Model fitting and Bayesian inference via power expectation propagation

Stima ed inferenza Bayesiana tramite power expectation propagation

Emanuele Degani, Luca Maestrini and Mauro Bernardi

Abstract We study a message passing approach to power expectation propagation for Bayesian model fitting and inference. Power expectation propagation is a class of variational approximations based on the notion of α-divergence that extends two notable approximations, namely mean field variational Bayes and expectation propagation. An illustration on a simple model allows to grasp benefits and complexities of this methodology and sets the basis for applications on more complex models.
Abstract *Studiamo l'approccio message passing al power expectation propagation per la stima e l'inferenza Bayesiana. Power expectation propagation è una classe di approssimazioni variazionali basata sulla nozione di divergenza α che estende due approssimazioni notevoli, mean field variational Bayes ed expectation propagation. Un'illustrazione su un semplice modello consente di cogliere benefici e complessità di questa metodologia, ponendo le basi per applicazioni su modelli più complessi.*

Key words: α-divergence, approximate Bayesian inference, factor graph, message passing, variational approximation.

1 Introduction

Bayesian inference deals with updating a prior distribution $p(\theta)$ on a parameter vector θ through the model likelihood $p(\mathbf{y}|\theta)$ for the observed data $\mathbf{y}$ to obtain the posterior distribution $p(\theta|\mathbf{y}) = p(\mathbf{y}|\theta)p(\theta)/p(\mathbf{y})$. Typically the marginal likelihood $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)\mathrm{d}\theta$ cannot be evaluated explicitly and Markov chain Monte Carlo (MCMC) methods have been the main toolkit to sample from the posterior

Emanuele Degani, Mauro Bernardi
Department of Statistical Sciences, University of Padua, Italy
e-mail: degani@stat.unipd.it, e-mail: mauro.bernardi@unipd.it,

Luca Maestrini
School of Mathematical and Physical Sciences, University of Technology Sydney, Australia
e-mail: luca.maestrini@uts.edu.au

density for decades. Nevertheless, MCMC algorithms may suffer of slow convergence and poor mixing behaviors that can compromise inferential conclusions [4].

Variational inference methods [3, 11] take a different perspective on the problem. Instead of sampling from $p(\theta|\mathbf{y})$, variational approaches are used to approximate the posterior density with an approximating density $q(\theta)$ chosen from a suitable family $\mathscr{Q}$ of distributions. The most common Bayesian variational methods find the optimal approximating density by solving

$$q^*(\theta) = \underset{q(\theta)\in\mathscr{Q}}{\arg\min}\, \mathrm{KL}(q(\theta)\|p(\theta|\mathbf{y})), \tag{1}$$

with $\mathrm{KL}(q(\theta)\|p(\theta|\mathbf{y}))$ denoting the Kullback–Leibler divergence between q and $p(\cdot|\mathbf{y})$. Practical solutions arise imposing a convenient partition $\{\theta_1,\ldots,\theta_M\}$ of θ such that $q(\theta)=\prod_{i=1}^{M} q(\theta_i)$ and employing a convex optimization scheme (see e.g. Section 10.1.1. of [2]) known as *mean field variational Bayes (MFVB)*.

Another variational inference technique, proposed in [8] and named *expectation propagation (EP)*, is built upon the optimization problem

$$q^*(\theta) = \underset{q(\theta)\in\mathscr{Q}}{\arg\min}\, \mathrm{KL}(p(\theta|\mathbf{y})\|q(\theta)), \tag{2}$$

where the arguments of the Kullback–Leibler divergence in (1) are reversed. This leads to a different class of iterative optimization schemes that [10] recasts into a *message passing* on a factor graph framework. The message passing paradigm allows for distributed and scalable fitting of variational approximations. [5] exploit the results of [10] and provide an explicit algorithm for performing EP on a simple statistical model, studying issues and challenges related to its implementation.

In this article we study a generalization of both MFVB and EP known as *power expectation propagation (Power-EP)* that was proposed by [9] to make (2) more tractable. This method yields a class of appealing message passing algorithms and we explore their use for statistical model fitting. Section 2 describes Power-EP and introduces a message passing technique to solve the optimization problem on models with factor graph representations. Section 3 provides explicit illustration on a simple model and Section 4 investigates the quality of the variational approximation via a simulation study. Final considerations and further developments are described in Section 5.

