Electronic Journal of Statistics
Vol. 10 (2016) 550–581
ISSN: 1935-7524
DOI: 10.1214/16-EJS1114

The explicit form of expectation propagation for a simple statistical model

Andy S. I. Kim and M. P. Wand

University of Technology Sydney and
Australian Research Council Centre of Excellence
for Mathematical and Statistical Frontiers,
School of Mathematical and Physical Sciences,
Broadway 2007, Australia
e-mail: sangil.kim@student.uts.edu.au; matt.wand@uts.edu.au
url: matt-wand.utsacademics.info

Abstract: We derive the explicit form of expectation propagation for approximate deterministic Bayesian inference in a simple statistical model. The model corresponds to a random sample from the Normal distribution. The explicit forms, and their derivation, allow a deeper understanding of the issues and challenges involved in practical implementation of expectation propagation for statistical analyses. No auxiliary approximations are used: we follow the expectation propagation prescription exactly. A simulation study shows expectation propagation to be more accurate than mean field variational Bayes for larger sample sizes, but at the cost of considerably more algebraic and computational effort.

MSC 2010 subject classifications: Primary62F15; secondary 62H12.
Keywords and phrases: Bayesian computing, factor graph, hierarchical Bayesian models, message passing algorithm, quadrature, variational message passing.

Received December 2014.

Contents

1. Introduction

Minka [8] and Minka and Winn [9] describe a general prescription for approximate inference in hierarchical Bayesian models known as *expectation propagation*, building on earlier work on topics such as *assumed density filtering* (e.g. Maybeck [6]) and *loopy belief propagation* (e.g. Frey and MacKay [3]). Expectation propagation is used to achieve fast deterministic inference in the Infer.NET software platform (Minka *et al.* [10]). Infer.NET also supports mean field variational Bayes (e.g. Wainwright and Jordan [14]) using the *variational message passing* formulation (Winn and Bishop [18]), for achieving similar aims. A small number of numerical studies (e.g. Minka [7]) have shown that expectation propagation is often more accurate than mean field variational Bayes.

Despite these developments, expectation propagation is virtually unknown in mainstream Statistics. Prescriptions such as those given in Minka [8] and Minka and Winn [9] use concepts such as factor graphs, message passing and Kullback-Leibler projection; which are unfamiliar to most statisticians. Our main contribution in this article is to obtain the explicit form of expectation propagation for a specific statistical model. By "explicit" we mean that a programmer could readily implement an expectation propagation fitting and inference algorithm based on the formulae given in Sections 2 and 4. We also avoid use of any auxiliary approximations – following Minka [8] and Minka and Winn [9] exactly. With succinctness in mind, we choose a particularly simple statistical scenario: Bayesian inference based on Normal random sample. Despite its simplicity, 12 pages of algebra, given in Appendix A, are required to derive the explicit forms from Minka [8]. Our contributions allow statistical analysts to see exactly what is involved in deriving and implementing expectation propagation. Zoeter and Heskes [19] provided details on expectation propagation for a stochastic volatility model. However, they used a least squares approximation to (Inverse) Gamma density projection, whereas we do this projection exactly via our Result 2.

As mentioned above, a prescription for expectation propagation for general Bayesian models is given in Minka [8]. This prescription involves: (1) specifying a product density form for approximating the joint posterior density function of the model parameters, latent and auxiliary variables, (2) forming the *factor graph* based on the model and the product density form, (3) computing *messages* for *passing* between the factors and nodes of the factor graph. The boxed algorithm at the end of Section 6 of Minka [8], together with his equations (54) and (83), is the formulation of expectation propagation that we use in this article.

In Section 2 we provide some preliminary definitions and results. Section 3 contains a summary of expectation propagation for general models. The centerpiece of the article is Section 4 in which we give the explicit form of expectation propagation for a Normal random sample model. In Section 5 we perform some comparisons with mean field variational Bayes approximate inference for the same model. A simulation study shows expectation propagation to usually be the more accurate of the two, although this has to be traded off against a much larger algebraic and computational overhead. All derivations are in Appendix A.

2. Preliminary definitions and results

The explicit form of expectation propagation for (4.2) under (4.4) depends on several definitions and results, which we lay out in this section.

2.1. Non-analytic function definitions

The following integral-defined functions are required:

$$\begin{aligned}\mathcal{A}(p,q,r,s,t,u) &\equiv \int_{-\infty}^{\infty} \frac{x^p \exp(qx - rx^2)\,dx}{(x^2+sx+t)^u}, \\ &\qquad p \geq 0,\ q \in \mathbb{R}, r > 0, s \in \mathbb{R}, t > \tfrac{1}{4}s^2, u > 0 \\ \text{and} \\ \mathcal{B}(p,q,r,s,t,u) &\equiv \int_{-\infty}^{\infty} \frac{x^p \exp\{qx - re^x - se^x/(t+e^x)\}\,dx}{(t+e^x)^u}, \\ &\qquad p \geq 0,\ q \in \mathbb{R}, r > 0, s \geq 0, t > 0, u > 0.\end{aligned} \tag{2.1}$$

Appendix B of Wand *et al.* [17] describes stable and efficient computation of functions of this type via quadrature. To avoid overflow and underflow, it is important to work with $\log|\mathcal{A}(p,q,r,s,t,u)|$ and $\text{sign}(\mathcal{A}(p,q,r,s,t,u))$ rather than $\mathcal{A}(p,q,r,s,t,u)$ itself. The same applies to computations involving the function $\mathcal{B}(p,q,r,s,t,u)$.

The only other non-analytic function required for algorithm specification is

$$(\log - \text{digamma})^{-1}(x), \quad x > 0, \tag{2.2}$$

the inverse of the function

$$(\log - \text{digamma})(x) \equiv \log(x) - \text{digamma}(x), \quad x > 0$$

with $\text{digamma}(x) \equiv \frac{d}{dx}\log\{\Gamma(x)\}$. This raises the question of existence and uniqueness of (2.2) over $\mathbb{R}^+ \equiv \{x \in \mathbb{R} : x > 0\}$. The following theorem shows that (2.2) is well-defined:

Theorem 1. *The function* $\log - \text{digamma}$ *is a bijective mapping from* $\mathbb{R}^+$ *onto* $\mathbb{R}^+$.

A proof of Theorem 1 is given in Appendix A.1.

The monotonicity and smoothness of the function $\log - \text{digamma}$ means that $(\log - \text{digamma})^{-1}$ can be computed rapidly via Newton-Raphson iteration. Good starting values can be obtained from

$$1/(2x) < (\log - \text{digamma})(x) < 1/x, \quad x > 0.$$

However it is worth noting that computation of $(\log - \text{digamma})(x)$ via subtraction can have round-off error problems for very large x, and lead to zero being returned erroneously. The R language (R Development Core Team [11]) function `logmdigamma()`, in the package `statmod` (Smyth [12]), overcomes this problem and accurately computes $(\log - \text{digamma})(x)$ for an input $x > 0$. In the simulations described in Section 5 we work with `logmdigamma()` in our computation of $(\log - \text{digamma})^{-1}$.

2.2. Distributional definitions and natural parametrization

The model specification and inference algorithms can be done in terms of two distributional families, the Normal distribution and the Inverse Gamma distribution.

The Normal distribution with mean μ and variance $\sigma^2 > 0$, denoted by $N(\mu, \sigma^2)$, has corresponding density function

$$p(x) = (2\pi\sigma^2)^{-1/2} \exp\{-(x-\mu)^2/(2\sigma^2)\}. \tag{2.3}$$

The Inverse Gamma distribution with shape parameter $\kappa > 0$ and rate parameter $\lambda > 0$, denoted by $\text{IG}(\kappa, \lambda)$, has corresponding density function

$$p(x) = \{\lambda^\kappa/\Gamma(\kappa)\}\, x^{-\kappa-1} \exp(-\lambda/x), \quad x > 0. \tag{2.4}$$

Each of (2.3) and (2.4) are *exponential family* density functions, since they can be written in the form

$$p(x) = \exp\{\boldsymbol{T}(x)^T \boldsymbol{\eta} - A(\boldsymbol{\eta}) + B(x)\}$$

where $\boldsymbol{T}(x)$ is the *natural statistic* and $\boldsymbol{\eta}$ is the *natural parameter*. For (2.3) we have

$$\boldsymbol{T}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}, \quad \boldsymbol{\eta} \equiv \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{bmatrix}, \quad A(\boldsymbol{\eta}) = -\frac{\eta_1^2}{4\eta_2} - \tfrac{1}{2}\log(-2\eta_2)$$

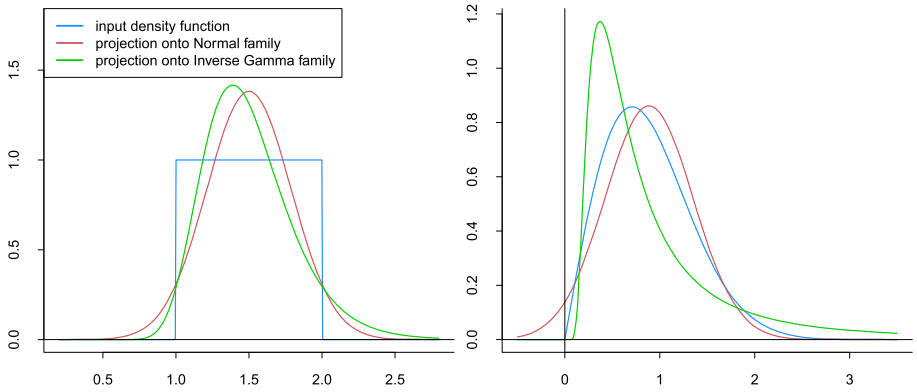

FIG 1. *Kullback Leibler projections of* $p(x) = 1,\ 1 < x < 2$, *(left panel) and* $p(x) = 2x\exp(-x^2),\ x > 0$, *(right panel) onto the Normal and Inverse Gamma families.*

and $B(x) = -\frac{1}{2}\log(2\pi)$. For (2.4) we have

$$\boldsymbol{T}(x) = \begin{bmatrix} \log(x) \\ 1/x \end{bmatrix},\ \boldsymbol{\eta} \equiv \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} -\kappa - 1 \\ -\lambda \end{bmatrix},$$
$$A(\boldsymbol{\eta}) = \log\Gamma(-\eta_1 - 1) + (\eta_1 + 1)\log(-\eta_2)$$

and $B(x) = 0$.

The inverse mappings from the natural parameters to the common parameters are

$$\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \equiv \begin{bmatrix} -\eta_1/(2\eta_2) \\ -1/(2\eta_2) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \kappa \\ \lambda \end{bmatrix} \equiv \begin{bmatrix} -\eta_1 - 1 \\ -\eta_2 \end{bmatrix}.$$

2.3. Kullback-Leibler divergence and projection

For arbitrary density functions p_1 and p_2 on $\mathbb{R}^d$,

$$\mathrm{KL}(p_1 \,\|\, p_2) \equiv \int_{\mathbb{R}^d} p_1(\boldsymbol{x}) \log\left\{p_1(\boldsymbol{x})/p_2(\boldsymbol{x})\right\} d\boldsymbol{x}$$

denotes the *Kullback-Leibler divergence* of p_2 from p_1. Note that $\mathrm{KL}(p_1 \,\|\, p_2) \geq 0$ for any p_1 and p_2 and that, in general, $\mathrm{KL}(p_1 \,\|\, p_2) \neq \mathrm{KL}(p_2 \,\|\, p_1)$.

Intrinsic to expectation propagation is the notion of *Kullback-Leibler projection*. Let $\mathcal{Q}$ be a family of univariate density functions. Then the Kullback-Leibler projection of the univariate density function p onto $\mathcal{Q}$ is given by

$$\mathrm{proj}[\,p\,] \equiv \operatorname*{argmin}_{q \in \mathcal{Q}} \mathrm{KL}(p \,\|\, q) \tag{2.5}$$

In the case where $\mathcal{Q}$ is an exponential family of density functions (2.5) simplifies to a convenient moment-matching problem. Suppose that

$$\mathcal{Q} = \{q : q(x) = \exp\{\boldsymbol{T}(x)^T\boldsymbol{\eta} - A(\boldsymbol{\eta}) + B(x)\}, \quad \boldsymbol{\eta} \in \boldsymbol{H}\}$$

where $\boldsymbol{H}$ is the space of allowable values of $\boldsymbol{\eta}$. Then substitution into (2.5) leads to

$$\begin{aligned} \text{proj}[\,p\,] &= \exp\{\boldsymbol{T}(x)^T\boldsymbol{\eta}^* - A(\boldsymbol{\eta}^*) + B(x)\} \\ \text{where } \boldsymbol{\eta}^* &\equiv \underset{\boldsymbol{\eta}\in\boldsymbol{H}}{\text{argmin}} \left\{ A(\boldsymbol{\eta}) - \boldsymbol{\eta}^T \int_{-\infty}^{\infty} \boldsymbol{T}(x)\, p(x)\, dx \right\}. \end{aligned}$$

However, an exponential family result that equates the derivative vector of $A(\boldsymbol{\eta})$ with the expectation of the natural statistic leads to $\boldsymbol{\eta}^*$ being the solution to

$$\int_{-\infty}^{\infty} \boldsymbol{T}(x)\, p(x)\, dx = \int_{-\infty}^{\infty} \boldsymbol{T}(x)\, \exp\{\boldsymbol{T}(x)^T\boldsymbol{\eta} - A(\boldsymbol{\eta}) + B(x)\}\, dx. \tag{2.6}$$

In other words, $\boldsymbol{\eta}^*$ is chosen so that p and proj$[\,p\,]$ have the same natural statistic moments. With relatively little algebra we then obtain:

Result 1. *Let x be non-degenerate random variable for which $E(x^2)$ exists and with density function p. The Kullback-Leibler projection of p onto the Normal family is the $N(\mu^*, (\sigma^*)^2)$ density function where*

$$\mu^* = E(x) \quad \textit{and} \quad (\sigma^2)^* = E(x^2) - (\mu^*)^2.$$

Result 2. *Let x be a positive-valued non-degenerate random variable for which $E(1/x)$ and $E\{\log(x)\}$ exist and with density function p. The Kullback-Leibler projection of p onto the Inverse Gamma family is the $IG(\kappa^*, \lambda^*)$ density function where*

$$\kappa^* = (\log - \text{digamma})^{-1}\Big(\log\, E(1/x) + E\{\log(x)\}\Big) \quad \textit{and} \quad \lambda^* = \kappa^*/E(1/x).$$

Figure 1 provides illustration of Kullback-Leibler projection onto the Normal and Inverse Gamma families. The left-hand panel shows the projections of $p(x) = 1,\ 1 < x < 2$, the density function of the Uniform distribution on $(1, 2)$. The input function for the right-hand panel is $p(x) = 2x\exp(-x^2),\ x > 0$, the Weibull density function with shape parameter 2.

