

PAPER • OPEN ACCESS

## Expectation propagation for Poisson data

To cite this article: Chen Zhang *et al* 2019 *Inverse Problems* **35** 085006

View the [article online](#) for updates and enhancements.



**IOP | ebooks**<sup>TM</sup>

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the [collection](#) - download the first chapter of every title for free.

# Expectation propagation for Poisson data

Chen Zhang, Simon Arridge<sup>1</sup> and Bangti Jin<sup>1</sup>

Department of Computer Science, University College London, London WC1E 6BT, United Kingdom

E-mail: [chen.zhang.16@ucl.ac.uk](mailto:chen.zhang.16@ucl.ac.uk), [s.arridge@ucl.ac.uk](mailto:s.arridge@ucl.ac.uk) and [b.jin@ucl.ac.uk](mailto:b.jin@ucl.ac.uk)

Received 18 October 2018, revised 19 February 2019

Accepted for publication 3 April 2019

Published 30 July 2019



CrossMark

## Abstract

The Poisson distribution arises naturally when dealing with data involving counts, and it has found many applications in inverse problems and imaging. In this work, we develop an approximate Bayesian inference technique based on expectation propagation for approximating the posterior distribution formed from the Poisson likelihood function and a Laplace type prior distribution, e.g. the anisotropic total variation prior. The approach iteratively yields a Gaussian approximation, and at each iteration, it updates the Gaussian approximation to one factor of the posterior distribution by moment matching. We derive explicit update formulas in terms of one-dimensional integrals, and also discuss stable and efficient quadrature rules for evaluating these integrals. The method is showcased on two-dimensional PET images.

Keywords: Poisson distribution, Laplace prior, expectation propagation, approximate Bayesian inference

(Some figures may appear in colour only in the online journal)

## 1. Introduction

The Poisson distribution is widely employed to describe inverse and imaging problems involving count data, e.g. emission computed tomography [40, 44], including positron emission tomography and single photon emission computed tomography. The corresponding likelihood function is a Poisson distribution with its parameter given by an affine transform (followed by a suitable link function). Over the past few decades, the mathematical theory and numerical algorithms for image reconstruction with Poisson data have witnessed impressive progresses. We refer interested readers to [22] for a comprehensive overview on variational regularization

<sup>1</sup> Author to whom any correspondence should be addressed.



Original content from this work may be used under the terms of the [Creative Commons Attribution 3.0 licence](https://creativecommons.org/licenses/by/3.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

techniques for Poisson data and [4] for mathematical modeling and numerical methods for Poisson data. It is worth noting that the Poisson model is especially important in the low-count regime, e.g. [0, 10] photons, whereas in the moderate (or high) count regime, heteroscedastic normal approximations can be employed in the reconstruction, leading to a weighted Gaussian likelihood function (e.g. via the so-called Anscombe transform [2]). In this work, we focus on the Poisson model.

To cope with the inherent ill-posed nature of the imaging problem, regularization plays an important role in image reconstruction. This can be achieved implicitly via early stopping during an iterative reconstruction procedure (e.g. EM algorithm or Richardson–Lucy iterations) or explicitly via suitable penalties, e.g. Sobolev penalty, sparsity and total variation. The penalized maximum likelihood (or equivalently maximum *a posteriori* (MAP)) is currently the most popular way for image reconstruction with Poisson models [12, 41]. However, these approaches can only provide point estimates, and the important issue of uncertainty quantification, which provides crucial reliability assessment on point estimates, is not fully addressed. The full Bayesian approach provides a principled yet very flexible framework for uncertainty quantification of inverse and imaging problems [27, 42]. The prior distribution acts as a regularizer, and the ill-posedness of the imaging problem is naturally dealt with. Due to the imprecise prior knowledge of the solution and the presence of the data noise, the posterior distribution contains an ensemble of inverse solutions consistent with the observed data, which can be used to quantify the uncertainties associated with a point estimator, via, e.g. credible interval or highest probability density regions.

For imaging problems with Poisson data, a full Bayesian treatment is challenging, due to the nonnegativity constraint and high-dimensionality of the parameter/data space. There are several possible strategies from the computational perspective. One idea is to use general-purposed sampling methods to explore the posterior state space, predominantly Markov chain Monte Carlo (MCMC) methods [32, 37]. Recent scalable variants, e.g. stochastic gradient Langevin dynamics [47], are very promising, although these techniques have not been applied to the Poisson model. Then the constraints on the signal can be incorporated directly by discarding samples violating the constraint. However, in order to obtain accurate statistical estimates, sampling methods generally require many samples and thus tend to suffer from high computational cost, due to the high problem dimensionality. Further, the MCMC convergence is challenging to diagnose. These observations have motivated intensive research works on developing approximate inference techniques (AITs). In the machine learning literature, a large number of AITs have been proposed, e.g. variational inference [3, 6, 9, 24, 26], expectation propagation [33, 34] and more recently Bayesian (deep) neural network [17]; see the survey [48] for a comprehensive overview of recent developments on variational inference. In all AITs, one aims at finding an optimal approximate yet tractable distribution within a family of parametric/nonparametric probability distributions (e.g. Gaussian), by minimizing the error in a certain probability metric, prominently Kullback–Leibler divergence. Empirically they can often produce reasonable approximations but at a much reduced computational cost than MCMC. However, there seem no systematic strategies for handling constraints in these approaches. For example, a straightforward truncation of the distribution due to the constraint often leads to elaborated distributions, e.g. truncated normal distribution, which tends to make the computation tedious or even completely intractable in variational Bayesian inference.

