Improving on Expectation Propagation

Ulrich Paquet[1], *Ole Winther*[2], *Manfred Opper*[3]

[1] *University of Stellenbosch, South Africa;* [2] *Technical University of Denmark, Denmark;* [3] *Technical University Berlin, Germany.*

Introduction

Expectation Propagation (EP) is a popular deterministic algorithm for approximating Bayesian averages.

- *The accuracy of these approximations can be systematically improved through perturbation corrections.*

We present an EP treatment for the gaussian mixture model. For predictive density and marginal likelihood, we illustrate how EP gives better accuracy than Variational Bayes (VB), and how higher order corrections present further improvements. A state of the art MCMC method—parallel tempering and thermodynamic integration—provide the base line for comparison.

Expectation Propagation in a Nutshell. In this illustration we observe i.i.d. data $\mathcal{D} = \{x_n\}_{n=1}^N$ generated by $p(x_n|\theta)$, with an exponential family prior $p(\theta) \propto \exp(\Lambda_0^T \phi(\theta))h(\theta)$ defined by the statistics $\phi(\theta)$.

- Approximate $p(\theta|\mathcal{D}) = \frac{1}{Z}\prod_n p(x_n|\theta)p(\theta)$ by a tractable density

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

which shares the same moments (i.e. an *expectation-consistent* approximation) with all the densities $q_n(\theta)$,

$$\langle \phi(\theta) \rangle_q = \langle \phi(\theta) \rangle_{q_n}, \quad n = 1, \ldots, N\ , \tag{2}$$

$$q_n(\theta) = \frac{1}{Z(\Lambda_{\backslash n}, 1_n)} p(x_n|\theta) \exp(\Lambda_{\backslash n}^T \phi(\theta))\, p(\theta)\ , \tag{3}$$

where $\Lambda = \sum_n \Lambda_n$ and $\Lambda_{\backslash n} = \Lambda - \Lambda_n$. If 1_n is a unit-vector in the nth direction, then

$$Z(\Lambda, a) = \int d\theta \prod_n [p(x_n|\theta)]^{a_n} \exp(\Lambda^T \phi(\theta))\, p(\theta)\ . \tag{4}$$

- Λ_n's are optimised to achieve consistency for moments in (2).
- The approximation to the marginal likelihood is given by

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

Corrections to EP

The *exact* posterior and the marginal likelihood are expressed in terms of q_n's, q and the normalising partition functions: solving (3) for $p(x_n|\theta)$ and using the definitions of the densities, we get

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

- The exact posterior and the marginal likelihood can be written as

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

$$\text{where} \quad R = \int d\theta q(\theta) \prod_n (1 + \varepsilon_n(\theta)) \quad \text{and} \quad \varepsilon_n(\theta) = \frac{q_n(\theta) - q(\theta)}{q(\theta)}\ . \tag{8}$$

The densities $q(\theta)$ and $q_n(\theta)$ share a set of generalised moments; we hope that they are close enough such that $\varepsilon_n(\theta)$ can be treated (in an average sense) as small.

- An expansion of the posterior and Z in terms of $\varepsilon_n(\theta)$ truncated at low orders might give the dominant corrections to EP:

$$R = 1 + \sum_{n_1<n_2} \langle \varepsilon_{n_1}(\theta)\varepsilon_{n_2}(\theta) \rangle_q + \sum_{n_1<n_2<n_3} \langle \varepsilon_{n_1}(\theta)\varepsilon_{n_2}(\theta)\varepsilon_{n_3}(\theta) \rangle_q + \ldots\ , \tag{9}$$

where the first order term $\sum_n \langle \varepsilon_n(\theta) \rangle_q = 0$ vanishes by the normalization of q_n and q.

Similarly, the predictive distribution $p(x|\mathcal{D}) = \int d\theta\, p(x|\theta)$ is