3. Expectation propagation in general

We first describe expectation propagation for general Bayesian statistical models with observed data $\boldsymbol{D}$ and parameter vector $\boldsymbol{\theta}$. Consider approximations to the joint posterior density function $p(\boldsymbol{\theta}|\boldsymbol{D})$ that have generic form

$$p(\boldsymbol{\theta}|\boldsymbol{D}) \approx \prod_{i=1}^{M} q^*(\boldsymbol{\theta}_i)$$

where

$$\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \tag{3.1}$$

is some partition of $\boldsymbol{\theta}$ and the $q^*(\boldsymbol{\theta}_i)$s are chosen to optimize a particular functional closeness criterion. For example, taking the $q^*(\boldsymbol{\theta}_i)$s to minimize the Kullback-Leibler divergence of $p(\boldsymbol{\theta}|\boldsymbol{D})$ from a product density over the elements of (3.1),

$$\mathrm{KL}\left(\prod_{i=1}^{M} q(\boldsymbol{\theta}_i) \,\middle\|\, p(\boldsymbol{\theta}|\boldsymbol{D})\right),$$

corresponds to mean field variational Bayes and the q^*s can be obtained using a convex optimization scheme (e.g. Section 10.1.1, Bishop [2]). Variational message passing solves the mean field variational Bayes optimization problem via iteratively updating messages on a factor graph as described in Minka [8] and Minka and Winn [9]. Expectation propagation is driven by the reverse Kullback-Leibler divergence

$$\mathrm{KL}\left(p(\boldsymbol{\theta}|\boldsymbol{D}) \,\middle\|\, \prod_{i=1}^{M} q(\boldsymbol{\theta}_i)\right). \tag{3.2}$$

The challenges of minimizing (3.2) are discussed in Section 6 of Minka [8]. The approximate inference method known as *belief propagation* (Frey and MacKay [3]) is based on (3.2) but leads to very complex approximating density functions.

Expectation propagation overcomes the complexity problem of belief propagation via Kullback-Leibler projection onto exponential density functions. Minka [8] develops a strategy for approximate minimization of (3.2) for general $p(\boldsymbol{\theta}|\boldsymbol{D})$ and $\prod_{i=1}^{M} q(\boldsymbol{\theta}_i)$ in terms of *messages* passed on an appropriate factor graph. We now provide details. A convenient notation for subsets S of $\{1,\ldots,M\}$ is

$$\boldsymbol{\theta}_S \equiv \{\boldsymbol{\theta}_i : i \in S\}.$$

Given the partition (3.1), the joint density function of $\boldsymbol{\theta}$ and $\boldsymbol{D}$ is expressible as

$$p(\boldsymbol{\theta},\boldsymbol{D}) = \prod_{j=1}^{N} f_j\left(\boldsymbol{\theta}_{S_j}\right) \quad \begin{array}{l}\text{for subsets } S_j \text{ of } \{1,\ldots,M\} \\ \text{and } \textit{factors } f_j,\ 1 \le j \le N.\end{array} \tag{3.3}$$

For example, if $p(\boldsymbol{\theta},\boldsymbol{D})$ is based on a directed acyclic graphical model with nodes $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M$ and $\boldsymbol{D}$ then

$$p(\boldsymbol{\theta},\boldsymbol{D}) = \left\{\prod_{i=1}^{M} p(\boldsymbol{\theta}_i|\text{parents of } \boldsymbol{\theta}_i)\right\} p(\boldsymbol{D}|\,\text{parents of } \boldsymbol{D}) \tag{3.4}$$

is an $N = M+1$ example of (3.3) with f_j, $1 \le j \le M$, corresponding to density function of θ_j conditional on its parents and f_{M+1} corresponding to the likelihood. Each factor is a function of the subset of (3.1) corresponding to parental relationships in the directed acyclic graph. Further factorization of (3.4) may be possible.

The factor graph in Figure 2 shows an $M = 9$, $N = 11$ example of (3.3). The edges link each factor to the stochastic nodes on which the factor depends.

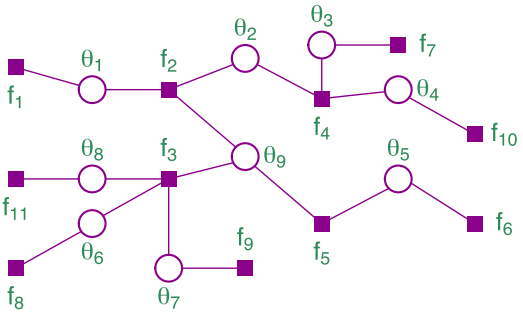

FIG 2. *A factor graph corresponding to a Bayesian model with stochastic nodes* $\theta_1, \ldots, \theta_9$ *and factors* $f_1, \ldots, f_{11}$.

The following notation is useful for describing the Minka [8] and Minka and Winn [9] expectation propagation algorithm:

$$\text{neighbors}(j) = \{1 \le i \le M : \boldsymbol{\theta}_i \text{ is a neighbor of } f_j\}$$

Examples of this notation for the Figure 2 factor graph are

$$\text{neighbors}(1) = \{1\}, \quad \text{neighbors}(2) = \{1, 2, 9\} \quad \text{and} \quad \text{neighbors}(3) = \{6, 7, 8, 9\}.$$

According to this notation, $p(\boldsymbol{\theta}, \boldsymbol{D}) = \prod_{j=1}^{N} f_j(\theta_{\text{neighbors}(j)})$. For each $1 \le i \le M$ and $1 \le j \le N$, the expectation propagation stochastic node to factor message updates are

$$m_{\boldsymbol{\theta}_i \to f_j}(\boldsymbol{\theta}_i) \longleftarrow \prod_{j' \ne j:\, i \in \text{neighbors}(j')} m_{f_{j'} \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \tag{3.5}$$

and the factor to stochastic node updates are

$$m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \longleftarrow \frac{\text{proj}\left[\begin{array}{l} Z^{-1} m_{\boldsymbol{\theta}_i \to f_j}(\boldsymbol{\theta}_i) \\ \times \displaystyle\int f_j(\boldsymbol{\theta}_{\text{neighbors}(j)}) \prod_{i' \in \text{neighbors}(j) \setminus \{i\}} m_{\boldsymbol{\theta}_{i'} \to f_j}(\boldsymbol{\theta}_{i'})\, d\boldsymbol{\theta}_{\text{neighbors}(j) \setminus \{i\}} \end{array}\right]}{m_{\boldsymbol{\theta}_i \to f_j}(\boldsymbol{\theta}_i)} \tag{3.6}$$

where Z is the normalizing factor that ensures that the function of $\boldsymbol{\theta}_i$ inside the proj[·] is a density function. The normalizing factor in (3.6) involves summation if some of the $\theta_{i'}$ have discrete components. The proj[·] in (3.6) denotes Kullback-Leibler projection onto an appropriate exponential family of density functions. The appropriate family is driven by conjugacy constraints. If $\text{neighbors}(j) = \{i\}$ then (3.6) reduces to

$$m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \longleftarrow \frac{\text{proj}\left[m_{\boldsymbol{\theta}_i \to f_j}(\boldsymbol{\theta}_i) f_j(\theta_i)/Z\right]}{m_{\boldsymbol{\theta}_i \to f_j}(\boldsymbol{\theta}_i)}. \tag{3.7}$$

If $f_j(\boldsymbol{\theta}_i)$ is proportional to an exponential density function and $m_{\boldsymbol{\theta}_i \to f_j}(\boldsymbol{\theta}_i)$ is initialized to be in the same family as $f_j(\boldsymbol{\theta}_i)$ then (3.7) becomes $m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) = f_j(\theta_i)$. As the stochastic node to factor messages get updated using (3.5) similar conjugacy contraints drive the choice of the family for the proj[·] operator for other $m_{f_j \to \boldsymbol{\theta}_i}$ updates. Upon convergence of the messages, the Kullback-Leibler optimal q-densities are obtained via

$$q^*(\boldsymbol{\theta}_i) \propto \prod_{j: i \in \text{neighbors}(j)} m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i). \tag{3.8}$$

A reasonable stopping criterion is the approximate marginal log-likelihood having a negligible relative change. An approximate marginal log-likelihood expression for general factor graphs is given in Appendix B of Minka and Winn [9], with justification from the arguments of Section 4.4 of Minka [8]. In terms of the notation of this section, the expression is:

$$\log\{\underset{\sim}{p}(\boldsymbol{D}; q)\} = \sum_{i=1}^{M} \log s_{\theta_i} + \sum_{j=1}^{N} \log s_{f_j} \tag{3.9}$$

where

$$s_{\theta_i} \equiv \int \prod_{j: i \in \text{neighbors}(j)} m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)\, d\theta_i \tag{3.10}$$

and

$$s_{f_j} \equiv \frac{\displaystyle\int f_j(\boldsymbol{\theta}_{\text{neighbors}(j)}) \prod_{i \in \text{neighbors}(j)} m_{\boldsymbol{\theta}_i \to f_j}(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_{\text{neighbors}(j)}}{\displaystyle\int \prod_{i \in \text{neighbors}(j)} m_{\boldsymbol{\theta}_i \to f_j}(\boldsymbol{\theta}_i)\, m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_{\text{neighbors}(j)}}. \tag{3.11}$$

This stopping criterion is not necessarily monotone, nor is convergence guaranteed (e.g. Bishop [2], Section 10.7). However, employment of damping strategies often leads to successful convergence (e.g. Minka [8]).

4. Expectation propagation for a normal random sample model

The general form of expectation propagation as described in the previous section is rather abstract. The actual computational steps are difficult to glean from expressions such as (3.5) and (3.6). We now focus on a specific simple Bayesian statistical model and make the updates as concrete as possible. Nevertheless, the updates are still complicated and require several pages of derivation which we provide in Appendix A.

We consider the following Bayesian Normal random sample model:

$$\begin{gathered} x_i \,|\, \mu, \sigma^2 \;\; \text{independently distributed} \;\; N(\mu, \sigma^2), \;\; 1 \le i \le n, \\ \mu \sim N(\mu_\mu, \sigma_\mu^2), \quad \sigma \sim \text{Half-Cauchy}(A). \end{gathered} \tag{4.1}$$

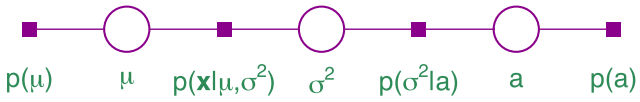

FIG 3. *Factor graph corresponding to (4.2) with stochastic nodes according to product restriction (4.4).*

where $\mu_\mu \in \mathbb{R}$, $\sigma_\mu > 0$ and $A > 0$ are user-specified hyperparameters. The Half-Cauchy(A) prior on σ corresponds to its density function being $p(\sigma) = \{2/(\pi A)\}/\{1 + (\sigma/A)^2\}$, $\sigma > 0$. However,

$$\sigma \sim \text{Half-Cauchy}(A) \quad \text{is equivalent to} \quad \sigma^2|a \sim \text{IG}(\tfrac{1}{2}, 1/a), \quad a \sim \text{IG}(\tfrac{1}{2}, 1/A^2)$$

where $\text{IG}(\kappa, \lambda)$ denotes the Inverse Gamma distribution with shape parameter $\kappa > 0$ and rate parameter $\lambda > 0$, with full definition given in Section 2.2. Therefore, an equivalent model to (4.1) is:

$$\begin{aligned} x_i \,|\, \mu, \sigma^2 \;\; &\text{independently distributed} \;\; N(\mu, \sigma^2), \;\; 1 \le i \le n, \\ \mu \sim N(\mu_\mu, \sigma_\mu^2), \quad &\sigma^2 |\, a \sim \text{IG}(\tfrac{1}{2}, 1/a), \quad a \sim \text{IG}(\tfrac{1}{2}, 1/A^2). \end{aligned} \tag{4.2}$$

Model (4.2) better lends itself to expectation propagation-based inference because all of the messages are in the Normal and Inverse Gamma families, and we work with it from now onwards.