In this work, we develop a computational strategy for exploring the posterior distribution for Poisson data (with two popular nonnegativity constraints) with a Laplace type prior based on expectation propagation [33, 34], in order to deliver a Gaussian approximation. Laplace prior promotes the sparsity of the image in a transformed domain, which is a valid assumption on most natural images. The main contributions of the work are as follows. First, we derive

explicit update formulas using one-dimensional integrals. It essentially exploits the rank-one projection form of the factors to reduce the intractable high-dimensional integrals to tractable one-dimensional ones. In this way, we arrive at two approximate inference algorithms, parameterized by either moment or natural parameters. Second, we derive stable and efficient quadrature rules for evaluating the resulting one-dimensional integrals, i.e. a recursive scheme for Poisson sites with large counts and an approximate expansion for Laplace sites, and discuss different schemes for the recursion, dependent of the integration interval, in order to achieve good numerical stability. Last, we illustrate the approach with comprehensive numerical experiments with the posterior distribution formed by Poisson likelihood and an anisotropic total variation prior, clearly showcasing the feasibility of the approach.

Last, we put the work in the context of Bayesian analysis of Poisson data. The predominant body of literature in statistics employs a log link function, commonly known as Poisson regression in statistics and machine learning (see, e.g. [3, 8]). This differs substantially from the one frequently arising in medical imaging, e.g. positron emission tomography, and in particular the crucial nonnegativity constraint becomes vacuous. The only directly relevant work we are aware of is the recent work [28]. The work [28] discussed a full Bayesian exploration with EP, by modifying the posterior distributions using a rectified linear function on the transformed domain of the signal, which induces singular measures on the region violating the constraint. However, the work [28] does not consider the background.

The rest of the paper is organized as follows. In section 2 we describe the posterior distribution for the Poisson likelihood function and a Laplace type prior. Then we give explicit expressions of the integrals involved in EP update and describe two algorithms in section 3. In section 4 we present stable and efficient numerical methods for evaluating one-dimensional integrals. Last, in section 5 we present numerical results for three benchmark images. In the appendices, we describe two useful parameterizations of a Gaussian distribution, Laplace approximation and additional comparative numerical results for a one-dimensional problem with MCMC and Laplace approximation to shed further insights into the performance of EP algorithms.

## 2. Problem formulation

In this part, we give the Bayesian formulation for Poisson data, i.e. the likelihood function  $p(y|x)$  and prior distribution  $p(x)$ , and discuss the nonnegativity constraint.

Let  $x \in \mathbb{R}^n$  be the (unknown) signal/image of interest,  $y \in \mathbb{R}_+^{m_1}$  be the observed Poisson data, and  $A = [a_{ij}] = [a_{ij}^t]_{i=1}^{m_1} \in \mathbb{R}_+^{m_1 \times n}$  be the forward map, where the superscript  $t$  denotes matrix/vector transpose. The entries of the matrix  $A$  are assumed to be nonnegative. For example, in emission computed tomography, it can be a discrete analogue of Radon transform, or probabilistically, the entry  $a_{ij}$  of the matrix  $A$  denotes the probability that the  $i$ th sensor pair records the photon emitted at the  $j$ th site.

The conditional probability density  $p(y_i|x)$  of observing  $y_i \in \mathbb{N}$  given the signal  $x$  is given by

$$p(y_i|x) = \frac{(a_i^t x + r_i)^{y_i} e^{-(a_i^t x + r_i)}}{y_i!},$$

where  $r = [r_i]_i \in \mathbb{R}_+^{m_1}$  is the background. That is, the entry  $y_i$  follows a Poisson distribution with a parameter  $a_i^t x + r_i$ . The Poisson model of this form is popular in the statistical modeling of inverse and imaging problems involving counts, e.g. positron emission tomography [44]. If

the entries of  $y$  are independent and identically distributed (i.i.d.), then the likelihood function  $p(y|x)$  is given by

$$p(y|x) = \prod_{i=1}^{m_1} p(y_i|x).$$

Note that the likelihood function  $p(y|x)$  is not well-defined for all  $x \in \mathbb{R}^n$ , and suitable constraints on  $x$  are needed in order to ensure the well-definedness of the factors  $p(y_i|x)$ 's. In the literature, there are three popular constraints:

1.  $\mathcal{C}_1 = \{x|x > 0\} := \cap_i \{x|x_i > 0\}$ ;
2.  $\mathcal{C}_2 = \{x|Ax > 0\} := \cap_i \{x|[Ax]_i = a_i^t x > 0\}$ ;
3.  $\mathcal{C}_3 = \{x|Ax + r > 0\} := \cap_i \{x|[Ax + r]_i = a_i^t x + r_i > 0\}$ .

Since the entries of  $A$  are nonnegative, there holds  $\mathcal{C}_1 \subset \mathcal{C}_2 \subset \mathcal{C}_3$ . In practice, the first assumption is most consistent with the physics in that it reflects the physical constraint that emission counts are non-negative. The last two assumptions were proposed to reduce positive bias in the cold region [30], i.e. the region that has zero count. In this work, we shall focus on the last two constraints.

The constraints  $\mathcal{C}_2$  and  $\mathcal{C}_3$  can be unified, which is useful for the discussions below.

**Definition 2.1.** For each likelihood factor  $p(y_i|x)$  with the constraint  $\mathcal{C}_2$ , let

$$V_i^+ = \{x|[Ax]_i = a_i^t x > 0\} \quad \text{and} \quad V_i^- = \mathbb{R}^n \setminus V_i^+.$$

For each likelihood factor  $p(y_i|x)$  with the constraint  $\mathcal{C}_3$ , let

$$V_i^+ = \{x|[Ax + r]_i = a_i^t x + r_i > 0\} \quad \text{and} \quad V_i^- = \mathbb{R}^n \setminus V_i^+.$$

Then the constraints  $\mathcal{C}_2$  and  $\mathcal{C}_3$  are both given by  $V^+ = \cap_i V_i^+$  and  $V^- = \mathbb{R}^n \setminus V^+$ .