2 Power-EP and message passing

Power-EP solves the following optimization problem:

$$q^*_\alpha(\theta) = \underset{q_\alpha(\theta)\in\mathscr{Q}}{\arg\min}\, \mathrm{D}_\alpha(p(\theta|\mathbf{y})\|q_\alpha(\theta)), \qquad \alpha\in(-\infty,\infty)\backslash\{0\}\,,$$

where $\mathrm{D}_\alpha(p(\theta|\mathbf{y})\|q_\alpha(\theta)) \equiv (\alpha(1-\alpha))^{-1}\{1-\int_\Theta p(\theta|\mathbf{y})^\alpha q_\alpha(\theta)^{1-\alpha}\mathrm{d}\theta\}$ is the α-divergence of Amari [1]. It possesses two notable limiting cases:

$$\mathrm{D}_\alpha(p(\theta|\mathbf{y})\|q(\theta)) \overset{\alpha\to 0}{\Longrightarrow} \mathrm{KL}(q(\theta)\|p(\theta|\mathbf{y})) \text{ and } \mathrm{D}_1(p(\theta|\mathbf{y}) = \mathrm{KL}(p(\theta|\mathbf{y})\|q(\theta)),$$

meaning that Power-EP reduces to MFVB and EP for $\alpha \to 0$ and $\alpha = 1$, respectively. Hence, the quality of Power-EP approximations varies with α, and for certain α values the approximations may outperform those obtained with MFVB and EP. We restrict our attention to approximations arising from $\alpha \in (0,1]$, that is to the class of approximations that has MFVB and EP as extreme and opposite cases.

[10] provides an approximate solution to the minimization in (2) based on message passing on factor graphs, for a given α. We employ this strategy and describe a message passing procedure for fitting models having a factor graph representation via Power-EP (see e.g. [6, §2.3] for a primer on factor graphs).

Consider a model whose joint density function can be factorized into N different factors $p(\theta, \mathbf{y}) = \prod_{j=1}^{N} f_j(\theta_{\text{neigh(j)}})$, with $\text{neigh(j)} \equiv \{1 \le i \le M : \theta_i \text{ is a neighbor of } f_j\}$. Introduce an approximating density to the posterior distribution $q_\alpha(\theta)$ that can be written as $q_\alpha(\theta) = \prod_{i=1}^{M} q_\alpha(\theta_i)$. Using a Power-EP approach, each density $q_\alpha(\theta_i)$ can be obtained as the product of *messages* reaching θ_i from the neighboring factors. For each $1 \le i \le M$ and $1 \le j \le N$, the Power-EP *factor to stochastic node message* updates are given by

$$m^{(\alpha)}_{f_j \to \theta_i}(\theta_i) \longleftarrow \text{proj}\Big\{ Z^{-1} [m^{(\alpha)}_{f_j \to \theta_i}(\theta_i)]^{1-\alpha} m^{(\alpha)}_{\theta_i \to f_j}(\theta_i) \int [f_j(\theta_{\text{neigh}(j)})]^{\alpha} \times \prod_{i' \in \text{neigh}(j)/\{i\}} [m^{(\alpha)}_{f_j \to \theta_{i'}}(\theta_{i'})]^{1-\alpha} m^{(\alpha)}_{\theta_{i'} \to f_j}(\theta_{i'}) \, \mathrm{d}\theta_{\text{neigh}(j)/\{i\}} \Big\} \Big/ m^{(\alpha)}_{\theta_i \to f_j}(\theta_i) \,, \tag{3}$$

where the $\longleftarrow$ symbol means that the function of θ_i on the left-hand side is updated according to the expression on the right-hand side, $\text{proj}\{p\}$ is the operator that projects the density function p onto an appropriate exponential family (see [5, §2.3]) and Z is the normalizing constant of p. For each $1 \le i \le M$ and $1 \le j \le N$, the Power-EP *stochastic node to factor message* updates have form

$$m^{(\alpha)}_{\theta_i \to f_j}(\theta_i) \longleftarrow \prod_{j' \ne j \,:\, i \in \text{neigh}(j')} m^{(\alpha)}_{f_{j'} \to \theta_i}(\theta_i) \,. \tag{4}$$