The joint density function of the observed data vector $\boldsymbol{x} \equiv (x_1, \ldots, x_n)$ and stochastic variables in (4.2) is

$$p(\boldsymbol{x}, \mu, \sigma^2, a) = p(\boldsymbol{x} |\, \mu, \sigma^2)\, p(\sigma^2 |\, a)\, p(\mu)\, p(a) \tag{4.3}$$

where, for example,

$$p(\boldsymbol{x}|\mu, \sigma^2) \equiv (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}.$$

We will derive the expectation propagation approximation to the joint posterior density function $p(\mu, \sigma^2, a \,|\, \boldsymbol{x})$, denoted by $q(\mu, \sigma^2, a)$, under the following product density restriction:

$$q(\mu, \sigma^2, a) = q(\mu)\, q(\sigma^2)\, q(a). \tag{4.4}$$

The relevant factor graph is shown in Figure 3. Notice that there is a circular node corresponding to each q-density factor on the right-hand side of (4.4). The solid square nodes correspond to the factors in (4.3) An edge connects each factor with the stochastic nodes that are included in that factor. Figure 3 is crucial to expectation propagation approximate inference for (4.2) as described in Minka [8].

With the function definitions of Appendix A.4 in place, the expectation propagation iteration algorithm boils down to updating the natural parameter vectors of messages between neighboring nodes on the factor graph in Figure 3. For example, the message from $p(x|\mu,\sigma^2)$ to μ is of the form

$$m_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu}(\mu) \propto \exp\left\{\begin{bmatrix}\mu\\ \mu^2\end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu}\right\}$$

for some 2×1 vector $\boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu}$. Details are given in Appendix A.5.3. We initialize $m_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu}(\mu)$ to be the $N(0,1)$ density function in μ, corresponding to $\boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu}$ being set to $[0\ \ -\frac{1}{2}]^T$. The Inverse Gamma messages are initialized at the IG$(1,1)$ density, which corresponds to their natural parameter vectors being set to $[-2\ \ -1]^T$.

The stopping criterion involves the approximate marginal log-likelihood, denoted here by $\log\{\underset{\sim}{p}(\boldsymbol{x};q)\}$. In Appendix A.5.9 we derive an expression for $\log\{\underset{\sim}{p}(\boldsymbol{x};q)\}$ in terms of the non-analytic functions given in Section 2.1.

The expectation propagation approximations to $p(\mu|\boldsymbol{x})$ and $p(\sigma^2|\boldsymbol{x})$ are, respectively, $q^*(\mu)$ and $q^*(\sigma^2)$ where

$$q^*(\mu) \text{ is the } N\Big(-\eta_{q(\mu),1}/(2\eta_{q(\mu),2}), -1/(2\,\eta_{q(\mu),2})\Big) \text{ density function in } \mu$$

and

$$q^*(\sigma^2) \text{ is the IG}\Big(-\eta_{q(\sigma^2),1}-1, -\eta_{q(\sigma^2),2}\Big) \text{ density function in } \sigma^2.$$

The boxed algorithm in Section 6 of Minka [8] also accommodates the possibility of applying a damping step-size $0 \le \varepsilon < 1$, for the factor to stochastic node messages. We did not find this to be necessary for the simple model at hand and set $\varepsilon = 0$ here, with convergence always achieved in our evaluation studies (Section 5).

Code for Algorithm 1 in the R language [11] is available on the web-site where this article resides.

5. Evaluation of accuracy

We conducted a simulation study to evaluate the inferential accuracy of Algorithm 1, as well as its relative accuracy compared with mean field variational Bayes. The accuracy of $q^*(\mu)$ is quantified via

$$\text{accuracy}\{q^*(\mu)\} \equiv 100\left(1-\tfrac{1}{2}\int_{-\infty}^{\infty}|\,q^*(\mu)-p(\mu\,|\,\boldsymbol{x})|\,d\mu\right)\%. \tag{5.1}$$

An analogous definition applies to accuracy$\{q^*(\sigma^2)\}$. This accuracy measurement has the advantage of being transformation invariant and ranging over 0%–100%.

Algorithm 1 *Expectation propagation algorithm for determining the natural parameter vectors* $\boldsymbol{\eta}_{q(\mu)}$, $\boldsymbol{\eta}_{q(\sigma^2)}$ *and* $\boldsymbol{\eta}_{q(a)}$ *of the optimal density functions* $q^*(\mu)$, $q^*(\sigma^2)$ *and* $q^*(a)$ *for approximate Bayesian inference in the Normal random sample model (4.2).*

Inputs: $x_1, \ldots, x_n$; $\mu_\mu \in \mathbb{R}$, $\sigma_\mu^2 > 0$, $A > 0$.

Obtain $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n x_i^2$.

Initialize:

$$\boldsymbol{\eta}_{p(\mu)\to\mu} \longleftarrow \begin{bmatrix} \frac{\mu_\mu}{\sigma_\mu^2} \\ -\frac{1}{2\sigma_\mu^2} \end{bmatrix} ; \boldsymbol{\eta}_{p(a)\to a} \longleftarrow \begin{bmatrix} -\frac{3}{2} \\ -\frac{1}{A^2} \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu} \longleftarrow \begin{bmatrix} 0 \\ -\frac{1}{2} \end{bmatrix} ; \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\sigma^2} \longleftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\sigma^2|\,a)\to\sigma^2} \longleftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix} ; \boldsymbol{\eta}_{p(\sigma^2|\,a)\to a} \longleftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

Cycle:

$$\boldsymbol{\eta}_{\mu\to p(\mu)} \longleftarrow \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu}$$
$$\boldsymbol{\eta}_{\mu\to p(\boldsymbol{x}|\mu,\sigma^2)} \longleftarrow \boldsymbol{\eta}_{p(\mu)\to\mu}$$
$$\boldsymbol{\eta}_{\sigma^2\to p(\boldsymbol{x}|\mu,\sigma^2)} \longleftarrow \boldsymbol{\eta}_{p(\sigma^2|\,a)\to\sigma^2}$$
$$\boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu} \longleftarrow G^N\left(\boldsymbol{\eta}_{\mu\to p(\boldsymbol{x}|\mu,\sigma^2)}, \boldsymbol{\eta}_{\sigma^2\to p(\boldsymbol{x}|\mu,\sigma^2)}; \begin{bmatrix} n \\ \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^2 \end{bmatrix}\right)$$
$$\boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\sigma^2} \longleftarrow G^{IG1}\left(\boldsymbol{\eta}_{\sigma^2\to p(\boldsymbol{x}|\mu,\sigma^2)}, \boldsymbol{\eta}_{\mu\to p(\boldsymbol{x}|\mu,\sigma^2)}; \begin{bmatrix} n \\ \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^2 \end{bmatrix}\right)$$
$$\boldsymbol{\eta}_{\sigma^2\to p(\sigma^2|\,a)} \longleftarrow \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\sigma^2}$$
$$\boldsymbol{\eta}_{a\to p(\sigma^2|\,a)} \longleftarrow \boldsymbol{\eta}_{p(a)\to a}$$
$$\boldsymbol{\eta}_{p(\sigma^2|\,a)\to\sigma^2} \longleftarrow G^{IG2}\left(\boldsymbol{\eta}_{\sigma^2\to p(\sigma^2|\,a)}, \boldsymbol{\eta}_{a\to p(\sigma^2|\,a)}; 3\right)$$
$$\boldsymbol{\eta}_{p(\sigma^2|\,a)\to a} \longleftarrow G^{IG2}\left(\boldsymbol{\eta}_{a\to p(\sigma^2|\,a)}, \boldsymbol{\eta}_{\sigma^2\to p(\sigma^2|\,a)}; 1\right)$$
$$\boldsymbol{\eta}_{a\to p(a)} \longleftarrow \boldsymbol{\eta}_{p(\sigma^2|\,a)\to a}$$

until the relative change in $\log \underset{\sim}{p}(\boldsymbol{x}; q)$ is negligible.

$$\boldsymbol{\eta}_{q(\mu)} \longleftarrow \boldsymbol{\eta}_{p(\mu)\to\mu} + \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu}$$
$$\boldsymbol{\eta}_{q(\sigma^2)} \longleftarrow \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\sigma^2} + \boldsymbol{\eta}_{p(\sigma^2|\,a)\to\sigma^2}$$
$$\boldsymbol{\eta}_{q(a)} \longleftarrow \boldsymbol{\eta}_{p(\sigma^2|\,a)\to a} + \boldsymbol{\eta}_{p(a)\to a}$$

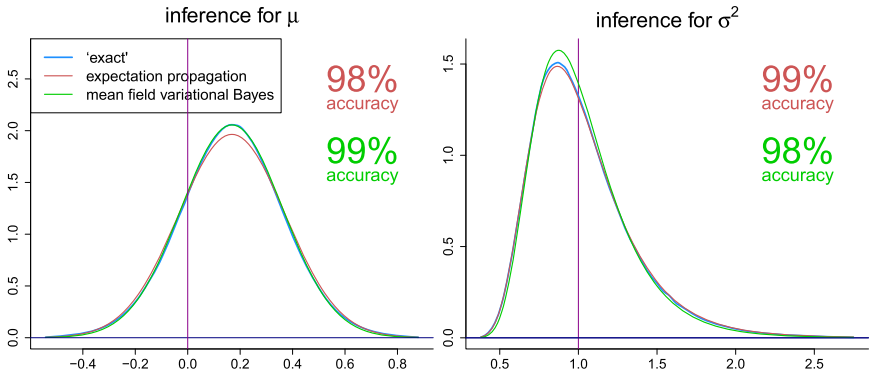

FIG 4. *Comparison among the expectation propagation approximate posterior density functions, mean field variational Bayes approximate density functions and the 'exact' Markov chain Monte Carlo-based posterior density function. The data correspond to the first* $n = 25$ *replication from the Normal random sample simulation study described in the text. The left panel is concerned with inference for* μ *based on* $p(\mu|\boldsymbol{x})$*. The right panel is concerned with inference for* σ^2 *based on* $p(\sigma^2|\boldsymbol{x})$*. The vertical lines show the true values of* μ *and* σ^2 *from which the data were generated. The accuracy values are as defined by (5.1).*

The sample sizes in the simulation study were

$$n \in \{25, 50, 100, 500, 1000, 5000\}.$$

For each sample size we replicated 100 random samples from the standard Normal distribution and obtained the expectation propagation approximations $q^*(\mu)$ and $q^*(\sigma^2)$ under model (4.2) with the hyperparameters set to be $\mu_\mu = 0$, $\sigma_\mu = A = 10^5$. We also obtained the mean field variational Bayes approximations using the special case of Algorithm 1 in Luts *et al.* [5] with $\boldsymbol{X}$ set to the $n \times 1$ vector of ones, $\boldsymbol{y}$ replaced by $\boldsymbol{x}$ and $\boldsymbol{\beta}$ replaced by μ. Exact computation of $p(\mu \,|\, \boldsymbol{x})$ and $p(\sigma^2 \,|\, \boldsymbol{x})$ is numerically challenging so we instead used binned kernel density estimation with direct plug-in bandwidth selection, as facilitated in the R package `KernSmooth` (Wand and Ripley [16]), applied to 1 million Markov chain Monte Carlo samples, following a burnin of size 1000. The R package `rstan` (Stan Development Team [13]) was used for Markov chain Monte Carlo. The very high sample size on which the kernel density estimates are based guarantees very good approximation of the required posterior density functions.

Figure 4 shows the 'exact' posterior density functions $p(\mu \,|\, \boldsymbol{x})$ and $p(\sigma^2 \,|\, \boldsymbol{x})$, their expectation propagation approximations $q^*(\mu)$ and $q^*(\sigma^2)$, and their mean field variational Bayes approximations, for the first replication from the simulation study with $n = 25$.

Summaries of accuracy scores of $q^*(\mu)$ and $q^*(\sigma^2)$ for all 100 replications are provided by the side-by-side boxplots of Figure 5. The accuracies are seen to be uniformly above 97% and mainly between 99% and 100%. This represents excellent performance for an approximate inference procedure. In Figure 6 we summarize the *difference* in accuracy of expectation propagation compared

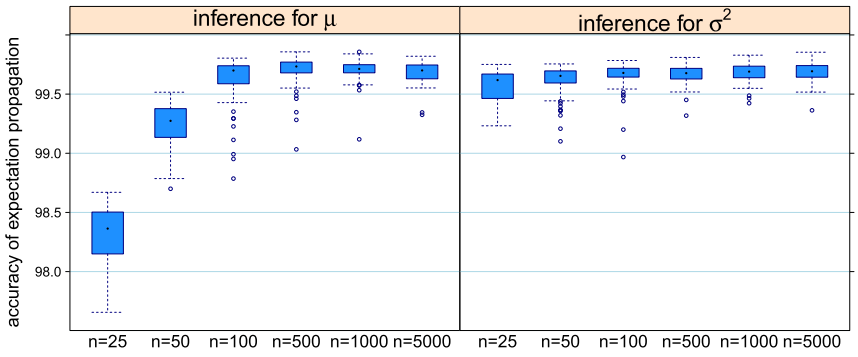

FIG 5. *Boxplots of accuracy values of expectation propagation obtained from the simulation study described in the text. The left panel summarizes accuracy values of* $q^*(\mu)$. *The right panel summarizes accuracy values of* $q^*(\sigma^2)$.