With the indicator function  $\mathbf{1}_{V^+}(x)$  of the set  $V^+$ , we modify the likelihood function  $p(y|x)$  by

$$\ell(x) = p(y|x)\mathbf{1}_{V^+}(x).$$

This extends the domain of  $p(y|x)$  from  $V^+$  to  $\mathbb{R}^n$ , and it facilitates a full Bayesian treatment. Since the indicator function  $\mathbf{1}_{V^+}(x)$  admits a separable form, i.e.  $\mathbf{1}_{V^+}(x) = \prod_{i=1}^{m_1} \mathbf{1}_{V_i^+}(x)$ ,  $\ell(x)$  factorizes into

$$\ell(x) = \prod_{i=1}^{m_1} \ell_i(x) \quad \text{with} \quad \ell_i(x) = p(y_i|x)\mathbf{1}_{V_i^+}(x).$$

To fully specify the Bayesian model, we have to stipulate the prior  $p(x)$ . We focus on a Laplace type prior. Let  $L \in \mathbb{R}^{m_2 \times n}$  and  $L_i^t \in \mathbb{R}^{n \times 1}$  be the  $i$ th row of  $L$ . Then a Laplace type prior  $p(x)$  is given by

$$p(x) = \prod_{i=1}^{m_2} p_i(x) \quad \text{with} \quad p_i(x) = \frac{\alpha}{2} e^{-\alpha|L_i^t x|}.$$

The parameter  $\alpha > 0$  determines the strength of the prior, playing the role of a regularization parameter in variational regularization [23]. The choice of the hyperparameter  $\alpha$  in the prior  $p(x)$  is notoriously challenging [23]. One may apply hierarchical Bayesian modeling in order to estimate it from the data simultaneously with the unknown  $x$  [3, 25, 46]. The prior  $p(x)$  is commonly known as a sparsity prior (in the transformed domain), which favors a candidate

with many small elements and few large elements in the vector  $Lx$ . The canonical total variation prior is recovered when the matrix  $L$  computes the discrete gradient. It is well known that the total variation penalty can preserve well edges in the image/signals, and hence it has been very popular for various imaging tasks [10, 38].

By Bayes' formula, we obtain the Bayesian solution to the Poisson inverse problem, i.e. the posterior probability density function:

$$p(x|y) = Z^{-1} \prod_{i=1}^{m_1} \ell_i(x) \prod_{i=1}^{m_2} p_i(x), \quad (2.1)$$

where  $Z$  is the normalizing constant, defined by  $Z = \int_{\mathbb{R}^n} \prod_{i=1}^{m_1} \ell_i(x) \prod_{i=1}^{m_2} p_i(x) dx$ . The computation of  $Z$  is generally intractable for high-dimensional problems, and  $p(x|y)$  has to be approximated.

### 3. Approximate inference by expectation propagation

In this section, we describe the basic concepts and algorithms of expectation propagation (EP), for exploring the posterior distribution (2.1). EP due to Minka [33, 34] is a popular variational type approximate inference method in the machine learning literature. It is especially suitable for approximating a distribution formed by a product of functions, with each factor being of projection form. Since its first appearance in 2001, EP has found many successful applications in practice, and it is reported to be very accurate, e.g. for Gaussian processes [36], and electrical impedance tomography with sparsity prior [19]. However, the theoretical understanding of EP remains quite limited [13, 14].

EP looks for an approximate Gaussian distribution  $q(x)$  to a target distribution by means of an iterative algorithm. It relies on the following factorization of the posterior distribution  $p(x|y)$  (with  $m = m_1 + m_2$  being the total number of factors):

$$p(x|y) = Z^{-1} \prod_{i=1}^m t_i(x), \quad \text{with } t_i(x) = \begin{cases} \ell_i(x), & i = 1, \dots, m_1, \\ p_{i-m_1}(x), & i = m_1 + 1, \dots, m. \end{cases} \quad (3.1)$$

Note that each factor  $t_i(x)$  is a function defined on the whole space  $\mathbb{R}^n$ . Likewise, we denote the Gaussian approximation  $q(x)$  to the posterior distribution  $p(x|y)$  by

$$q(x) = \tilde{Z}^{-1} \prod_{i=1}^m \tilde{t}_i(x),$$

with each factor  $\tilde{t}_i(x)$  being a Gaussian distribution  $\mathcal{N}(x|\mu_i, C_i)$ , and  $\tilde{Z}$  is the corresponding normalizing constant. Below we use two different parameterizations of a Gaussian distribution, i.e. moment parameters (mean and covariance)  $(\mu, C)$  and natural parameters  $(h, \Lambda)$ ; see appendix A. Both parameterizations have their pros and cons: the moment one does not require solving linear systems, and the natural one allows singular covariances for the Gaussians  $\tilde{t}_i(x)$ . The rest of this section is devoted to the derivation of the algorithms and their complexity.

#### 3.1. Reduction to one-dimensional integrals

There are two main steps of one EP iteration: (a) form a tilted distribution  $\hat{q}_i(x)$ , and (b) update the Gaussian approximation  $q(x)$  by matching its moments with that of  $\hat{q}_i(x)$ . The moment matching step can be interpreted as minimizing Kullback–Leibler divergence  $\text{KL}(\hat{q}_i||q)$  [19,

33, 34]. Recall that the Kullback–Leibler divergence from one probability distribution  $p(x)$  to another  $q(x)$  is defined by [29]

$$D_{\text{KL}}(p||q) = \int_{\mathbb{R}^n} p(x) \log \frac{p(x)}{q(x)} dx.$$

By Jensen's inequality, the divergence  $D_{\text{KL}}(p||q)$  is always nonnegative, and it vanishes if and only if  $p(x) = q(x)$  almost everywhere.

