backslash n}}{v_{k\backslash n}+1}(m_{k\backslash n} - x_n)(m_{k\backslash n} - x_n)^T \ .
\end{aligned} \tag{35}$$

The normalizer of $q_n(\theta)$ follows from (27) to be $\int d\theta' p(x_n|\theta')q_{\backslash n}(\theta') = \sum_k \langle \pi_k \rangle_{q_{\backslash n}} p(x_n|k\backslash n) = Z_n$.

F.1 Corrections for the Marginal Likelihood

A single second-order term in (33) can be evaluated with

$$\begin{aligned}
\Big\langle \varepsilon_{n_1}(\theta)\varepsilon_{n_2}(\theta) \Big\rangle_q &= \int d\theta\, \frac{q_{n_1}(\theta) q_{n_2}(\theta)}{q(\theta)} - 1 \\
&= -1 + \frac{1}{Z_{n_1}} \frac{1}{Z_{n_2}} (2\pi)^{-d} \sum_{k=1}^K \sum_{l=1}^K \left\{ \frac{Z_{\mathcal{D}}(\lambda)\, Z_{\mathcal{D}}(\lambda^{(k,l)})}{Z_{\mathcal{D}}(\lambda_{\backslash n_1})\, Z_{\mathcal{D}}(\lambda_{\backslash n_2})} \right.
\end{aligned}$$

$$\cdots \times \prod_{j=1}^{K} \frac{Z_{\mathcal{NW}}(m_j, v_j, a_j, B_j)\, Z_{\mathcal{NW}}(m_j^{(k,l)}, v_j^{(k,l)}, a_j^{(k,l)}, B_j^{(k,l)})}{Z_{\mathcal{NW}}(m_{j\backslash n_1}, v_{j\backslash n_1}, a_{j\backslash n_1}, B_{j\backslash n_1})\, Z_{\mathcal{NW}}(m_{j\backslash n_2}, v_{j\backslash n_2}, a_{j\backslash n_2}, B_{j\backslash n_2})} \Bigg\} .$$

The above sum over k relates x_{n_1} to coming from a particular mixture component k, while the sum over l does the same for x_{n_2}. For a particular element in the sum over k and l we need parameters relating to each of the $j = 1, \ldots, K$ mixture components. For the Dirichlet normalizer the parameters $\lambda_j^{(k,l)}$ depend on whether $k = l$, implying that both x_{n_1} and x_{n_2} were generated from the same mixture component, or whether $k \neq l$, implying that x_{n_1} and x_{n_2} came from different mixture components. The elements of $\lambda^{(k,l)}$ are:

$$\lambda_j^{(k,l)} = \lambda_{j\backslash n_1} + \lambda_{j\backslash n_2} - \lambda_j \qquad \text{for } j \neq k, l \ .$$

When $k \neq l$ two indices j remain to be defined; if $k = l$ we will have one remaining index to take care of:

$$\begin{aligned}
\lambda_j^{(k,l)} &= \lambda^*_{j\backslash n_1} + \lambda_{j\backslash n_2} - \lambda_j && \text{for } j = k \text{ and } k \neq l, \\
\lambda_j^{(k,l)} &= \lambda_{j\backslash n_1} + \lambda^*_{j\backslash n_2} - \lambda_j && \text{for } j = l \text{ and } k \neq l, \\
\lambda_j^{(k,l)} &= \lambda^*_{j\backslash n_1} + \lambda^*_{j\backslash n_2} - \lambda_j && \text{for } j = k = l \ .
\end{aligned}$$

For each element in the sum over k and l the Normal-Wishart parameters are similarly defined. When $j \neq k, l$ we have:

$$\begin{aligned}
v_j^{(k,l)} &= v_{j\backslash n_1} + v_{j\backslash n_2} - v_j, \\
m_j^{(k,l)} &= \frac{v_{j\backslash n_1} m_{j\backslash n_1} + v_{j\backslash n_2} m_{j\backslash n_2} - v_j m_j}{v_{j\backslash n_1} + v_{j\backslash n_2} - v_j}, \\
a_j^{(k,l)} &= a_{j\backslash n_1} + a_{j\backslash n_2} - a_j, \\
B_j^{(k,l)} &= B_{j\backslash n_1} + \frac{1}{2} v_{j\backslash n_1} m_{j\backslash n_1} m_{j\backslash n_1}^T + B_{j\backslash n_2} + \frac{1}{2} v_{j\backslash n_2} m_{j\backslash n_2} m_{j\backslash n_2}^T \\
&\quad \cdots - B_j - \frac{1}{2} v_j m_j m_j^T - \frac{1}{2} v_j^{(k,l)} m_j^{(k,l)} m_j^{(k,l)T} \ .
\end{aligned}$$

As was seen for the mixture weights, we will need further definitions: when $j = k$ and $k \neq l$ we shall use $v_j^{(k,l)} = v^*_{j\backslash n_1} + v_{j\backslash n_2} - v_j$; a similar definition follows when $j = l$. Finally, when $j = k = l$ we find that $v_j^{(k,l)} = v^*_{j\backslash n_1} + v^*_{j\backslash n_2} - v_j$. The other Normal-Wishart parameters follow the same route. The correction evaluates in $\mathcal{O}(N^2K^2)$ complexity.