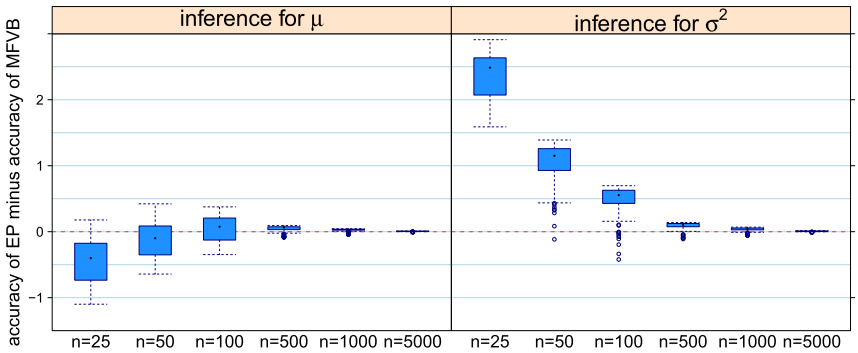

FIG 6. *Boxplots of the improvement in the accuracy of expectation propagation compared with mean field variational Bayes obtained from the simulation study described in the text. The boxplots are obtained from the subtracting the accuracy of the mean field variational Bayes (MFVB) approximate posterior density function from the accuracy of the expectation propagation (EP) approximate posterior density function based on the same sample. The left panel summarizes improvement in accuracy of* $q^*(\mu)$. *The right panel summarizes improvement in accuracy of* $q^*(\sigma^2)$.

with mean field variational Bayes. The boxplots in the left panel are obtained by subtracting the accuracy of mean field variational Bayes, based on the analogue of (5.1), from the accuracy of $q^*(\mu)$ for approximations based on the same sample in the simulation study. For low to moderate sample sizes expectation propagation is seen to be less accurate than mean field variational Bayes regarding inference for μ. For $n = 500$ there is a slight advantage of expectation propagation, but the differences diminish to zero as the sample size increases into the thousands. In the case of inference for σ^2, expectation propagation is 1–2 percentage points better than mean field variational Bayes for low to moderate samples. This advantage subsides for sample sizes in the thousands.

TABLE 1
Mean (standard deviation) computing times in seconds for expectation propagation and mean field variational Bayes for the simulation study described in text

sample size	times for expec. propagation	times for mean field variational Bayes
25	5.17 (0.372)	0.00081 (0.000394)
50	4.94 (0.347)	0.00084 (0.000368)
100	4.88 (0.331)	0.00088 (0.000383)
500	4.88 (0.349)	0.00096 (0.000374)
1000	4.88 (0.411)	0.00126 (0.000441)
5000	4.89 (0.371)	0.00309 (0.000473)

Expectation propagation is seen to have a slight edge over mean field variational Bayes because of the improvement it offers for inference concerning σ^2. This finding is in keeping with the simulation studies of Minka [7] and Bakker and Heskes[1] where versions of expectation propagation were shown to outperform variational approximations in specific contexts.

We also kept track of computational times and Table 1 provides summaries. The timings correspond to running 100 iterations of both expectation propagation and mean field variational Bayes in Version 3.2.0 of R [11] on a MacBook Air laptop with 8 gigabytes of random access memory and 1.7 gigahertz processor. As expected mean field variational Bayes is much faster since it involves purely algebraic updates, whereas expectation propagation requires time-consuming quadrature. Since both approaches depend only on the sum and sum of squares sufficient statistics the changes in computing time for higher sample sizes is negligible. Expectation propagation is slightly slower for $n = 25$, which is probably due to the numerical integrals being slower to converge in this more difficult low data situation.

6. Conclusions

We have carried out a concrete study of expectation propagation for a specific statistical model. The algorithm in Section 4 shows precisely what is involved for practical implementation. For the Bayesian Normal random sample model with Half-Cauchy standard deviation prior expectation propagation is shown to provide excellent accuracy, and offers improvements over mean field variational Bayes for larger sample sizes. This improvement in accuracy needs to be traded off against computational complexity. Expectation propagation requires several numerical integration evaluations whereas mean field variational Bayes involves simple arithmetic computations, as listed in Algorithm 1 of Luts *et al.* [5].

Appendix A: Appendix: Proofs and derivations

A.1. Proof of Theorem 1

From Lemma 1 of Guo and Qi [4]

$$\log(x) - \frac{1}{x} < \text{digamma}(x) < \log(x) - \frac{1}{2x} \quad \text{for all } x > 0 \tag{A.1}$$

and

$$\frac{1}{x} + \frac{1}{2x^2} < \text{trigamma}(x) < \frac{1}{x} + \frac{1}{x^2} \quad \text{for all } x > 0 \tag{A.2}$$

where $\text{trigamma}(x) \equiv \frac{d}{dx}\text{digamma}(x)$. From (A.1) it follows that

$$(\log - \text{digamma})(x) > \frac{1}{2x} \quad \text{for all } x > 0,$$

implying that $\log - \text{digamma}$ is a mapping from $\mathbb{R}^+$ to $\mathbb{R}^+$. Next note that

$$\frac{d}{dx}(\log - \text{digamma})(x) = \frac{1}{x} - \text{trigamma}(x) < -\frac{1}{2x^2} < 0 \quad \text{for all } x > 0$$

where we have used (A.2). Hence $\log - \text{digamma}$ is strictly monotonically decreasing on $\mathbb{R}^+$. Lastly, a rearrangement of (A.1) is

$$\frac{1}{2x} < (\log - \text{digamma})(x) < \frac{1}{x} \quad \text{for all } x > 0$$

which leads immediately to

$$\lim_{x\to\infty}(\log - \text{digamma})(x) = 0 \quad \text{and} \quad \lim_{x\to 0^+}(\log - \text{digamma})(x) = +\infty.$$

Therefore $\log - \text{digamma}$ is a one-to-one function that maps $\mathbb{R}^+$ onto $\mathbb{R}^+$.

A.2. Derivation of Result 1

In the case of projection onto the Normal family, (10) becomes

$$\int_{-\infty}^{\infty} \begin{bmatrix} x \\ x^2 \end{bmatrix} p(x)\,dx = \begin{bmatrix} \mu \\ \mu^2 + \sigma^2 \end{bmatrix}$$

which corresponds to the system of equations

$$\begin{aligned} E(x) &= \mu \\ E(x^2) &= \mu^2 + \sigma^2. \end{aligned} \tag{A.3}$$

The optimal parameters μ^* and $(\sigma^2)^*$ are the solutions of (A.3). The solution is easily found to be

$$\begin{aligned} \mu^* &= E(x) \\ (\sigma^2)^* &= E(x^2) - (\mu^*). \end{aligned}$$

A.3. Derivation of Result 2

In the case of projection onto the Inverse Gamma family, (10) becomes

$$\int_{-\infty}^{\infty} \begin{bmatrix} \log(x) \\ 1/x \end{bmatrix} p(x)\,dx = \begin{bmatrix} \log(\lambda) - \text{digamma}(\kappa) \\ \kappa/\lambda \end{bmatrix}$$

which corresponds to the system of equations

$$\begin{aligned} E\{\log(x)\} &= \log(\lambda) - \text{digamma}(\kappa) \\ E(1/x) &= \kappa/\lambda. \end{aligned} \tag{A.4}$$

The optimal parameters κ^* and λ^* are the solutions of (A.4). The second equation of (A.4) gives the relationship

$$\lambda^* = \kappa^*/E(1/x).$$

and substitution into the first equation of (A.4) leads to

$$\begin{aligned} E\{\log(x)\} &= \log(\kappa^*/E(1/x)) - \text{digamma}(\kappa^*) \\ &= (\log - \text{digamma})(\kappa^*) - \log(E(1/x)). \end{aligned}$$

Hence

$$(\log - \text{digamma})(\kappa^*) = \log(E(1/x)) + E\{\log(x)\}$$

and Result 2 immediately follows.

A.4. Further function definitions

The following functions, which depend on the integral functions $\mathcal{A}$ and $\mathcal{B}$ defined in Section 2.1, are useful for describing the expectation propagation updates:

$$\alpha\left(k, \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}\right) \equiv \mathcal{A}\left(k, a_1, -a_2, \frac{-2c_2}{c_1}, \frac{c_3 - 2b_2}{c_1}, \frac{c_1 - 2b_1 - 2}{2}\right)$$

and

$$\beta\left(k, \ell, v, w, \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}\right) \equiv \mathcal{B}\left(k, \frac{\ell + c_1 - 1}{2} - a_1, \frac{c_1 c_3 - c_2^2}{2c_1} - a_2, -b_2\left(\frac{c_2}{c_1} + \frac{b_1}{2b_2}\right)^2, v, w\right).$$

Then let

$$g(\ell, v, w, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \equiv (\log - \text{digamma})^{-1}\left(\log\left\{\frac{\beta(0, \ell+1, v, w, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}{\beta(0, \ell-1, v, w, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}\right\} - \frac{\beta(1, \ell-1, v, w, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}{\beta(0, \ell-1, v, w, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}\right),$$

$$G^N(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{c}) \equiv \left[\frac{\alpha(2, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}{\alpha(0, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})} - \left\{\frac{\alpha(1, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}{\alpha(0, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}\right\}^2\right]^{-1} \begin{bmatrix} \alpha(1, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})/\alpha(0, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \\ -1/2 \end{bmatrix} - \boldsymbol{a},$$

$$G^{\mathrm{IG1}}\left(\boldsymbol{a}, \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}; \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}\right) \equiv$$
$$\begin{bmatrix} -1 - g(0, -2b_2/c_1, \frac{1}{2}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \\ \dfrac{-g(0, -2b_2/c_1, \frac{1}{2}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})\, \beta(0, -1, -2b_2/c_1, \frac{1}{2}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}{\beta(0, 1, -2b_2/c_1, \frac{1}{2}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})} \end{bmatrix} - \boldsymbol{a}$$

and

$$G^{\mathrm{IG2}}\left(\boldsymbol{a}, \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}; k\right) \equiv$$
$$\begin{bmatrix} -1 - g\left(k-2, -b_2, 1-k/2-b_1, \boldsymbol{a}, \begin{bmatrix} 0 \\ b_2 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}\right) \\ \dfrac{\left\{ \begin{array}{l} -g\left(k-2, -b_2, 1-k/2-b_1, \boldsymbol{a}, \begin{bmatrix} 0 \\ b_2 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}\right) \\ \times\beta\left(0, k-3, -b_2, 1-k/2-b_1, \boldsymbol{a}, \begin{bmatrix} 0 \\ b_2 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}\right) \end{array} \right\}}{\beta\left(0, k-1, -b_2, 1-k/2-b_1, \boldsymbol{a}, \begin{bmatrix} 0 \\ b_2 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}\right)} \end{bmatrix} - \boldsymbol{a}.$$

For stable computation of the fractions appearing in the above expressions it is imperative to work with logarithms. For example,

$$\frac{\alpha(1, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}{\alpha(0, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})} = \mathrm{sign}\{\alpha(1, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})\} \exp[\log|\alpha(1, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})| - \log\{\alpha(0, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})\}].$$

The numerator and denominator components each depend of versions of the function $\log|\mathcal{A}(p, q, r, s, t, u)|$ which, as discussed in Section 2.1, can be computed accurately using the strategy given in Appendix B of Wand *et al.* [17].

A.5. Derivation of Algorithm 1

Under product restriction (4.4) expectation propagation for the Normal random sample model is driven by minimization of

$$\mathrm{KL}\Big(q(\mu)q(\sigma^2)\, q(a) \,\Big\|\, p(\mu, \sigma^2, a|\boldsymbol{x})\Big)$$

From (3.7) we have

$$m_{p(\mu)\to\mu}(\mu) \propto \frac{\mathrm{proj}\left[m_{\mu\to p(\mu)}(\mu) \exp\left\{ \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \begin{bmatrix} \mu_\mu/\sigma_\mu^2 \\ -1/(2\sigma_\mu^2) \end{bmatrix} \right\} \Big/ Z\right]}{m_{\mu\to p(\mu)}(\mu)}$$

where Z is the normalizing factor such that the function of μ inside the proj$[\cdot]$ is a density function. Conjugacy of

$$m_{\mu \to p(\mu)}(\mu) \quad \text{with} \quad \exp\left\{ \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \begin{bmatrix} \mu_\mu/\sigma_\mu^2 \\ -1/(2\sigma_\mu^2) \end{bmatrix} \right\}$$

implies that $m_{\mu \to p(\mu)}(\mu)$ is proportional to a Normal density function which, in turn implies that $m_{p(\mu) \to \mu}(\mu)$ is proportional to a Normal density function. Application of the updates (3.5) and (3.6) and enforcement of conjugacy constraints leads to:

> messages involving μ are proportional to Normal density functions and messages involving σ^2 and a are proportional to Inverse Gamma density functions. (A.5)