The task at step (a) is to construct the  $i$ th tilted distribution  $\hat{q}_i(x)$ . Let  $q_{\setminus i}(x)$  be the  $i$ th cavity distribution, i.e. the product of all but the  $i$ th factor, and defined by

$$q_{\setminus i}(x) = Z_i^{-1} \prod_{j \neq i} \tilde{t}_j(x)$$

with  $Z_i = \int_{\mathbb{R}^n} \prod_{j \neq i} \tilde{t}_j(x) dx$ . It is Gaussian, i.e.  $q_{\setminus i}(x) = \mathcal{N}(x|\mu_{\setminus i}, C_{\setminus i})$ , whose moment and natural parameters are denoted by  $(\mu_{\setminus i}, C_{\setminus i})$  and  $(h_{\setminus i}, \Lambda_{\setminus i})$ , respectively. Then the  $i$ th tilted distribution  $\hat{q}_i(x)$  of the approximation  $q(x)$  is given by

$$\hat{q}_i(x) = \hat{Z}_i^{-1} t_i(x) \prod_{j \neq i} \tilde{t}_j(x),$$

where  $\hat{Z}_i = \int_{\mathbb{R}^n} t_i(x) \prod_{j \neq i} \tilde{t}_j(x) dx$  is the corresponding normalizing constant. With the exclusion-inclusion step, one replaces the  $i$ th factor  $\tilde{t}_i(x)$  in the approximation  $q$  with the exact one  $t_i(x)$ .

The task at step (b) is to compute moments of the  $i$ th tilde distribution  $\hat{q}_i(x)$ , which are then used to update the approximation  $q(x)$ . This requires integration over  $\mathbb{R}^n$ , which is generally numerically intractable, if  $\hat{q}_i(x)$  were arbitrary. Fortunately, each factor  $t_i(x)$  in (3.1) is of projection form and depends only on the scalar  $u^t x$ , with the vector  $u \in \mathbb{R}^n$  being either  $a_i$  or  $L_i$ . This is the key fact rendering relevant high-dimensional integrals numerically tractable. Below we write the factor  $t_i(x)$  as  $t_i(u^t x)$  and accordingly, the  $i$ th cavity function  $\hat{q}_i(x)$  as

$$\hat{q}_i(x) = \hat{Z}_i^{-1} t_i(u^t x) \mathcal{N}(x|\mu_{\setminus i}, C_{\setminus i}),$$

upon replacing  $\prod_{j \neq i} \tilde{t}_j(x)$  with its normalized version  $\mathcal{N}(x|\mu_{\setminus i}, C_{\setminus i})$ , and accordingly the normalizing constant  $\hat{Z}_i$ .

Since a Gaussian is determined by its mean and covariance, it suffices to evaluate the 0th to 2nd moments of  $\hat{q}_i(x)$ . The projection form of the factor  $t_i$  allows reducing the moment evaluation of  $\hat{q}_i(x)$  to 1D integrals. Theorem 3.1 gives the update scheme for  $q(x)$  from  $\hat{q}_i(x)$ .

**Theorem 3.1.** *The normalizing constant  $\hat{Z}_i := \int_{\mathbb{R}^n} t_i(u^t x) \mathcal{N}(x|\mu_{\setminus i}, C_{\setminus i}) dx$  is given by*

$$\hat{Z}_i = \int_{\mathbb{R}} t_i(s) \mathcal{N}(s|u_i^t \mu_{\setminus i}, u_i^t C_{\setminus i} u_i) ds =: Z_s$$

Then with the auxiliary variables  $\bar{s} \in \mathbb{R}$  and  $C_s$  defined by

$$\bar{s} = Z_s^{-1} \int_{\mathbb{R}} t_i(s) \mathcal{N}(s|u_i^t \mu_{\setminus i}, u_i^t C_{\setminus i} u_i) s ds \quad \text{and} \quad C_s = Z_s^{-1} \int_{\mathbb{R}} t_i(s) \mathcal{N}(s|u_i^t \mu_{\setminus i}, u_i^t C_{\setminus i} u_i) s^2 ds - \bar{s}^2, \quad (3.2)$$

the mean  $\mu = \mathbb{E}_{\hat{q}_i}[x]$  and covariance  $C = \mathbb{V}_{\hat{q}_i}[x]$  are given respectively by

$$\begin{aligned} \mu &= \mu_{\setminus i} + C_{\setminus i} u_i (u_i^t C_{\setminus i} u_i)^{-1} (\bar{s} - u_i^t \mu_{\setminus i}), \\ C &= C_{\setminus i} + (u_i^t C_{\setminus i} u_i)^{-2} (C_s - u_i^t C_{\setminus i} u_i) C_{\setminus i} u_i u_i^t C_{\setminus i}. \end{aligned}$$

Similarly, the precision mean  $h_{\hat{q}_i}$  and precision  $\Lambda_{\hat{q}_i}$  are given respectively by

$$\begin{aligned} h_{\hat{q}_i} &= h_{\setminus i} + \lambda_{1,i} u_i \quad \text{with } \lambda_{1,i} = \frac{\bar{s}}{C_s} - \frac{u_i^t \mu_{\setminus i}}{u_i^t C_{\setminus i} u_i}, \\ \Lambda_{\hat{q}_i} &= \Lambda_{\setminus i} + \lambda_{2,i} u_i u_i^t \quad \text{with } \lambda_{2,i} = \frac{1}{C_s} - \frac{1}{u_i^t C_{\setminus i} u_i}. \end{aligned}$$

**Proof.** The expressions for  $\hat{Z}_i$ ,  $\mu$  and  $C$  were given in [19, section 3]. Thus it suffices to derive the formulas for  $(h, \Lambda)$ . Recall the Sherman–Morrison formula [21, p 65]: for any invertible  $B \in \mathbb{R}^{n \times n}$ ,  $u, v \in \mathbb{R}^n$ , there holds

$$(B + uv^t)^{-1} = B^{-1} - \frac{B^{-1}uv^tB^{-1}}{1 + v^tB^{-1}u}. \quad (3.3)$$