F.2 Corrections for the Predictive Distribution

From (34) we can compute a first-order correction to the predictive distribution with $p(x|\mathcal{D}) \approx \int d\theta\, q(\theta) p(x|\theta)(1 + \sum_n \varepsilon_n(\theta))$, which we rewrite as

$$p(x|\mathcal{D}) \approx \sum_n \int d\theta\, p(x|\theta) q_n(\theta) - (N-1) \int d\theta\, p(x|\theta) q(\theta) \ .$$

Each predictive density in the above equation simplifies as

$$\int d\theta\, p(x|\theta) q_n(\theta) = \frac{1}{Z_n} \sum_k \sum_l \begin{cases} \dfrac{\lambda_{k\backslash n}\, \lambda_{l\backslash n}}{(\sum_{k'} \lambda_{k'\backslash n} + 1) \sum_{k'} \lambda_{k'\backslash n}} p(x_n|k\backslash n)\, p(x|l\backslash n) & \text{if } k \neq l \\ \dfrac{(\lambda_{k\backslash n} + 1)\lambda_{k\backslash n}}{(\sum_{k'} \lambda_{k'\backslash n} + 1) \sum_{k'} \lambda_{k'\backslash n}} p(x_n|k\backslash n)\, p(x|x_n, k\backslash n) & \text{if } k = l. \end{cases}$$

We have seen how to compute $p(x_n|k\backslash n)$ in (26) and the discussion that followed it; we similarly define $p(x|l\backslash n)$. Density $p(x|x_n, k\backslash n)$ is again the Student-t distribution of (26), but now Λ is replaced with $\Lambda^*_{\backslash n}$ from (35). The correction evaluates in $\mathcal{O}(NK^2)$ complexity.

References

H. Attias. A variational Bayesian framework for graphical models. In T. Leen et al., editor, *Advances in Neural Information Processing Systems 12*. MIT Press, Cambridge, 2000.

M. J. Beal and Z. Ghahramani. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, 7:453–464, 2003.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

C. M. Bishop and M. Svensén. Bayesian hierarchical mixtures of experts. In U. Kjaerulff and C. Meek, editors, *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, pages 57–64. Morgan Kaufmann, 2003.

M. Chertkov and V. Y. Chernyak. Loop series for discrete statistical models on graphs. *Journal of Statistical Mechanics: Theory and Experiment*, 2006:P06009, 2006.

A. Corduneanu and C. M. Bishop. Variational Bayesian model selection for mixture distributions. In T. Richardson and T. Jaakkola, editors, *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, pages 27–34. Morgan Kaufmann, 2001.

J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society B*, 56(2):363–375, 1994.

A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius. Storage capacity and learning algorithms for two-layer neural networks. *Phys. Rev. A*, 45(10):7590–7609, May 1992.

V. Gómez, J. M. Mooij, and H. J. Kappen. Truncating the loop series expansion for belief propagation. *Journal of Machine Learning Research*, 8:1987–2016, 2007.

P. Gregory. *Bayesian Logical Data Analysis for the Physical Sciences*. Cambridge University Press, 2005.

T. Heskes and O. Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In A. Darwiche and N. Friedman, editors, *Proceedings UAI-2002*, pages 216–233, 2002.

Y. Iba. Extended ensemble Monte Carlo. *International Journal of Modern Physics C*, 12 (5):623–656, 2001.

S . Ikeda, T. Tanaka, and S. Amari. Information geometry of turbo and low-density parity-check codes. *IEEE Transactions on Information Theory*, 50(6):1097, 2004.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

D. A. Kofke. On the acceptance probability of replica-exchange Monte Carlo trials. *Journal of Chemical Physics*, 117(15):6911–6914, 2002.

D. J. C. MacKay. A problem with variational free energy minimization. Technical report, Department of Physics, University of Cambridge, 2001.

H. Matsui and T. Tanaka. Analysis on equilibrium point of expectation propagation using information geometry. In *International Conference on Neural Information Processing (ICONIP)*, 2008.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *UAI 2001*, pages 362–369, 2001a.

T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT Media Lab, 2001b.

T. P. Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, Cambridge, UK, 2005.

J. M. Mooij, B. Wemmenhove, H. J. Kappen, and T. Rizzo. Loop corrected belief propagation. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

M. Opper, U. Paquet, and O. Winther. Improving on expectation propagation. In *Neural Information Processing Systems*, 2008.

M. Opper and O. Winther. Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12:2655–2684, 2000.

M. Opper and O. Winther. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6:2177–2204, 2005.

C. E. Rasmussen and Z. Ghahramani. Occam's razor. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2001.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society (B)*, 59(4):731–792, 1997.

M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, 2003.

J. Skilling. Probabilistic data analysis: an introductory guide. *Journal of Microscopy*, 190 (1):28–36, 1998.

E. Sudderth, M. Wainwright, and A. Willsky. Loop series and Bethe variational bounds in attractive graphical models. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1425–1432. MIT Press, Cambridge, MA, 2008.