Under (A.5) we then have the messages between neighboring nodes on the factor graph in Figure 1 assuming the following forms:

$$\begin{aligned}
m_{p(\mu) \to \mu}(\mu) &= \exp\left\{ \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \boldsymbol{\eta}_{p(\mu) \to \mu} \right\}, \\
m_{\mu \to p(\mu)}(\mu) &= \exp\left\{ \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \boldsymbol{\eta}_{\mu \to p(\mu)} \right\}, \\
m_{\mu \to p(\boldsymbol{x}|\mu,\sigma^2)}(\mu) &= \exp\left\{ \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \boldsymbol{\eta}_{\mu \to p(\boldsymbol{x}|\mu,\sigma^2)} \right\}, \\
m_{p(\boldsymbol{x}|\mu,\sigma^2) \to \mu}(\mu) &= \exp\left\{ \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2) \to \mu} \right\}, \\
m_{p(\boldsymbol{x}|\mu,\sigma^2) \to \sigma^2}(\sigma^2) &= \exp\left\{ \begin{bmatrix} \log(\sigma^2) \\ 1/\sigma^2 \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2) \to \sigma^2} \right\}, \\
m_{\sigma^2 \to p(\boldsymbol{x}|\mu,\sigma^2)}(\sigma^2) &= \exp\left\{ \begin{bmatrix} \log(\sigma^2) \\ 1/\sigma^2 \end{bmatrix}^T \boldsymbol{\eta}_{\sigma^2 \to p(\boldsymbol{x}|\mu,\sigma^2)} \right\}, \\
m_{\sigma^2 \to p(\sigma^2|\,a)}(\sigma^2) &= \exp\left\{ \begin{bmatrix} \log(\sigma^2) \\ 1/\sigma^2 \end{bmatrix}^T \boldsymbol{\eta}_{\sigma^2 \to p(\sigma^2|\,a)} \right\}, \\
m_{p(\sigma^2|\,a) \to \sigma^2}(\sigma^2) &= \exp\left\{ \begin{bmatrix} \log(\sigma^2) \\ 1/\sigma^2 \end{bmatrix}^T \boldsymbol{\eta}_{p(\sigma^2|\,a) \to \sigma^2} \right\}, \\
m_{p(\sigma^2|\,a) \to a}(a) &= \exp\left\{ \begin{bmatrix} \log(a) \\ 1/a \end{bmatrix}^T \boldsymbol{\eta}_{p(\sigma^2|\,a) \to a} \right\}, \\
m_{a \to p(\sigma^2|\,a)}(a) &= \exp\left\{ \begin{bmatrix} \log(a) \\ 1/a \end{bmatrix}^T \boldsymbol{\eta}_{a \to p(\sigma^2|\,a)} \right\}, \\
m_{a \to p(a)}(a) &= \exp\left\{ \begin{bmatrix} \log(a) \\ 1/a \end{bmatrix}^T \boldsymbol{\eta}_{a \to p(a)} \right\} \\
\text{and} \quad m_{p(a) \to a}(a) &= \exp\left\{ \begin{bmatrix} \log(a) \\ 1/a \end{bmatrix}^T \boldsymbol{\eta}_{p(a) \to a} \right\}
\end{aligned} \tag{A.6}$$

where, for example, $\boldsymbol{\eta}_{p(\mu) \to \mu}$ is the natural parameter vector of $m_{p(\mu) \to \mu}(\mu)$. This notation has the advantage of making it easy to match a natural parameter vector with its corresponding message. However, it is cumbersome to use in the derivations of the updates and we also adopt the following abbreviated notation for some of the natural parameters:

$$\begin{aligned} \boldsymbol{\eta}^{\star} &\equiv \boldsymbol{\eta}_{\mu \to p(\boldsymbol{x}|\mu,\sigma^2)}, & \boldsymbol{\eta}^{\#} &\equiv \boldsymbol{\eta}_{\sigma^2 \to p(\boldsymbol{x}|\mu,\sigma^2)}, \\ \boldsymbol{\eta}^{\otimes} &\equiv \boldsymbol{\eta}_{\sigma^2 \to p(\sigma^2|\, a)} \quad \text{and} & \boldsymbol{\eta}^{\boxplus} &\equiv \boldsymbol{\eta}_{a \to p(\sigma^2|\, a)}. \end{aligned}$$

Additional useful notation is

$$\begin{aligned} A_{\text{N}}\left(\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}\right) &\equiv -\tfrac{1}{4}\,(\eta_1^2/\eta_2) - \tfrac{1}{2}\log(-2\eta_2), \\ \text{and} \qquad A_{\text{IG}}\left(\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}\right) &\equiv \log\Gamma(-\eta_1 - 1) - (-\eta_1 - 1)\,\log(-\eta_2) \end{aligned} \tag{A.7}$$

for the log-partition functions of the Normal and Inverse Gamma density functions with natural parameters η_1 and η_2. Also, we define

$$\overline{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \|\boldsymbol{x}\|^2 \equiv \sum_{i=1}^{n} x_i^2 \quad \text{and} \quad s^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Expectation propagation for the Normal random sample model reduces to updating the density functions in (A.6) which, in turn, reduces to updating each of their natural parameter vectors.

The stochastic node to factor message updates are very simple, and are summarized in Appendix A.5.1. The factor to stochastic node message updates are quite involved, and Appendices A.5.2–A.5.7 describe their derivations based on (3.6).

A.5.1. Derivations of stochastic node to factor message updates

The messages from stochastic nodes to factors have much simpler update derivations, based on (54) of Minka [8]. For example, the message from σ^2 to $p(\boldsymbol{x}|\,\mu,\sigma^2)$ is proportional to the product of the factor to σ^2 messages other than the message passed from $p(\boldsymbol{x}|\,\mu,\sigma^2)$. The only other factor neighboring σ^2 is $p(\sigma^2|a)$, so we get

$$m_{\sigma^2 \to p(\boldsymbol{x}|\mu,\sigma^2)}(\sigma^2) = m_{p(\sigma^2|\, a) \to \sigma^2}(\sigma^2).$$

This implies that the update for the $m_{\sigma^2 \to p(\boldsymbol{x}|\mu,\sigma^2)}(\sigma^2)$ natural parameter should be

$$\boldsymbol{\eta}_{\sigma^2 \to p(\boldsymbol{x}|\mu,\sigma^2)} \longleftarrow \boldsymbol{\eta}_{p(\sigma^2|\, a) \to \sigma^2}.$$

Similarly, the natural parameter updates for stochastic node to factor messages are

$$\boldsymbol{\eta}_{\mu \to p(\mu)} \longleftarrow \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2) \to \mu}, \quad \boldsymbol{\eta}_{\mu \to p(\boldsymbol{x}|\mu,\sigma^2)} \longleftarrow \boldsymbol{\eta}_{p(\mu) \to \mu}$$

$$\boldsymbol{\eta}_{\sigma^2 \to p(\sigma^2 | a)} \longleftarrow \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2) \to \sigma^2}, \quad \boldsymbol{\eta}_{a \to p(\sigma^2 | a)} \longleftarrow \boldsymbol{\eta}_{p(a) \to a}$$

and

$$\boldsymbol{\eta}_{a \to p(a)} \longleftarrow \boldsymbol{\eta}_{p(\sigma^2 | a) \to a}.$$

A.5.2. Derivation of the $m_{p(\mu) \to \mu}(\mu)$ update

From (3.6),

$$m_{p(\mu) \to \mu}(\mu) \propto \frac{\text{proj}[m_{\mu \to p(\mu)}(\mu)\, p(\mu)/Z]}{m_{\mu \to p(\mu)}(\mu)}.$$

Then, from (A.6),

$$m_{\mu \to p(\mu)}(\mu)\, p(\mu) \propto \exp\left\{ \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \left(\boldsymbol{\eta}_{\mu \to p(\mu)} + \begin{bmatrix} \mu_\mu/\sigma_\mu^2 \\ -1/(2\sigma_\mu^2) \end{bmatrix} \right) \right\}.$$

Since $m_{\mu \to p(\mu)}(\mu)\, p(\mu)$ is proportional to a Normal density function, its projection onto the Normal family is the same function up to multiplicative factors. Hence

$$\text{proj}[m_{\mu \to p(\mu)}(\mu)\, p(\mu)] \propto \exp\left\{ \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \left(\boldsymbol{\eta}_{\mu \to p(\mu)} + \begin{bmatrix} \mu_\mu/\sigma_\mu^2 \\ -1/(2\sigma_\mu^2) \end{bmatrix} \right) \right\}$$

and so, dividing by $m_{\mu \to p(\mu)}(\mu)$, we get

$$m_{p(\mu) \to \mu}(\mu) = \exp\left\{ \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \begin{bmatrix} \mu_\mu/\sigma_\mu^2 \\ -1/(2\sigma_\mu^2) \end{bmatrix} \right\}.$$

Hence

$$\boldsymbol{\eta}_{p(\mu) \to \mu} = \begin{bmatrix} \mu_\mu/\sigma_\mu^2 \\ -1/(2\sigma_\mu^2) \end{bmatrix}$$

which remains constant throughout the iterations.

A.5.3. Derivation of the $m_{p(\boldsymbol{x}|\mu,\sigma^2) \to \mu}(\mu)$ update

Equation (3.6) applied to the message $m_{p(\boldsymbol{x}|\mu,\sigma^2) \to \mu}(\mu)$ is

$$m_{p(\boldsymbol{x}|\mu,\sigma^2) \to \mu}(\mu) \propto \frac{\text{proj}\left[m_{\mu \to p(\boldsymbol{x}|\mu,\sigma^2)}(\mu) \int_0^\infty p(\boldsymbol{x}|\mu,\sigma^2) m_{\sigma^2 \to p(\boldsymbol{x}|\mu,\sigma^2)}(\sigma^2) d\sigma^2 / Z \right]}{m_{\mu \to p(\boldsymbol{x}|\mu,\sigma^2)}(\mu)}. \tag{A.8}$$

Then the integral in (A.8) is

$$\int_0^\infty p(\boldsymbol{x}|\mu,\sigma^2) m_{\sigma^2 \to p(\boldsymbol{x}|\mu,\sigma^2)}(\sigma^2)\, d\sigma^2$$

$$
\begin{aligned}
&= \int_0^\infty (2\pi\sigma^2)^{-n/2} \exp\left\{-\|\boldsymbol{x}-\mathbf{1}\mu\|^2/(2\sigma^2)\right\} \exp\left\{ \begin{bmatrix} \log\sigma^2 \\ 1/\sigma^2 \end{bmatrix}^T \boldsymbol{\eta}^{\#} \right\} d\sigma^2 \\
&= (2\pi)^{-n/2} \int_0^\infty \exp\left\{ \begin{bmatrix} \log\sigma^2 \\ 1/\sigma^2 \end{bmatrix}^T \left(\begin{bmatrix} -n/2 \\ -\|\boldsymbol{x}-\mathbf{1}\mu\|^2/2 \end{bmatrix} + \boldsymbol{\eta}^{\#} \right) \right\} d\sigma^2 \\
&\propto \exp\left\{ A_{\text{IG}} \left(\begin{bmatrix} -n/2 \\ -\|\boldsymbol{x}-\mathbf{1}\mu\|^2/2 \end{bmatrix} + \boldsymbol{\eta}^{\#} \right) \right\} \\
&\propto \left(\frac{\|\boldsymbol{x}-\mathbf{1}\mu\|^2}{2} - \eta_2^{\#} \right)^{-n/2+\eta_1^{\#}+1}.
\end{aligned}
$$

Noting that

$$
m_{\mu \to p(\boldsymbol{x}|\mu,\sigma^2)}(\mu) = \exp\left\{ \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \boldsymbol{\eta}^{\star} \right\},
$$

the density function inside the proj operator in (A.8)

$$
p_\bullet(\mu) \equiv \frac{\exp\left(\eta_1^\star \mu + \eta_2^\star \mu^2\right)}{Z_\bullet \left(\mu^2 - 2\overline{x}\mu + \frac{\|\boldsymbol{x}\|^2 - 2\eta_2^{\#}}{n}\right)^{\frac{n}{2}-\eta_1^{\#}-1}}.
$$

where $Z_\bullet$ is the normalizing factor. From Result 1, the projection of $p_\bullet$ onto the family of Normal density functions is the $N(\mu_\bullet^*, (\sigma^2)_\bullet^*)$ density function where

$$
\mu_\bullet^* = \textstyle\int_{-\infty}^{\infty} \mu\, p_\bullet(\mu)\, d\mu \;\text{ and }\; (\sigma^2)_\bullet^* = \int_{-\infty}^{\infty} (\mu^2) p_\bullet(\mu)\, d\mu - (\mu_\bullet^*)^2. \tag{A.9}
$$

The integrals in (A.9) can be presented in terms of the function $\mathcal{A}$, defined in equation (5), as follows:

$$
\int_{-\infty}^{\infty} \mu\, p_\bullet(\mu)\, d\mu = \frac{\mathcal{A}\left(1, \eta_1^\star, -\eta_2^\star, -\frac{2c_2}{c_1}, \frac{c_3-2\eta_2^{\#}}{c_1}, \frac{c_1-2\eta_1^{\#}-2}{2}\right)}{\mathcal{A}\left(0, \eta_1^\star, -\eta_2^\star, -\frac{2c_2}{c_1}, \frac{c_3-2\eta_2^{\#}}{c_1}, \frac{c_1-2\eta_1^{\#}-2}{2}\right)}
$$

$$
\text{and} \quad \int_{-\infty}^{\infty} (\mu^2) p_\bullet(\mu)\, d\mu = \frac{\mathcal{A}\left(2, \eta_1^\star, -\eta_2^\star, -\frac{2c_2}{c_1}, \frac{c_3-2\eta_2^{\#}}{c_1}, \frac{c_1-2\eta_1^{\#}-2}{2}\right)}{\mathcal{A}\left(0, \eta_1^\star, -\eta_2^\star, -\frac{2c_2}{c_1}, \frac{c_3-2\eta_2^{\#}}{c_1}, \frac{c_1-2\eta_1^{\#}-2}{2}\right)}
$$

where $(c_1, c_2, c_3) = (n, \sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$. Since the natural parameter vector of proj$[p_\bullet]$ is

$$
\begin{bmatrix} \mu_\bullet^*/(\sigma^2)_\bullet^* \\ -1/\{2(\sigma^2)_\bullet^*\} \end{bmatrix},
$$