Let  $\lambda = (u_i^t C_{\setminus i} u_i)^{-2} (C_s - u_i^t C_{\setminus i} u_i)$ . Then the precision matrix  $\Lambda$  is given by

$$\begin{aligned} \Lambda &= (C_{\setminus i} + C_{\setminus i} u_i \lambda u_i^t C_{\setminus i})^{-1} \\ &= C_{\setminus i}^{-1} - u_i (\lambda^{-1} + u_i^t C_{\setminus i} u_i)^{-1} u_i^t \\ &= \Lambda_{\setminus i} + \left( \frac{1}{C_s} - \frac{1}{u_i^t C_{\setminus i} u_i} \right) u_i u_i^t. \end{aligned}$$

Similarly, the precision mean  $h := \Lambda \mu$  is given by

$$\begin{aligned} h &= \left[ \Lambda_{\setminus i} + \left( \frac{1}{C_s} - \frac{1}{u_i^t C_{\setminus i} u_i} \right) u_i u_i^t \right] [\mu_{\setminus i} + C_{\setminus i} u_i (u_i^t C_{\setminus i} u_i)^{-1} (\bar{s} - u_i^t \mu_{\setminus i})] \\ &= \Lambda_{\setminus i} \mu_{\setminus i} + u_i \left( \frac{\bar{s}}{C_s} - \frac{u_i^t \mu_{\setminus i}}{u_i^t C_{\setminus i} u_i} \right) = h_{\setminus i} + u_i \left( \frac{\bar{s}}{C_s} - \frac{u_i^t \mu_{\setminus i}}{u_i^t C_{\setminus i} u_i} \right). \end{aligned}$$

This completes the proof of the theorem.  $\square$

In both approaches, the 1D integrals  $(Z_s, \bar{s}, C_s)$  are needed, which depend on  $u_i^t \mu_{\setminus i}$  and  $u_i^t C_{\setminus i} u_i$ . A direct approach is first to downdate (the Cholesky factor of)  $\Lambda$  and then to solve a linear system. In practice, this can be expensive and the cost can be mitigated. Indeed, they can be computed without the downdating step; see lemma 3.1 below. Below we use the super- or subscript  $n$  and  $o$  to denote a variable updated at current iteration from that of the last iteration.

**Lemma 3.1.** Let  $c = u_i^t \Lambda_o^{-1} u_i = u_i^t C_o u_i$ ,  $(h, \Lambda)$  be the natural parameter of  $q(x)$  and  $(\lambda_{1,i}, \lambda_{2,i})$  be defined in theorem 3.1. Then the mean  $u_i^t \mu_{\setminus i}$  and variance  $u_i^t C_{\setminus i} u_i$  of the Gaussian distribution  $\mathcal{N}(s | u_i^t \mu_{\setminus i}, u_i^t C_{\setminus i} u_i)$  are respectively given by

$$u_i^t \mu_{\setminus i} = \frac{u_i^t \Lambda_o^{-1} h - c \lambda_{1,i}^o}{1 - c \lambda_{2,i}^o} \quad \text{and} \quad u_i^t C_{\setminus i} u_i = \frac{c}{1 - c \lambda_{2,i}^o}.$$

**Proof.** We suppress the sub/superscript  $o$ . By the definition of  $u_i^t C_{\setminus i} u_i$  and the Sherman–Morrison formula (3.3), we have

$$\begin{aligned} u_i^t C_{\setminus i} u_i &= u_i^t (\Lambda - \lambda_{2,i} u_i u_i^t)^{-1} u_i \\ &= u_i^t [\Lambda^{-1} - \Lambda^{-1} u_i (-\lambda_{2,i}^{-1} + c)^{-1} u_i^t \Lambda^{-1}] u_i \\ &= c - c (-\lambda_{2,i}^{-1} + c)^{-1} c = \frac{c}{1 - c \lambda_{2,i}}, \end{aligned}$$

and similarly, we have

$$\begin{aligned} u_i^t \mu_{\setminus i} &= u_i^t (\Lambda - \lambda_{2,i} u_i u_i^t)^{-1} (h - \lambda_{1,i} u_i) \\ &= u_i^t [\Lambda^{-1} - \Lambda^{-1} u_i (-\lambda_{2,i}^{-1} + c)^{-1} u_i^t \Lambda^{-1}] (h - \lambda_{1,i} u_i) = \frac{u_i^t \Lambda^{-1} h - c \lambda_{1,i}}{1 - c \lambda_{2,i}}. \end{aligned}$$

This completes the proof of the lemma.  $\square$

Since the quantities for the 1D integrals can be calculated from variables updated in the last iteration, it is unnecessary to form cavity distributions. Indeed, the cavity precision is formed by  $\Lambda_{\setminus i} = \Lambda_o - \lambda_{2,i}^o u_i u_i^t$ , and the updated precision is given by  $\Lambda_n = \Lambda_{\setminus i} + \lambda_{2,i}^n u_i u_i^t$ ; and similarly for  $h$ . Thus, we can update  $\Lambda$  directly with  $(\lambda_{2,i}^o, \lambda_{2,i}^n)$  and  $h$  with  $(\lambda_{1,i}^o, \lambda_{1,i}^n)$ ; this is summarized in the next remark.

**Remark 3.1.** The differences  $\lambda_{k,i}^n - \lambda_{k,i}^o$ ,  $k = 1, 2$ , can be used to update the natural parameter  $(h, \Lambda)$ :

$$\lambda_{1,i}^n - \lambda_{1,i}^o = \frac{\bar{s}}{C_s} - \frac{u_i^t \mu_o}{u_i^t C_o u_i} \quad \text{and} \quad \lambda_{2,i}^n - \lambda_{2,i}^o = \frac{1}{C_s} - \frac{1}{u_i^t C_o u_i}.$$

Moreover, the sign of  $\lambda_{2,i}^n - \lambda_{2,i}^o$  determines whether to update or downdate the Cholesky factor of  $\Lambda$ .