Optimization can be performed by iteratively updating the factor graph messages via (3) and (4) upon convergence. Convergence can be assessed by monitoring the *α-approximate marginal log-likelihood* defined as

$$\log \tilde{p}(\mathbf{y}; q_\alpha) \equiv \sum_{i=1}^{M} \log s^{(\alpha)}_{\theta_i} + \frac{1}{\alpha} \sum_{j=1}^{N} \log s^{(\alpha)}_{f_j}, \text{ with } s^{(\alpha)}_{\theta_i} \equiv \int \prod_{j : i \in \text{neigh}(j)} m^{(\alpha)}_{f_j \to \theta_i}(\theta_i) \, \mathrm{d}\theta_i$$
$$\text{and } s^{(\alpha)}_{f_j} \equiv \frac{\int \big(f_j(\theta_{\text{neigh}(j)})\big)^{\alpha} \prod_{i \in \text{neigh}(j)} m^{(\alpha)}_{\theta_i \to f_j}(\theta_i) \Big(m^{(\alpha)}_{f_j \to \theta_i}(\theta_i)\Big)^{1-\alpha} \mathrm{d}\theta_{\text{neigh}(j)}}{\int \prod_{i \in \text{neigh}(j)} m^{(\alpha)}_{\theta_i \to f_j}(\theta_i) \, m^{(\alpha)}_{f_j \to \theta_i}(\theta_i) \, \mathrm{d}\theta_{\text{neigh}(j)}} . \tag{5}$$

At convergence, the optimal approximating densities can be obtained from

$$q^{*}_{\alpha}(\theta_i) \propto \prod_{j \,:\, i \in \text{neigh}(j)} m^{(\alpha)}_{f_j \to \theta_i}(\theta_i) = m^{(\alpha)}_{\theta_i \to f_j}(\theta_i) \, m^{(\alpha)}_{f_j \to \theta_i}(\theta_i) \,. \tag{6}$$

It is worth noting that when $\alpha = 1$, the resulting $q_1^*(\theta)$ approximation matches the one from EP. Consequently, expressions (3.5)–(3.11) of [5], and results of [6] can be immediately retrieved fixing $\alpha = 1$ in expressions (3)–(6).

3 Simple illustrative example

The general expressions of Section 2 providing a message passing solution to Power-EP are anything but intuitive and the computational steps behind (3)–(6) are difficult to glean. Therefore, we make explicit illustration on the simple Bayesian Normal random sample model studied in [5]. The model we consider is:

$$y_i \overset{\text{i.i.d.}}{\sim} \text{N}(\mu, \sigma^2),\ \mu \sim \text{N}(\mu_\mu, \sigma_\mu^2),\ \sigma^2|a \sim \text{Inv-Gamma}\Big(\frac{1}{2}, \frac{1}{a}\Big),\ a \sim \text{Inv-Gamma}\Big(\frac{1}{2}, \frac{1}{A^2}\Big), \quad (7)$$

for $1 \leq i \leq n$, where $\mu_\mu \in \mathbb{R}$, $\sigma_\mu > 0$ and $A > 0$ are fixed hyperparameters, and the hierarchical specification on σ^2 is such that $\sigma \sim \text{Half-Cauchy}(A)$. The joint density function then factorizes as $p(\mathbf{y}, \mu, \sigma^2, a) = p(\mathbf{y}|\mu, \sigma^2)p(\mu)p(\sigma^2|a)p(a)$.

Consider the approximation $q_\alpha(\mu, \sigma^2, a) = q_\alpha(\mu)q_\alpha(\sigma^2)q_\alpha(a)$ to the posterior density. Application of (3) and enforcement of conjugacy constraints give rise to the following expressions for the Power-EP factor to stochastic node messages:

$$m^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\to\mu}(\mu) \propto \exp\left(\begin{bmatrix}\mu \\ \mu^2\end{bmatrix}^T \eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\to\mu}\right),\ m^{(\alpha)}_{p(\sigma^2|a)\to a}(a) \propto \exp\left(\begin{bmatrix}\log a \\ 1/a\end{bmatrix}^T \eta^{(\alpha)}_{p(\sigma^2|a)\to a}\right),$$
$$m^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\to\sigma^2}(\sigma^2) \propto \exp\left(\begin{bmatrix}\log\sigma^2 \\ 1/\sigma^2\end{bmatrix}^T \eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\to\sigma^2}\right),\ m^{(\alpha)}_{p(\mu)\to\mu}(\mu) \propto \exp\left(\begin{bmatrix}\mu \\ \mu^2\end{bmatrix}^T \eta^{(\alpha)}_{p(\mu)\to\mu}\right),$$
$$m^{(\alpha)}_{p(\sigma^2|a)\to\sigma^2}(\sigma^2) \propto \exp\left(\begin{bmatrix}\log\sigma^2 \\ 1/\sigma^2\end{bmatrix}^T \eta^{(\alpha)}_{p(\sigma^2|a)\to\sigma^2}\right),\ m^{(\alpha)}_{p(a)\to a}(a) \propto \exp\left(\begin{bmatrix}\log a \\ 1/a\end{bmatrix}^T \eta^{(\alpha)}_{p(a)\to a}\right).$$

Here the symbol η denotes natural parameter vectors of exponential families. Straightforward application of (4) leads to similar and conjugate expressions for the Power-EP stochastic node to factor messages. Application of (6) leads to the optimal approximating densities for the parameters of interest $q^*_\alpha(\mu)$ and $q^*_\alpha(\sigma^2)$:

$$q^*_\alpha(\mu) \propto \exp\left(\begin{bmatrix}\mu \\ \mu^2\end{bmatrix}^T \eta_{q^*_\alpha(\mu)}\right) \quad \text{and} \quad q^*_\alpha(\sigma^2) \propto \exp\left(\begin{bmatrix}\log\sigma^2 \\ 1/\sigma^2\end{bmatrix}^T \eta_{q^*_\alpha(\sigma^2)}\right), \quad (8)$$

which correspond to a $\text{N}(-[\eta_{q^*_\alpha(\mu)}]_1/(2[\eta_{q^*_\alpha(\mu)}]_2), -1/(2[\eta_{q^*_\alpha(\mu)}]_2))$ density function for μ and an $\text{Inv-Gamma}(-[\eta_{q^*_\alpha(\sigma^2)}]_1 - 1, -[\eta_{q^*_\alpha(\sigma^2)}]_2)$ density function for σ^2, respectively, with $\eta_{q^*_\alpha(\mu)}$ and $\eta_{q^*_\alpha(\sigma^2)}$ vectors of length 2.

Given that the resulting Power-EP messages belong to exponential families, their updates can be performed just by updating their η natural parameter vectors. Derivations of these updates and the explicit expression of $\log \tilde{p}(\mathbf{y}; q_\alpha)$ from (5) are not provided here for brevity, but follow steps similar to those in [5, §A.5].

Algorithm 1 lists the iterative natural parameter updates for fitting the Bayesian random sample model via Power-EP message passing. Within the expressions of the natural parameter updates, $\eta^{(\alpha)}_{f_j\leftrightarrow\theta_i} \equiv (1-\alpha)\eta^{(\alpha)}_{f_j\to\theta_i} + \eta^{(\alpha)}_{\theta_i\to f_j}$ for meaningful com-

binations of $f_j \in \{p(\mu), p(\mathbf{y}|\mu,\sigma^2), p(\sigma^2|a), p(a)\}$ and $\theta_i \in \{\mu, \sigma^2, a\}$. Functions $G^N(\cdot)$, $G^{IG1}(\cdot)$ and $G^{IG2}(\cdot)$ are defined in [5, §A.4] and involve quadrature methods for evaluating non-analytic functions that are described in [5, §2.1].

Algorithm 1 *Power-Expectation Propagation message passing algorithm for determining the parameter of the optimal density functions $q^*_\alpha(\mu)$ and $q^*_\alpha(\sigma^2)$ of interest for approximate Bayesian inference on the Normal random sample model* (7).