(A.8) can be expressed as:

$$
m_{p(\boldsymbol{x}|\mu,\sigma^2) \to \mu}(\mu) = \exp\left\{ \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2) \to \mu} \right\}
$$

where

$$\boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu} =$$

$$\left[\begin{array}{c} \left\{\dfrac{\mathcal{A}\left(1,\eta_1^\star,-\eta_2^\star,-\frac{2c_2}{c_1},\frac{c_3-2\eta_2^\#}{c_1},\frac{c_1-2\eta_1^\#-2}{2}\right)}{\mathcal{A}\left(0,\eta_1^\star,-\eta_2^\star,-\frac{2c_2}{c_1},\frac{c_3-2\eta_2^\#}{c_1},\frac{c_1-2\eta_1^\#-2}{2}\right)}\right\}\Bigg/ \\ \left[\dfrac{\mathcal{A}\left(2,\eta_1^\star,-\eta_2^\star,-\frac{2c_2}{c_1},\frac{c_3-2\eta_2^\#}{c_1},\frac{c_1-2\eta_1^\#-2}{2}\right)}{\mathcal{A}\left(0,\eta_1^\star,-\eta_2^\star,-\frac{2c_2}{c_1},\frac{c_3-2\eta_2^\#}{c_1},\frac{c_1-2\eta_1^\#-2}{2}\right)}- \right. \\ \left. \left\{\left\{\dfrac{\mathcal{A}\left(1,\eta_1^\star,-\eta_2^\star,-\frac{2c_2}{c_1},\frac{c_3-2\eta_2^\#}{c_1},\frac{c_1-2\eta_1^\#-2}{2}\right)}{\mathcal{A}\left(0,\eta_1^\star,-\eta_2^\star,-\frac{2c_2}{c_1},\frac{c_3-2\eta_2^\#}{c_1},\frac{c_1-2\eta_1^\#-2}{2}\right)}\right\}\right\}^2\right] \\ \\ -\frac{1}{2}\Big/ \\ \left[\dfrac{\mathcal{A}\left(2,\eta_1^\star,-\eta_2^\star,-\frac{2c_2}{c_1},\frac{c_3-2\eta_2^\#}{c_1},\frac{c_1-2\eta_1^\#-2}{2}\right)}{\mathcal{A}\left(0,\eta_1^\star,-\eta_2^\star,-\frac{2c_2}{c_1},\frac{c_3-2\eta_2^\#}{c_1},\frac{c_1-2\eta_1^\#-2}{2}\right)}- \right. \\ \left. \left\{\left\{\dfrac{\mathcal{A}\left(1,\eta_1^\star,-\eta_2^\star,-\frac{2c_2}{c_1},\frac{c_3-2\eta_2^\#}{c_1},\frac{c_1-2\eta_1^\#-2}{2}\right)}{\mathcal{A}\left(0,\eta_1^\star,-\eta_2^\star,-\frac{2c_2}{c_1},\frac{c_3-2\eta_2^\#}{c_1},\frac{c_1-2\eta_1^\#-2}{2}\right)}\right\}\right\}^2\right] \end{array}\right] - \boldsymbol{\eta}^\star$$

$$= G^N(\boldsymbol{\eta}^\star, \boldsymbol{\eta}^\#; c_1, c_2, c_3).$$

Therefore, updates of $m_{\mu\to p(\boldsymbol{x}|\mu,\sigma^2)}(\mu)$ correspond to the natural parameter update

$$\boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu} \longleftarrow G^N\left(\boldsymbol{\eta}_{\mu\to p(\boldsymbol{x}|\mu,\sigma^2)}, \boldsymbol{\eta}_{\sigma^2\to p(\boldsymbol{x}|\mu,\sigma^2)}; n, \sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right).$$

A.5.4. Derivation of the $m_{p(\sigma^2|\,a)\to\sigma^2}(\sigma^2)$ update

From (3.6),

$$m_{p(\sigma^2|\,a)\to\sigma^2}(\sigma^2) \propto \frac{\text{proj}\left[\begin{array}{l} Z^{-1}m_{\sigma^2\to p(\sigma^2|\,a)}(\sigma^2) \\ \times\int_0^\infty p(\sigma^2|a)m_{a\to p(\sigma^2|\,a)}(a)\,da \end{array}\right]}{m_{\sigma^2\to p(\sigma^2|\,a)}(\sigma^2)}. \tag{A.10}$$

The integral in (A.10) is

$$\int_0^\infty p(\sigma^2|a)m_{a\to p(\sigma^2|\,a)}(a)\,da$$

$$
\begin{aligned}
&= \int_0^\infty \frac{(1/a)^{\frac{1}{2}}}{\Gamma(\frac{1}{2})}(\sigma^2)^{-\frac{1}{2}-1}\exp\{-1/(a\sigma^2)\}\,\exp\left\{\begin{bmatrix}\log(a)\\ 1/a\end{bmatrix}^T \boldsymbol{\eta}^{\boxplus}\right\} da \\
&= \Gamma(\tfrac{1}{2})^{-1}\,(\sigma^2)^{-3/2}\int_0^\infty \exp\left\{\begin{bmatrix}\log(a)\\ 1/a\end{bmatrix}^T \left(\begin{bmatrix}-1/2\\ -1/\sigma^2\end{bmatrix} + \boldsymbol{\eta}^{\boxplus}\right)\right\} da \\
&\propto (\sigma^2)^{-3/2}\exp\left\{A_{\mathrm{IG}}\left(\begin{bmatrix}-1/2\\ -1/\sigma^2\end{bmatrix} + \boldsymbol{\eta}^{\boxplus}\right)\right\} \\
&\propto (\sigma^2)^{-3/2}\left(\frac{1}{\sigma^2} - \eta_2^{\boxplus}\right)^{\frac{1}{2}+\eta_1^{\boxplus}}.
\end{aligned}
$$

Since

$$
m_{\sigma^2 \to p(\sigma^2|\,a)}(\sigma^2) = \exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma^2\end{bmatrix}^T \boldsymbol{\eta}^{\otimes}\right\}
$$

the density function inside the proj operator in (A.10) is

$$
p_\circ(\sigma^2) \equiv (\sigma^2)^{\eta_1^{\otimes}-3/2}\left(\frac{1}{\sigma^2} - \eta_2^{\boxplus}\right)^{\frac{1}{2}+\eta_1^{\boxplus}} \exp(\eta_2^{\otimes}/\sigma^2)/Z_\circ, \quad \sigma^2 > 0,
$$

where $Z_\circ$ is the normalizing factor. From Result 2, proj$[p_\circ]$ is the IG$(\kappa_\circ^*, \lambda_\circ^*)$ density function where

$$
\begin{aligned}
\kappa_\circ^* = &(\log - \text{digamma})^{-1}\Big(\log\left\{\textstyle\int_0^\infty (\frac{1}{\sigma^2}) p_\circ(\sigma^2)\,d\sigma^2\right\} \\
&\qquad + \textstyle\int_0^\infty \log(\sigma^2) p_\circ(\sigma^2)\,d\sigma^2\Big)
\end{aligned}
\qquad \text{(A.11)}
$$

and

$$
\lambda_\circ^* = \kappa_\circ^* \Big/ \textstyle\int_0^\infty (1/\sigma^2) p_\circ(\sigma^2)\,d\sigma^2.
$$

Using the change of variable $\sigma^2 = e^{-x}$ and straightforward algebra, the integrals in (A.11) can be expressed in terms of the $\mathcal{B}$ function, defined at equation (5), as follows:

$$
\int_0^\infty (1/\sigma^2) p_\circ(\sigma^2)\,d\sigma^2 = \frac{\mathcal{B}(0, \frac{3}{2} - \eta_1^{\otimes}, -\eta_2^{\otimes}, 0, -\eta_2^{\boxplus}, -\frac{1}{2} - \eta_1^{\boxplus})}{\mathcal{B}(0, \frac{1}{2} - \eta_1^{\otimes}, -\eta_2^{\otimes}, 0, -\eta_2^{\boxplus}, -\frac{1}{2} - \eta_1^{\boxplus})}
$$

and

$$
\int_0^\infty \log(\sigma^2)\, p_\circ(\sigma^2)\,d\sigma^2 = \frac{-\mathcal{B}(1, \frac{1}{2} - \eta_1^{\otimes}, -\eta_2^{\otimes}, 0, -\eta_2^{\boxplus}, -\frac{1}{2} - \eta_1^{\boxplus})}{\mathcal{B}(0, \frac{1}{2} - \eta_1^{\otimes}, -\eta_2^{\otimes}, 0, -\eta_2^{\boxplus}, -\frac{1}{2} - \eta_1^{\boxplus})}.
$$

Noting that the natural parameter vector of proj$[p_\circ]$ is $[-\kappa_\circ^* - 1 \quad -\lambda_\circ^*]^T$ the message from $p(\sigma^2|a)$ to σ^2 is

$$
m_{p(\sigma^2|\,a) \to \sigma^2}(\sigma^2) = \exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma^2\end{bmatrix}^T \boldsymbol{\eta}_{p(\sigma^2|\,a) \to \sigma^2}\right\}
$$

where

$$\boldsymbol{\eta}_{p(\sigma^2\,|\,a)\,\to\,\sigma^2} =$$

$$\left[\begin{array}{l} -(\log -\text{digamma})^{-1}\left\{\log\left(\dfrac{\mathcal{B}(0,\frac{3}{2}-\eta_1^{\otimes},-\eta_2^{\otimes},0,-\eta_2^{\boxplus},-\frac{1}{2}-\eta_1^{\boxplus})}{\mathcal{B}(0,\frac{1}{2}-\eta_1^{\otimes},-\eta_2^{\otimes},0,-\eta_2^{\boxplus},-\frac{1}{2}-\eta_1^{\boxplus})}\right)\right. \\ \qquad\qquad \left. -\dfrac{\mathcal{B}(1,\frac{1}{2}-\eta_1^{\otimes},-\eta_2^{\otimes},0,-\eta_2^{\boxplus},-\frac{1}{2}-\eta_1^{\boxplus})}{\mathcal{B}(0,\frac{1}{2}-\eta_1^{\otimes},-\eta_2^{\otimes},0,-\eta_2^{\boxplus},-\frac{1}{2}-\eta_1^{\boxplus})}\right\} - 1 \\ \\ (\log -\text{digamma})^{-1}\left\{\log\left(\dfrac{\mathcal{B}(0,\frac{3}{2}-\eta_1^{\otimes},-\eta_2^{\otimes},0,-\eta_2^{\boxplus},-\frac{1}{2}-\eta_1^{\boxplus})}{\mathcal{B}(0,\frac{1}{2}-\eta_1^{\otimes},-\eta_2^{\otimes},0,-\eta_2^{\boxplus},-\frac{1}{2}-\eta_1^{\boxplus})}\right)\right. \\ \qquad\qquad \left. -\dfrac{\mathcal{B}(1,\frac{1}{2}-\eta_1^{\otimes},-\eta_2^{\otimes},0,-\eta_2^{\boxplus},-\frac{1}{2}-\eta_1^{\boxplus})}{\mathcal{B}(0,\frac{1}{2}-\eta_1^{\otimes},-\eta_2^{\otimes},0,-\eta_2^{\boxplus},-\frac{1}{2}-\eta_1^{\boxplus})}\right\} \\ \qquad\qquad \times\dfrac{\mathcal{B}(0,\frac{1}{2}-\eta_1^{\otimes},-\eta_2^{\otimes},0,-\eta_2^{\boxplus},-\frac{1}{2}-\eta_1^{\boxplus})}{\mathcal{B}(0,\frac{3}{2}-\eta_1^{\otimes},-\eta_2^{\otimes},0,-\eta_2^{\boxplus},-\frac{1}{2}-\eta_1^{\boxplus})} \end{array}\right] - \boldsymbol{\eta}^{\otimes}$$

$$= G^{\text{IG2}}(\boldsymbol{\eta}^{\otimes},\boldsymbol{\eta}^{\boxplus};3) = G^{\text{IG2}}(\boldsymbol{\eta}_{\sigma^2\,\to\,p(\sigma^2\,|\,a)},\boldsymbol{\eta}_{a\,\to\,p(\sigma^2\,|\,a)};3).$$

Therefore, updates of $m_{p(\sigma^2\,|\,a)\,\to\,\sigma^2}(\sigma^2)$ correspond to the natural parameter update

$$\boldsymbol{\eta}_{p(\sigma^2\,|\,a)\,\to\,\sigma^2} \longleftarrow G^{\text{IG2}}(\boldsymbol{\eta}_{\sigma^2\,\to\,p(\sigma^2\,|\,a)},\boldsymbol{\eta}_{a\,\to\,p(\sigma^2\,|\,a)};3).$$

A.5.5. Derivation of the $m_{p(\boldsymbol{x}|\mu,\,\sigma^2)\,\to\,\sigma^2}(\sigma^2)$ update