### 3.2. Update schemes and algorithms

Now we state the direct update scheme, i.e. without explicitly constructing the intermediate cavity distribution  $q_{\setminus i}(x)$ , for both natural and moment parameterizations.

**Theorem 3.2.** Let  $(h, \Lambda)$  and  $(\mu, C)$  be the natural and moment parameters of the Gaussian approximation  $q(x)$ , respectively. The following update schemes hold.

(i) The precision mean  $h$  and precision  $\Lambda$  can be updated by

$$h_n = h_o + \left( \frac{\bar{s}}{C_s} - \frac{u_i^t \Lambda_o^{-1} h_o}{u_i^t \Lambda_o^{-1} u_i} \right) u_i \quad \text{and} \quad \Lambda_n = \Lambda_o + \left( \frac{1}{C_s} - \frac{1}{u_i^t \Lambda_o^{-1} u_i} \right) u_i u_i^t.$$

(ii) The mean  $\mu$  and covariance  $C$  can be updated by

$$\mu_n = \mu_o + \frac{\bar{s} - u_i^t \mu_o}{u_i^t C_o u_i} C_o u_i \quad \text{and} \quad C_n = C_o + \left( \frac{C_s}{(u_i^t C_o u_i)^2} - \frac{1}{u_i^t C_o u_i} \right) (C_o u_i)(u_i^t C_o).$$

**Proof.** The first assertion is direct from theorem 3.1 and remark 3.1, and it can be rewritten as

$$\Lambda_n = \Lambda_o + (\lambda_{2,i}^n - \lambda_{2,i}^o) u_i u_i^t \quad \text{and} \quad h_n = h_o + (\lambda_{1,i}^n - \lambda_{1,i}^o) u_i.$$

By Sherman–Morrison formula (3.3), the covariance  $C_n = \Lambda_n^{-1}$  is given by

$$\begin{aligned}
C_n &= (\Lambda_o + (\lambda_{2,i}^n - \lambda_{2,i}^o)u_i u_i^t)^{-1} \\
&= \Lambda_o^{-1} - \Lambda_o^{-1}u_i \left( \frac{1}{\lambda_{2,i}^n - \lambda_{2,i}^o} + u_i^t C_o u_i \right)^{-1} u_i^t \Lambda_o^{-1} \\
&=: C_o + \eta_2(C_o u_i)(u_i^t C_o),
\end{aligned}$$

where the scalar  $\eta_2 := -(\frac{1}{\lambda_{2,i}^n - \lambda_{2,i}^o} + u_i^t C_o u_i)^{-1}$  can be simplified to

$$\eta_2 = -\frac{\lambda_{2,i}^n - \lambda_{2,i}^o}{1 + (\lambda_{2,i}^n - \lambda_{2,i}^o)u_i^t C_o u_i} = -\frac{1}{u_i^t C_o u_i} + \frac{C_s}{(u_i^t C_o u_i)^2},$$

where the second identity follows from remark 3.1. Similarly, the mean  $\mu_n := C_n h_n$  is given by

$$\begin{aligned}
\mu_n &= [C_o + \eta_2(C_o u_i)(u_i^t C_o)] [h_o + (\lambda_{1,i}^n - \lambda_{1,i}^o)u_i] \\
&= \mu_o + (\lambda_{1,i}^n - \lambda_{1,i}^o)C_o u_i + \eta_2 u_i^t \mu_o C_o u_i + \eta_2 (\lambda_{1,i}^n - \lambda_{1,i}^o)u_i^t C_o u_i C_o u_i =: \mu_o + \eta_1 C_o u_i,
\end{aligned}$$

where, in view of remark 3.1,  $\eta_1 := (\lambda_{1,i}^n - \lambda_{1,i}^o) + \eta_2 u_i^t \mu_o + \eta_2 (\lambda_{1,i}^n - \lambda_{1,i}^o)u_i^t C_o u_i$  can be simplified to

$$\eta_1 = \frac{(\lambda_{1,i}^n - \lambda_{1,i}^o) - (\lambda_{2,i}^n - \lambda_{2,i}^o)u_i^t \mu_o}{1 + (\lambda_{2,i}^n - \lambda_{2,i}^o)u_i^t C_o u_i} = \frac{\bar{s} - u_i^t \mu_o}{u_i^t C_o u_i}.$$

This completes the proof of the theorem.  $\square$

All matrix operations in theorem 3.2 are of rank one type, which can be implemented stably and efficiently with the Cholesky factors and their update/downdate; see section 3.3 for details. Thus, in practice, we employ Cholesky factors of the precision  $\Lambda$  and covariance  $C$ , denoted by  $\Lambda_{\text{chol}}$  and  $C_{\text{chol}}$ , respectively, instead of  $\Lambda$  and  $C$ . Further, we also use the auxiliary variables  $(\lambda_{1,i}, \lambda_{2,i})$  defined in theorem 3.1, and stack  $\{(\lambda_{1,i}, \lambda_{2,i})\}_{i=1}^{m_1+m_2}$  into two vectors

$$\lambda_1 = [\lambda_{1,i}]_i, \quad \lambda_2 = [\lambda_{2,i}]_i \in \mathbb{R}^{m_1+m_2},$$

which are initialized to zeros. Thus, we obtain two inference procedures for Poisson data with a Laplace type prior in algorithms 1 and 2.

The rigorous convergence analysis of EP is outstanding. Nonetheless, empirically, it often converges very fast, which is also observed in our numerical experiments in section 5. In practice, one can terminate the iteration by monitoring the relative change of the parameters or fixing the maximum number  $K$  of iterations. The important task of computing 1D integrals will be discussed in section 4 below.