Input: $\mathbf{y} = (y_1, \ldots, y_n)^T$, μ_μ, $\sigma_\mu > 0$ and $A > 0$. Create: $\mathbf{c} = (n, \sum_{i=1}^n y_i, \sum_{i=1}^n y_i^2)^T$.
Select: Power-EP factor $\alpha \in (0, 1]$.
Initialize: $\eta^{(\alpha)}_{p(\mu)\to\mu} \leftarrow \begin{bmatrix} \mu_\mu/\sigma^2_\mu \\ -1/(2\sigma^2_\mu) \end{bmatrix}$, $\eta^{(\alpha)}_{p(a)\to a} \leftarrow \begin{bmatrix} -3/2 \\ -1/A^2 \end{bmatrix}$, $\eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\to\mu} \leftarrow \begin{bmatrix} 0 \\ -1/2 \end{bmatrix}$,
$\eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\to\sigma^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}$, $\eta^{(\alpha)}_{p(\sigma^2|a)\to\sigma^2} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}$, $\eta^{(\alpha)}_{p(\sigma^2|a)\to a} \leftarrow \begin{bmatrix} -2 \\ -1 \end{bmatrix}$,
$\eta^{(\alpha)}_{\mu\to p(\mathbf{y}|\mu,\sigma^2)} \leftarrow \eta^{(\alpha)}_{p(\mu)\to\mu}$, $\eta^{(\alpha)}_{a\to p(\sigma^2|a)} \leftarrow \eta^{(\alpha)}_{p(a)\to a}$.
Cycle until the relative change in $\log \tilde{p}(\mathbf{y}; q_\alpha)$ is negligible:

$$\eta^{(\alpha)}_{\sigma^2\to p(\mathbf{y}|\mu,\sigma^2)} \leftarrow \eta^{(\alpha)}_{p(\sigma^2|a)\to a},$$
$$\eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\to\mu} \leftarrow G^N\left(\eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\leftrightarrow\mu}, \eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\leftrightarrow\sigma^2}; \alpha\mathbf{c}\right) + (1-\alpha)\eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\to\mu},$$
$$\eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\to\sigma^2} \leftarrow G^{IG1}\left(\eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\leftrightarrow\sigma^2}, \eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\leftrightarrow\mu}; \alpha\mathbf{c}\right) + (1-\alpha)\eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\to\sigma^2},$$
$$\eta^{(\alpha)}_{\sigma^2\to p(\sigma^2|a)} \leftarrow \eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\to\sigma^2},$$
$$\eta^{(\alpha)}_{p(\sigma^2|a)\to\sigma^2} \leftarrow G^{IG2}\left(\eta^{(\alpha)}_{p(\sigma^2|a)\leftrightarrow\sigma^2}, \begin{bmatrix} [\eta^{(\alpha)}_{p(\sigma^2|a)\leftrightarrow a}]_1 + 2(1-\alpha) \\ [\eta^{(\alpha)}_{p(\sigma^2|a)\leftrightarrow a}]_2/\alpha \end{bmatrix}; 3\alpha\right) + (1-\alpha)\eta^{(\alpha)}_{p(\sigma^2|a)\to\sigma^2},$$
$$\eta^{(\alpha)}_{p(\sigma^2|a)\to a} \leftarrow G^{IG2}\left(\eta^{(\alpha)}_{p(\sigma^2|a)\leftrightarrow a}, \begin{bmatrix} [\eta^{(\alpha)}_{p(\sigma^2|a)\leftrightarrow\sigma^2}]_1 + 2(1-\alpha) \\ [\eta^{(\alpha)}_{p(\sigma^2|a)\leftrightarrow\sigma^2}]_2/\alpha \end{bmatrix}; \alpha\right) + (1-\alpha)\eta^{(\alpha)}_{p(\sigma^2|a)\to a}.$$

Output for (8): $\eta_{q^*_\alpha(\mu)} = \eta^{(\alpha)}_{p(\mu)\to\mu} + \eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\to\mu}$, $\eta_{q^*_\alpha(\sigma^2)} = \eta^{(\alpha)}_{p(\mathbf{y}|\mu,\sigma^2)\to\sigma^2} + \eta^{(\alpha)}_{p(\sigma^2|a)\to\sigma^2}$.