The message from $p(\boldsymbol{x}\,|\,\mu,\sigma^2)$ to σ^2 is, according to (3.6),

$$m_{p(\boldsymbol{x}|\mu,\,\sigma^2)\,\to\,\sigma^2}(\sigma^2) \propto \frac{\text{proj}\left[\begin{array}{l} Z^{-1}m_{\sigma^2\,\to\,p(\boldsymbol{x}|\mu,\,\sigma^2)}(\sigma^2) \\ \times\int_{-\infty}^{\infty} p(\boldsymbol{x}|\mu,\sigma^2)\,m_{\mu\,\to\,p(\boldsymbol{x}|\mu,\,\sigma^2)}(\mu)\,d\mu \end{array}\right]}{m_{\sigma^2\,\to\,p(\boldsymbol{x}|\mu,\,\sigma^2)}(\sigma^2)}. \tag{A.12}$$

First note that

$$p(\boldsymbol{x}\,|\,\mu,\sigma^2) = n^{-1/2}(2\pi\sigma^2)^{(1-n)/2}\exp\left\{\frac{-(n-1)s^2}{2\sigma^2}\right\}\{2\pi(\sigma^2/n)\}^{-1/2}$$

$$\times\exp\left\{\frac{-(\mu-\overline{x})^2}{2(\sigma^2/n)}\right\}$$

and

$$m_{\mu\,\to\,p(\boldsymbol{x}|\mu,\,\sigma^2)}(\mu) \propto \{2\pi(\sigma^\star)^2\}^{-1/2}\exp\left\{\frac{-(\mu-\mu^\star)^2}{2(\sigma^\star)^2}\right\}$$

where $\mu^\star = -\eta_1^\star/(2\eta_1^\star)$ and $(\sigma^\star)^2 = -1/(2\eta_1^\star)$. Then by (A.2) of Wand and Jones [15],

$$\begin{aligned}\int_{-\infty}^{\infty} & p(\boldsymbol{x}|\mu,\sigma^2)\,\{2\pi(\sigma^\star)^2\}^{-1/2}\exp\left\{\frac{-(\mu-\mu^\star)^2}{2(\sigma^\star)^2}\right\}d\mu \\ &= n^{-1/2}(2\pi\sigma^2)^{(1-n)/2}\exp\left\{\frac{-(n-1)s^2}{2\sigma^2}\right\}[2\pi\{(\sigma^2/n)+(\sigma^\star)^2\}]^{-1/2} \\ &\quad\times\exp\left\{\frac{-(\overline{x}-\mu^\star)^2}{2\{(\sigma^2/n)+(\sigma^\star)^2\}}\right\}.\end{aligned}$$

Using the fact that

$$m_{\sigma^2\to p(\boldsymbol{x}|\mu,\sigma^2)}(\sigma^2) = \exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma^2\end{bmatrix}^T\boldsymbol{\eta}^{\#}\right\},$$

the function inside the proj operator in (A.12) is then

$$\begin{aligned}p_\diamond(\sigma^2) \equiv \frac{1}{Z_\diamond}(\sigma^2)^{\eta_1^{\#}+(1-n)/2}\exp\left\{\frac{\eta_2^{\#}-\frac{1}{2}(n-1)s^2}{\sigma^2}\right\} \\ \times\left(\sigma^2-\frac{n}{2\eta_2^\star}\right)^{-1/2}\exp\left\{\frac{-n\left(\overline{x}+\frac{\eta_1^\star}{2\eta_2^\star}\right)^2}{2(\sigma^2-\frac{n}{2\eta_2^\star})}\right\},\quad \sigma^2>0,\end{aligned}$$

where $Z_\diamond$ is the normalizing factor. Kullback-Leibler projection of $p_\diamond$ onto the Inverse Gamma family of density functions requires the integrals

$$\int_0^\infty \log(\sigma^2)\,p_\diamond(\sigma^2)\,d\sigma^2 \quad\text{and}\quad \int_0^\infty (1/\sigma^2)\,p_\diamond(\sigma^2)\,d\sigma^2.$$

Via the change of variable $\sigma^2 = e^{-x}$ and some algebra, these integrals may be written in terms of the $\mathcal{B}$ function as follows:

$$\begin{aligned}\int_0^\infty & \log(\sigma^2)\,p_\diamond(\sigma^2)\,d\sigma^2 \\ &= \frac{-\mathcal{B}\left(1,\frac{1}{2}(n-1)-\eta_1^{\#},\frac{1}{2}(n-1)s^2-\eta_2^{\#},-\eta_2^\star\left(\overline{x}+\frac{\eta_1^\star}{2\eta_2^\star}\right)^2,\frac{-2\eta_2^\star}{n},\frac{1}{2}\right)}{\mathcal{B}\left(0,\frac{1}{2}(n-1)-\eta_1^{\#},\frac{1}{2}(n-1)s^2-\eta_2^{\#},-\eta_2^\star\left(\overline{x}+\frac{\eta_1^\star}{2\eta_2^\star}\right)^2,\frac{-2\eta_2^\star}{n},\frac{1}{2}\right)}\end{aligned}$$

and

$$\begin{aligned}\int_0^\infty & (1/\sigma^2)\,p_\diamond(\sigma^2)\,d\sigma^2 \\ &= \frac{\mathcal{B}\left(0,\frac{1}{2}(n+1)-\eta_1^{\#},\frac{1}{2}(n-1)s^2-\eta_2^{\#},-\eta_2^\star\left(\overline{x}+\frac{\eta_1^\star}{2\eta_2^\star}\right)^2,\frac{-2\eta_2^\star}{n},\frac{1}{2}\right)}{\mathcal{B}\left(0,\frac{1}{2}(n-1)-\eta_1^{\#},\frac{1}{2}(n-1)s^2-\eta_2^{\#},-\eta_2^\star\left(\overline{x}+\frac{\eta_1^\star}{2\eta_2^\star}\right)^2,\frac{-2\eta_2^\star}{n},\frac{1}{2}\right)}.\end{aligned}$$

Arguments similar to those used in Appendix A.5.4, which involve Kullback-Leibler projection of $p_\circ$ onto the Inverse Gamma family of density functions, lead to the natural parameter update

$$\boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\,\to\,\sigma^2} \longleftarrow G^{\text{IG1}}\left(\boldsymbol{\eta}_{\sigma^2\,\to\,p(\boldsymbol{x}|\mu,\sigma^2)}, \boldsymbol{\eta}_{\mu\,\to\,p(\boldsymbol{x}|\mu,\sigma^2)}; n, \sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right).$$

A.5.6. Derivation of the $m_{p(\sigma^2|\,a)\,\to\,a}(a)$ update

Equation (3.6) gives

$$m_{p(\sigma^2|\,a)\,\to\,a}(a) \propto \frac{\text{proj}\left[m_{a\,\to\,p(\sigma^2|\,a)}(a)\int_0^\infty p(\sigma^2|a) m_{\sigma^2\,\to\,p(\sigma^2|\,a)}(\sigma^2)\,d\sigma^2/Z\right]}{m_{a\,\to\,p(\sigma^2|\,a)}(a)}.$$

Arguments analogous to those given in the derivation of $m_{p(\sigma^2|\,a)\,\to\,\sigma^2}(\sigma^2)$ lead to

$$m_{p(\sigma^2|\,a)\,\to\,a}(a) = \exp\left\{\begin{bmatrix}\log(a)\\ 1/a\end{bmatrix}^T \boldsymbol{\eta}_{p(\sigma^2|\,a)\,\to\,a}\right\}$$

with natural parameter update

$$\boldsymbol{\eta}_{p(\sigma^2|\,a)\,\to\,a} \longleftarrow G^{\text{IG2}}(\boldsymbol{\eta}_{a\,\to\,p(\sigma^2|\,a)}, \boldsymbol{\eta}_{\sigma^2\,\to\,p(\sigma^2|\,a)}; 1).$$

A.5.7. Derivation of the $m_{p(a)\,\to\,a}(a)$ update

Using arguments similar to those used in Appendix A.5.2,

$$\begin{aligned} m_{p(a)\,\to\,a}(a) &\propto \frac{\text{proj}[m_{a\,\to\,p(a)}(a)\,p(a)/Z]}{m_{a\,\to\,p(a)}(a)} \\ &\propto p(a) \propto \exp\left\{\begin{bmatrix}\log(a)\\ 1/a\end{bmatrix}^T \begin{bmatrix}-3/2\\ -1/A^2\end{bmatrix}\right\}. \end{aligned}$$

Hence

$$\boldsymbol{\eta}_{p(a)\,\to\,a} = \begin{bmatrix}-3/2\\ -1/A^2\end{bmatrix}$$

which remains constant throughout the iterations.

A.5.8. Derivation of q-density construction

On convergence of the messages, the optimal q-densities for each of μ, σ^2 and a can be found via equation (44) of Minka and Winn [9].

Specifically

$$\begin{aligned} q(\mu) &\propto m_{p(\mu)\to\mu}(\mu)\, m_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu}(\mu), \\ q(\sigma^2) &\propto m_{p(\boldsymbol{x}|\mu,\sigma^2)\to\sigma^2}(\sigma^2)\, m_{p(\sigma^2|\,a)\to\sigma^2}(\sigma^2) \\ \text{and}\quad q(a) &\propto m_{p(\sigma^2|\,a)\to a}(a)\, m_{p(a)\to a}(a). \end{aligned}$$

These lead to

$$\begin{aligned} q(\mu) &\propto \exp\left\{ \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \left(\boldsymbol{\eta}_{p(\mu)\to\mu} + \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu} \right) \right\}, \\ q(\sigma^2) &\propto \exp\left\{ \begin{bmatrix} \log(\sigma^2) \\ 1/\sigma^2 \end{bmatrix}^T \left(\boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\sigma^2} + \boldsymbol{\eta}_{p(\sigma^2|\,a)\to\sigma^2} \right) \right\} \\ \text{and}\quad q(a) &\propto \exp\left\{ \begin{bmatrix} \log(a) \\ 1/a \end{bmatrix}^T \left(\boldsymbol{\eta}_{p(\sigma^2|\,a)\to a} + \boldsymbol{\eta}_{p(a)\to a} \right) \right\} \end{aligned}$$

and the optimal q-density functions follow immediately.

A.5.9. Derivation of the approximate log-likelihood expression

For the factor graph depicted in Figure 1, the approximate marginal log-likelihood expression (3.9) is

$$\begin{aligned} \log\{\underset{\sim}{p}(\boldsymbol{x};q)\} \;=\; & \log s_{p(\mu)} + \log s_{\mu} + \log s_{p(\boldsymbol{x}\,|\,\mu,\sigma^2)} + \log s_{\sigma^2} \\ & + \log s_{p(\sigma^2\,|\,a)} + \log s_{a} + \log s_{p(a)} \end{aligned} \tag{A.13}$$

where s_{μ}, s_{σ^2} and s_a are given by (3.10) and $s_{p(\mu)}$, $s_{p(\boldsymbol{x}\,|\,\mu,\sigma^2)}$ and $s_{p(\sigma^2\,|\,a)}$ are given by (3.11).

We first treat the terms corresponding to the stochastic nodes μ, σ^2 and a. From (46) of Minka and Winn [9],

$$\begin{aligned} s_{\mu} &= \int_{-\infty}^{\infty} m_{p(\mu)\to\mu}(\mu) m_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu}(\mu) d\mu \\ &= \int_{-\infty}^{\infty} \exp\left\{ \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \left(\boldsymbol{\eta}_{p(\mu)\to\mu} + \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu} \right) \right\} d\mu \end{aligned} \tag{A.14}$$

which leads to

$$\log\, s_{\mu} = A_{\mathrm{N}}\left(\boldsymbol{\eta}_{p(\mu)\to\mu} + \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\mu} \right) - \tfrac{1}{2}\log(2\pi)$$

Analogous arguments result in

$$\log s_{\sigma^2} = A_{\mathrm{IG}}\left(\boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\to\sigma^2} + \boldsymbol{\eta}_{p(\sigma^2|\,a)\to\sigma^2} \right)$$

and

$$\log s_a = A_{\text{IG}}\left(\boldsymbol{\eta}_{p(a)\,\to\,a} + \boldsymbol{\eta}_{p(\sigma^2|\,a)\,\to\,a}\right).$$

Recall that the functions A_{N} and A_{IG} are defined at (A.7).

The first term in (A.13) corresponding to a factor is, according to (48) of Minka and Winn [9],

$$s_{p(\mu)} = \frac{\int_{-\infty}^{\infty} m_{\mu\,\to\,p(\mu)}(\mu)\,p(\mu)\,d\mu}{\int_{-\infty}^{\infty} m_{\mu\,\to\,p(\mu)}(\mu) m_{p(\mu)\,\to\,\mu}(\mu)\,d\mu} = 1 \tag{A.15}$$

where we have used $m_{p(\mu)\,\to\,\mu}(\mu) = p(\mu)$. Hence $\log s_{p(\mu)} = 0$.

The expression for $s_{p(a)}$ is similar in nature to (A.15), and leads to $s_{p(a)} = 1$, which implies that $\log s_{p(a)} = 0$.