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

Results and illustrations

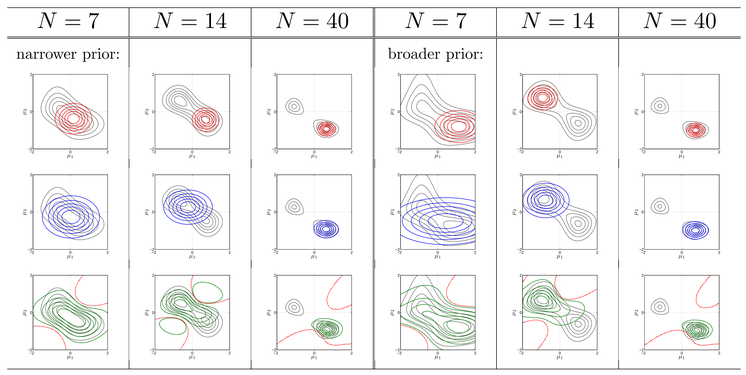
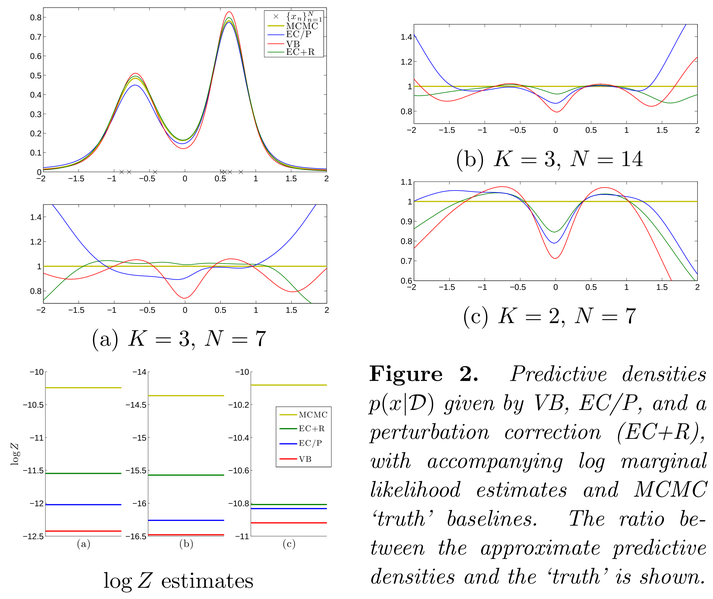


Figure 1. *A comparison between the VB (top row) and EC/P (middle) $q(\theta)$, and a first-order EC correction $p(\theta|\mathcal{D}) \approx \sum_n q_n(\theta) - (N-1)q(\theta)$ (bottom). Data is assumed to come from a two-component mixture $p(x_n|\theta) = 0.4\mathcal{N}(x_n|\mu_1, 1) + 0.6\mathcal{N}(x_n|\mu_2, 1)$, with only the means $\theta = \{\mu_1, \mu_2\}$ unknown. 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.*

- In the figures presented below we assume that the likelihood for data point x_n is a mixture of K d-dimensional gaussians, $p(x_n|\theta) = \sum_{k=1}^K p(k)p(x_n|k) = \sum_{k=1}^K \pi_k \mathcal{N}(x_n|\mu_k, \Gamma_k^{-1})$, with Dirichlet and Normal-Wishart priors on π and $\{\mu_k, \Gamma_k\}_{k=1}^K$.
- The approximating density $q(\theta) = q(\pi)\prod_k q(\mu_k, \Gamma_k)$ is a factorized product of a Dirichlet and Normal-Wishart distributions.



(a) $K = 3$, $N = 7$ (b) $K = 3$, $N = 14$ (c) $K = 2$, $N = 7$

log Z estimates

Figure 2. *Predictive densities $p(x|\mathcal{D})$ given by VB, EC/P, and a perturbation correction (EC+R), with accompanying log marginal likelihood estimates and MCMC 'truth' baselines. The ratio between the approximate predictive densities and the 'truth' is shown.*

- Under posteriors with many symmetries general arguments suggest that we can correct the marginal likelihood estimate by a factor of $K!$ for large N; it is unclear what this behaviour will be for smaller N.

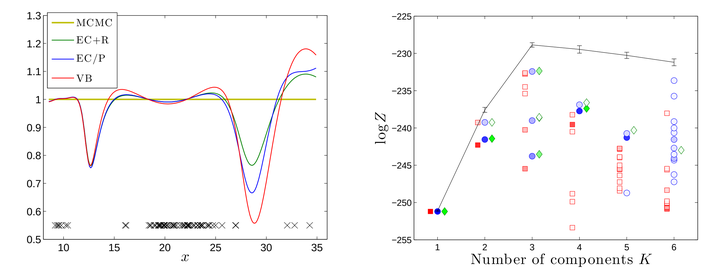


Figure 3. *Left: The ratio between the predictive densities given by VB, EC/P, and EC+R, and the 'truth' $p(x|\mathcal{D})$ for the* **galaxy** *data set, with $K = 3$. Right: $\log Z$ estimates for choices of K, given for VB (squares), EC/P (circles), and EC+R (diamonds), with an MCMC baseline.*