4 Simulation study

We assess the performances of Power-EP for fitting model (7) through a simulation study. For each sample size $n \in \{25, 50, 100, 500, 1000\}$, we generate 100 random samples from the $\mathrm{N}(0,1)$ distribution and obtain the optimal Power-EP approximating densities of interest $q^*_\alpha(\mu)$ and $q^*_\alpha(\sigma^2)$ for $\alpha \in \{0.25, 0.5, 0.75, 1\}$ via Algorithm 1, and *MFVB* approximations using Algorithm 1 of [7]. We set diffuse priors with hyperparameters $\mu_\mu = 0$ and $\sigma_\mu = A = 10^5$. For each replicate, we evaluate the quality of the approximation computing, for $\theta = \mu, \sigma^2$, accuracy$\{q^*_\alpha(\theta)\} \equiv 100\,(1 - 0.5 \int |q^*_\alpha(\theta) - p(\theta|\mathbf{y})| d\theta)$. The 'true' marginal posterior densities are obtained via kernel density estimation applied to MCMC samples obtained with the `rstan` library [12], after excluding an appropriate burn-in sample. Figure 1 summarizes the results and compares the approximations. For small sample sizes, Power-EP approximations with $\alpha = 0.25, 0.5, 0.75$ overperform both EP and MFVB in terms of accuracy for μ, whereas EP provides a better approxi-

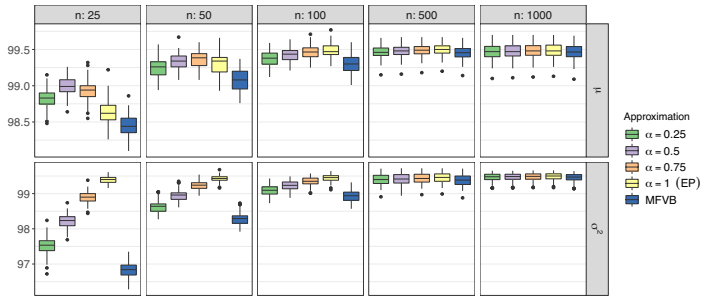

Fig. 1 Accuracy values of the approximating q^*'s for μ and σ^2, at different sample sizes.

mation for σ^2. As n increases, the accuracy of the approximations becomes more uniform for both μ and σ^2.

5 Conclusions and further developments

We studied Power-EP as a message passing approach for fitting models that have a factor graph representation through the minimization of the α-divergence between the posterior and an approximating density. Power-EP includes the more common MFVB and EP approximations, which can be outperformed by approximations based on appropriate choice of α, especially when the number of observations is limited. Implementation of Power-EP for a wide set of α values comes with a higher computational cost, that could be reduced applying optimization strategies based on automatic differentiation. Further directions include the exploration of methods for automatic selection of α values that produce better approximations and application to more complex statistical models.

References

1. Amari, S.: Differential-Geometrical Methods in Statistics. Springer, New York (1985)
2. Bishop, C. M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
3. Blei, M. D., Kucukelbir, A., McAuliffe, J. D.: Variational Inference: A Review for Statisticians. J. Am. Stat. Assoc. **112**, 859–877 (2017)
4. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B.: Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC, (2013)
5. Kim, A. S. I., Wand, M. P.: The Explicit Form of Expectation Propagation for a Simple Statistical Model. Electron. J. Stat. **10**, 550–581 (2016)
6. Kim, A. S. I., Wand, M. P.: On Expectation Propagation for Generalised, Linear and Mixed Models. Aust. N. Z. J. Stat. **60**, 75–102 (2018)
7. Luts, J., Broderick, T., Wand, M. P.: Real-time semiparametric regression. J. Comput. Graph. Stat. **23**, 589–615 (2014)
8. Minka, T. P.: Expectation propagation for approximate Bayesian inference. Proc. of the XVII Conf. on Uncert. in Art. Intel., Morgan Kaufmann Publishers Inc., 362–239 (2001)
9. Minka, T. P.: Power EP. Micr. Res. Tech. Rep. **149** (2004)
10. Minka, T. P.: Divergence measures and message passing. Micr. Res. Tech. Rep. **173** (2005)
11. Ormerod, J. T., Wand, M. P.: Explaining variational approximations. Am. Stat. **64**, 140–153 (2010)
12. Stan Development Team.: `rstan` 2.21.2: the R interface to Stan. http://mc-stan.org/ (2020)