Next is $s_{p(\boldsymbol{x}|\mu,\sigma^2)}$ which, according to (48) of Minka and Winn [9], has logarithm

$$\begin{aligned}
&\log s_{p(\boldsymbol{x}|\mu,\sigma^2)} = \\
&\quad \log\left\{\int_{-\infty}^{\infty}\int_{0}^{\infty} m_{\mu\,\to\,p(\boldsymbol{x}|\mu,\sigma^2)}(\mu)\, m_{\sigma^2\,\to\,p(\boldsymbol{x}|\mu,\sigma^2)}(\sigma^2)\; p(\boldsymbol{x}|\mu,\sigma^2)\,d\,\sigma^2\,d\,\mu\right\} \\
&\qquad - \log\left\{\int_{-\infty}^{\infty} m_{\mu\,\to\,p(\boldsymbol{x}|\mu,\sigma^2)}(\mu) m_{p(\boldsymbol{x}|\mu,\sigma^2)\,\to\,\mu}(\mu)\,d\mu\right\} \\
&\qquad - \log\left\{\int_{0}^{\infty} m_{\sigma^2\,\to\,p(\boldsymbol{x}|\mu,\sigma^2)}(\sigma^2) m_{p(\boldsymbol{x}|\mu,\sigma^2)\,\to\,\sigma^2}(\sigma^2)\,d\sigma^2\right\}.
\end{aligned} \tag{A.16}$$

From arguments given in Appendix A.5.3

$$\begin{aligned}
&\int_0^{\infty} m_{\sigma^2\,\to\,p(\boldsymbol{x}|\mu,\sigma^2)}(\sigma^2)\; p(\boldsymbol{x}|\mu,\sigma^2) d\,\sigma^2 \\
&\qquad = (2\pi)^{-n/2} \exp\left\{A_{\text{IG}}\left(\begin{bmatrix} -n/2 \\ -\|\boldsymbol{x} - \mathbf{1}\mu\|^2/2 \end{bmatrix} + \boldsymbol{\eta}^{\#}\right)\right\}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\log\left\{\int_{-\infty}^{\infty}\int_{0}^{\infty} m_{\mu\,\to\,p(\boldsymbol{x}|\mu,\sigma^2)}(\mu)\, m_{\sigma^2\,\to\,p(\boldsymbol{x}|\mu,\sigma^2)}(\sigma^2)\; p(\boldsymbol{x}|\mu,\sigma^2)\,d\,\sigma^2\,d\,\mu\right\} \\
&\quad = -\frac{n}{2}\log(2\pi) - \log\left[\int_{-\infty}^{\infty} \exp\left\{\begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^T \boldsymbol{\eta}^{\star}\right\}\right. \\
&\qquad\qquad \left.\times \exp\left\{A_{\text{IG}}\left(\begin{bmatrix} -n/2 \\ -\|\boldsymbol{x} - \mathbf{1}\mu\|^2/2 \end{bmatrix} + \boldsymbol{\eta}^{\#}\right)\right\}\, d\mu\right] \\
&\quad = -\frac{n+1}{2}\log(2\pi) - \log\left\{\Gamma\left(\tfrac{n}{2} - \eta_1^{\#} - 1\right)\right\} \\
&\qquad + \log\left\{\alpha\left(\boldsymbol{\eta}^{\#}, \boldsymbol{\eta}^{\star}; \begin{bmatrix} n \\ \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^2 \end{bmatrix}\right)\right\}
\end{aligned}$$

where we have used algebra similar to that used in Appendix A.5.3. Substitution of this expression into (A.16) and simplification of the second and third terms, analogous to (A.14), then leads to

$$\begin{aligned}\log s_{p(\boldsymbol{x}|\mu,\sigma^2)} = &-\frac{n+1}{2}\log(2\pi) - \log\left\{\Gamma\left(\tfrac{n}{2} - \eta_1^{\#} - 1\right)\right\} \\ &+ \log\left\{\alpha\left(\boldsymbol{\eta}^{\#}, \boldsymbol{\eta}^{\star}; \begin{bmatrix} n \\ \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^2 \end{bmatrix}\right)\right\} \\ &- A_{\mathrm{N}}\left(\boldsymbol{\eta}^{\star} + \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\,\to\,\mu}\right) - A_{\mathrm{IG}}\left(\boldsymbol{\eta}^{\#} + \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\,\to\,\sigma^2}\right).\end{aligned}$$

The remaining term in (A.13) is

$$\begin{aligned}&\log s_{p(\sigma^2|\,a)} = \\ &\quad\log\left\{\int_0^\infty\int_0^\infty m_{\sigma^2\,\to\,p(\sigma^2|\,a)}(\sigma^2)\, m_{a\,\to\,p(\sigma^2|\,a)}(a)\; p(\sigma^2|\,a) d\,a\,d\,\sigma^2\right\} \\ &\quad-\log\left\{\int_0^\infty m_{\sigma^2\,\to\,p(\sigma^2|\,a)}(\sigma^2) m_{p(\sigma^2|\,a)\,\to\,\sigma^2}(\sigma^2)\, d\sigma^2\right\} \\ &\quad-\log\left\{\int_0^\infty m_{a\,\to\,p(\sigma^2|\,a)}(a) m_{p(\sigma^2|\,a)\,\to\,a}(a)\, da\right\}.\end{aligned} \tag{A.17}$$

Applying the algebraic steps used in Appendix A.5.4 to the first term of (A.17) results in

$$\begin{aligned}&\log\left\{\int_0^\infty\int_0^\infty m_{\sigma^2\,\to\,p(\sigma^2|\,a)}(\sigma^2)\, m_{a\,\to\,p(\sigma^2|\,a)}(a)\; p(\sigma^2|\,a) d\,a\,d\,\sigma^2\right\} = \\ &\qquad-\log\{\Gamma(-\eta_1^{\boxplus} - \tfrac{1}{2})\} + \log\{\mathcal{B}(0, \tfrac{1}{2} - \eta_1^{\otimes}, -\eta_2^{\otimes}, 0, -\eta_2^{\boxplus}, -\eta_1^{\boxplus} - \tfrac{1}{2})\}.\end{aligned}$$

Substitution into (A.17) and simplification of the second and third terms gives

$$\begin{aligned}\log s_{p(\sigma^2|\,a)} \;=\; &-\tfrac{1}{2}\log(\pi) - \log\{\Gamma(-\eta_1^{\boxplus} - \tfrac{1}{2})\} \\ &+ \log\{\mathcal{B}(0, \tfrac{1}{2} - \eta_1^{\otimes}, -\eta_2^{\otimes}, 0, -\eta_2^{\boxplus}, -\eta_1^{\boxplus} - \tfrac{1}{2})\} \\ &- A_{\mathrm{IG}}(\boldsymbol{\eta}^{\otimes} + \boldsymbol{\eta}_{p(\sigma^2|\,a)\,\to\,\sigma^2}) - A_{\mathrm{IG}}(\boldsymbol{\eta}^{\boxplus} + \boldsymbol{\eta}_{p(\sigma^2|\,a)\,\to\,a}).\end{aligned}$$

Adding each of the simplified expressions for the terms in (A.13) we obtain

$$\begin{aligned}\log\{\underset{\sim}{p}(\boldsymbol{x}; q)\} = &-\tfrac{1}{2}(n+2)\log(2\pi) - \tfrac{1}{2}\log(\pi) \\ &- \log\left\{\Gamma\left(\tfrac{n}{2} - \eta_{\sigma^2\,\to\,p(\boldsymbol{x}|\,\mu,\,\sigma^2),\,1} - 1\right)\right\} \\ &- \log\left\{\Gamma(-\eta_{a\,\to\,p(\sigma^2|\,a),\,1} - \tfrac{1}{2})\right\} \\ &+ A_{\mathrm{N}}\left(\boldsymbol{\eta}_{p(\mu)\,\to\,\mu} + \boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\,\to\,\mu}\right) \\ &+ A_{\mathrm{IG}}\left(\boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\,\to\,\sigma^2} + \boldsymbol{\eta}_{p(\sigma^2|\,a)\,\to\,\sigma^2}\right)\end{aligned}$$

$$
\begin{aligned}
&+A_{\text{IG}}\left(\boldsymbol{\eta}_{p(a)\,\to\,a}+\boldsymbol{\eta}_{p(\sigma^2|\,a)\,\to\,a}\right)\\
&-A_{\text{N}}\left(\boldsymbol{\eta}_{\mu\,\to\,p(\boldsymbol{x}|\mu,\sigma^2)}+\boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\,\to\,\mu}\right)\\
&-A_{\text{IG}}\left(\boldsymbol{\eta}_{\sigma^2\,\to\,p(\boldsymbol{x}|\mu,\sigma^2)}+\boldsymbol{\eta}_{p(\boldsymbol{x}|\mu,\sigma^2)\,\to\,\sigma^2}\right)\\
&-A_{\text{IG}}\left(\boldsymbol{\eta}_{a\,\to\,p(\sigma^2|\,a)}+\boldsymbol{\eta}_{p(\sigma^2|\,a)\,\to\,a}\right)\\
&-A_{\text{IG}}\left(\boldsymbol{\eta}_{\sigma^2\,\to\,p(\sigma^2|\,a)}+\boldsymbol{\eta}_{p(\sigma^2|\,a)\,\to\,\sigma^2}\right)\\
&+\log\left\{\alpha\left(\boldsymbol{\eta}_{\sigma^2\,\to\,p(\boldsymbol{x}|\mu,\sigma^2)},\boldsymbol{\eta}_{\mu\,\to\,p(\boldsymbol{x}|\mu,\sigma^2)};\begin{bmatrix}n\\ \sum_{i=1}^n x_i\\ \sum_{i=1}^n x_i^2\end{bmatrix}\right)\right\}\\
&+\log\{\mathcal{B}(0,\tfrac{1}{2}-\eta_{\sigma^2\,\to\,p(\sigma^2|\,a),\,1},-\eta_{\sigma^2\,\to\,p(\sigma^2|\,a),\,2},0,\\
&\qquad -\eta_{a\,\to\,p(\sigma^2|\,a),\,2},-\eta_{a\,\to\,p(\sigma^2|\,a),\,1}-\tfrac{1}{2})\}.
\end{aligned}
$$

Acknowledgements

This research was partially supported by the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers. The authors are grateful for comments from Tom Minka, the editor and the reviewers.

References

[1] Bakker, B. & Heskes, T. (2007), 'Learning and approximate inference in dynamic hierarchical models', *Computational Statistics and Data Analysis*, **52**, 821–839. MR2418531

[2] Bishop, C.M. (2006), *Pattern Recognition and Machine Learning*, Springer, New York. MR2247587

[3] Frey, B.J. & MacKay, D.J.C. (1998). 'A revolution: Belief propagation in graphs with cycles', In Jordan, M.I., Kearns, M.J., &, Solla S.A. (eds.) *Advances in Neural Information Processing Systems 10*, pp. 479–485.

[4] Guo, B.-N. & Qi, F. (2013), 'Refinements of lower bounds for polygamma functions', *Proceedings of the American Mathematical Society*, **141**, 1007–1015. MR3003692

[5] Luts, J., Broderick, T. & Wand, M.P. (2014), 'Real-time semiparametric regression', *Journal of Computational and Graphical Statistics*, **23**, 589–615. MR3224647

[6] Maybeck, P.S. (1982), *Stochastic Models, Estimation and Control*, Academic Press, New York.

[7] Minka, T.P. (2001), 'Expectation propagation for approximate Bayesian inference'. In Breese, J.S. & Koller, D. (eds) *Proceedings of the Sev-*

enteenth Conference on Uncertainty in Artificial Intelligence, pp. 362–369. Morgan Kaufmann, Burlington, Massachusetts.

[8] Minka, T. (2005), 'Divergence measures and message passing', *Microsoft Research Technical Report Series*, **MSR-TR-2005-173**, 1–17.

[9] Minka, T. & Winn, J. (2008), 'Gates: A graphical notation for mixture models', *Microsoft Research Technical Report Series*, **MSR-TR-2008-185**, 1–16.

[10] Minka, T., Winn, J., Guiver, J. & Knowles, D. (2013), Infer.NET 2.5, http://research.microsoft.com/infernet.

[11] R Development Core Team (2015), 'R: A language and environment for statistical computing', R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org

[12] Smyth, G. (2013), 'statmod 1.4. Statistical modeling', http://cran.r-project.org

[13] Stan Development Team (2016), 'rstan 2.9: R interface to Stan', Version 2.9, http://mc-stan.org

[14] Wainwright, M.J. & Jordan, M.I. (2008), 'Graphical models, exponential families, and variational inference', *Foundations and Trends in Machine Learning*, **1**, 1–305.

[15] Wand, M.P. & Jones, M.C. (1993), 'Comparison of smoothing parameterizations in bivariate kernel density estimation', *Journal of the American Statistical Association*, **88**, 520–528. MR1224377

[16] Wand M.P. & Ripley, B.D. (2010), 'KernSmooth 2.23. Functions for kernel smoothing for Wand & Jones (1995)', http://cran.r-project.org

[17] Wand, M.P., Ormerod, J.T., Padoan, S.A. & Frühwirth, R. (2011), 'Mean field variational Bayes for elaborate distributions', *Bayesian Analysis*, **6**, 847–900. MR2869967

[18] Winn, J. & Bishop, C.M. (2005), 'Variational message passing', *Journal of Machine Learning Research*, **6**, 661–694. MR2249835

[19] Zoeter, O. & Heskes, T. (2005), 'Gaussian quadrature based expectation propagation', In Ghahramani, Z. and Cowell, R. (eds.) *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics* pp. 445–452. The Society for Artificial Intelligence and Statistics.