---

**Algorithm 1.** Expectation propagation for Poisson data (natural parametrization).

---

- 1: Input:  $(A, y)$ , hyper-parameter  $\alpha$ , and maximum number  $K$  of iterations
  - 2: Initialize  $h$ ,  $\Lambda_{\text{chol}}$ ,  $\lambda_1$  and  $\lambda_2$ ;
  - 3: **for**  $k = 1, 2, \dots, K$  **do**
  - 4: Randomly choose an index  $i$  to update;
  - 5: Compute the mean and variance for 1D Gaussian integral by lemma 3.1;
  - 6: Evaluate  $\bar{s}$  and  $C_s$  in (3.2);
  - 7: Calculate and update  $\lambda_{1,i}$  and  $\lambda_{2,i}$ ;
  - 8: Update  $h$  and  $\Lambda_{\text{chol}}$  by theorem 3.2;
  - 9: Check the stopping criterion.
  - 10: **end for**
  - 11: Output:  $(h, \Lambda_{\text{chol}})$
-

### 3.3. Efficient implementation and complexity estimate

The rank-one matrix update  $A \pm \beta uu^t$ , for  $A \in \mathbb{R}^{n \times n}$ ,  $u \in \mathbb{R}^n$  and  $\beta > 0$ , can be stably and efficiently updated/downdated with the Cholesky factor of  $A$  with  $\sqrt{\beta}u$ . The update step of  $A$  can be viewed as an iteration from  $A_k$  to  $A_{k+1}$ . Let the upper triangular matrices  $R_k$  and  $R_{k+1}$  be the Cholesky factors of  $A_k$  and  $A_{k+1}$  respectively, i.e.  $A_k = R_k^t R_k$  and  $A_{k+1} = R_{k+1}^t R_{k+1}$ . There are two possible cases:

- (i) If  $A_{k+1} = A_k + \beta uu^t$ ,  $R_{k+1}$  is the Cholesky rank one update of  $R_k$  with  $\sqrt{\beta}u$ .
- (ii) If  $A_{k+1} = A_k - \beta uu^t$ ,  $R_{k+1}$  is the Cholesky rank one downdate of  $R_k$  with  $\sqrt{\beta}u$ .

The update/downdate is available in several packages. For example, in MATLAB, the function `cholupdate` implements the update/downdate of Cholesky factors, based on LAPACK subroutines ZCHUD and ZCHDD.

Next, we discuss the computational complexity per inner iteration. The first step picks one index  $i$ , which is of constant complexity. For the second step, i.e. computing the mean and variance for 1D integrals, the dominant part is linear solve involving upper triangular matrices and matrix-vector product for natural and moment parameters. For either parameterization, it incurs  $\mathcal{O}(n^2)$  operations. The third step computes  $\bar{s}$  and  $C_s$  from the one dimensional integrals. For Poisson site, the complexity is  $\mathcal{O}(y_i)$ , and for Laplace site, it is  $\mathcal{O}(1)$ . Last, the fourth step is dominated by Cholesky factor modifications, and its complexity is  $\mathcal{O}(n^2)$ . Overall, the computational complexity per inner iteration is  $\mathcal{O}(n^2 + y_i)$ . In a large data setting,  $y_i \ll n$ , and thus the complexity is about  $\mathcal{O}(n^2)$ .

In passing, we note that in practice, the covariance/precision matrix may admit additional structures, e.g. sparsity, which translate into structures on Cholesky factors. For the general sparsity assumption, it seems unclear how to effectively exploit it for Cholesky update/downdate for enhanced efficiency, except the diagonal case, which can be incorporated into the algorithm straightforwardly.

## 4. Stable evaluation of 1d integrals

Now we develop a stable implementation for the three 1D integrals:  $Z_s$ ,  $\bar{s}$  and  $C_s$  in theorem 3.1. These integrals form the basic components of algorithms 1 and 2, and their stable, accurate and efficient evaluation is crucial to the performance of the algorithms. By suppressing the subscript  $i$ , we can write the integrals in a unified way:

$$J_j = \int_{\mathbb{R}} t(s) \mathcal{N}(s|m, \sigma^2) s^j ds, \quad j = 0, 1, 2,$$

where the factor  $t(s)$  is either Poisson likelihood or Laplace prior. Then we can express  $\bar{s}$  and  $C_s$  in terms of  $J_j$  by

$$\bar{s} = \frac{J_1}{J_0} \quad \text{and} \quad C_s = \frac{J_2}{J_0} - \bar{s}^2.$$

Note that the normalizing constants in  $J_j$  cancel out in  $\bar{s}$  and  $C_s$ , and thus they can be ignored when evaluating the integrals. In essence, the computation boils down to stable evaluation of moments of a (truncated) Gaussian distribution. This task was studied in several works [11, 39]: [11] focuses on Gaussian moments, and [39] discusses also evaluating the integrals involving Laplace distributions. Below we derive the formulas for the (constrained) Poisson likelihood and Laplace prior separately.

**Algorithm 2.** Expectation propagation for Poisson data (moment parametrization).

---

1: Input:  $(A, y)$ , hyper-parameter  $\alpha$ , and maximum number  $K$  of iterations  
2: Initialize  $\mu$ ,  $C_{\text{chol}}$ ,  $\lambda_1$  and  $\lambda_2$ ;  
3: **for**  $k = 1, 2, \dots, K$  **do**  
4:   Randomly choose an index  $i$  to update;  
5:   Compute the mean and variance for 1D Gaussian integral by lemma 3.1;  
6:   Evaluate  $\bar{s}$  and  $C_s$  in (3.2);  
7:   Calculate and update  $\lambda_{1,i}$  and  $\lambda_{2,i}$ ;  
8:   Update  $\mu$  and  $C_{\text{chol}}$  by theorem 3.2;  
9:   Check the stopping criterion.  
10: **end for**  
11: Output:  $(\mu, C_{\text{chol}})$

---

**4.1. Poisson likelihood**

Throughout, we suppress the subscript  $i$ , write  $V_+$  etc in place of  $V_i^+$  etc and introduce the scalar variable  $s = a^t x$ . Then the constraint on  $x$  transfers to that on  $s$ :  $a^t x > 0$  corresponds to  $s > 0$  and  $a^t x + r > 0$  to  $s > -r$ , respectively. We shall slightly abuse the notation and use  $\mathbf{1}_{V_+}(s)$  as the indicator for the constraint on  $s$ . Then the Poisson likelihood  $t(x)$  can be equivalently written in either  $x$  or  $s$  as

$$t(x) = \frac{(a^t x + r)^y e^{-(a^t x + r)}}{y!} \mathbf{1}_{V_+}(x) \quad \text{and} \quad t(s) = \frac{(s + r)^y e^{-(s+r)}}{y!} \mathbf{1}_{V_+}(s).$$

Note that the factorial  $y!$  cancels out when computing  $\bar{s}$  and  $C_s$ , so it is omitted in the derivation below. For a fixed  $\mathcal{N}(s|m, \sigma^2)$ , the integrals  $J_{y,j}$  depend on the observed count data  $y$  and moment order  $j$ :

$$J_{y,j} = \int_b^\infty (s + r)^y s^j e^{-(s+r)} \mathcal{N}(s|m, \sigma^2) ds.$$

where the lower integral bound  $b = 0$  or  $b = -r$ , which is evident from the context. Note that the terms  $e^{-(s+r)}$  and  $\mathcal{N}(s|m, \sigma^2)$  in  $J_{y,j}$  together give an unnormalized Gaussian density. This allows us to reduce the integrals  $J_{y,j}$  into (truncated) Gaussian moment evaluations of the type:

$$I_y = \int_b^\infty (s + r)^y \mathcal{N}(s|m - \sigma^2, \sigma^2) ds,$$

and accordingly  $\bar{s}$  and  $C_s$ . This is given in the next result.

**Theorem 4.1.** *The scalars  $\bar{s}$  and  $C_s$  can be computed by*

$$\bar{s} = \frac{I_{y+1}}{I_y} - r \quad \text{and} \quad C_s = \frac{I_{y+2}}{I_y} - \left( \frac{I_{y+1}}{I_y} \right)^2.$$

**Proof.** First, we claim that with  $\alpha = e^{\frac{\sigma^2}{2} - m - r}$ , there hold the following identities

$$J_{y,0} = \alpha I_y, \quad J_{y,1} = \alpha(I_{y+1} - r I_y), \quad \text{and} \quad J_{y,2} = \alpha(I_{y+2} - 2r I_{y+1} + r^2 I_y). \quad (4.1)$$

Let  $c_\sigma = (2\pi\sigma^2)^{-\frac{1}{2}}$ . Then by completing the square, we obtain

$$e^{-(s+r)} \mathcal{N}(s|m, \sigma^2) = c_\sigma e^{-r-s-\frac{(s-m)^2}{2\sigma^2}} = c_\sigma e^{\frac{\sigma^2}{2} - m - r} e^{-\frac{(s-(m-\sigma^2))^2}{2\sigma^2}}.$$

The claim follows directly from the trivial identities

$$\begin{aligned}(s+r)^y s &= (s+r)^{y+1} - r(s+r)^y, \\ (s+r)^y s^2 &= (s+r)^{y+2} - 2r(s+r)^{y+1} + r^2(s+r)^y.\end{aligned}$$

The desired identities follow from the definitions and the recursions in (4.1) by

$$\begin{aligned}\bar{s} &= \frac{J_{y,1}}{J_{y,0}} = \frac{\alpha(I_{y+1} - rI_y)}{\alpha I_y} = \frac{I_{y+1}}{I_y} - r, \\ C_s &= \frac{J_{y,2}}{J_{y,0}} - \bar{s}^2 = \frac{\alpha(I_{y+2} - 2rI_{y+1} + r^2I_y)}{\alpha I_y} - \left(\frac{I_{y+1}}{I_y} - r\right)^2 = \frac{I_{y+2}}{I_y} - \left(\frac{I_{y+1}}{I_y}\right)^2.\end{aligned}$$

This completes the proof.  $\square$

However, directly evaluating  $I_y$  can still be numerically unstable for large  $y$ . To avoid the potential instability, we develop a stable recursive scheme on  $I_y$ .

**Lemma 4.1.** For  $y \geq 2$ , the following recursion holds

$$I_y = (m - \sigma^2 + r)I_{y-1} + \sigma^2(y-1)I_{y-2} + \frac{\sigma^2(b+r)^{y-1}}{\sqrt{2\pi\sigma^2}} e^{-\frac{(b-m+\sigma^2)^2}{2\sigma^2}}.$$

**Proof.** Let  $c = m - \sigma^2$ ,  $d = \sigma^2$  and  $f(s) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s-c)^2}{2d}}$ . The definition of  $I_y$  implies

$$\begin{aligned}I_y &= \int_b^\infty (s+r)^y f(s) ds = \int_b^\infty (s+r)^{y-1} \left(d \frac{s-c}{d} + c+r\right) f(s) ds \\ &= -d \int_b^\infty (s+r)^{y-1} \left(-\frac{s-c}{d}\right) f(s) ds + (c+r) \int_b^\infty (s+r)^{y-1} f(s) ds.\end{aligned}$$

Next we employ the trivial identity  $\frac{d}{ds} f(s) = -\frac{s-c}{d} f(s)$  and apply integration by parts to the first term

$$\begin{aligned}&\int_b^\infty (s+r)^{y-1} \left(-\frac{s-c}{d}\right) f(s) ds \\ &= (s+r)^{y-1} f(s) \Big|_b^\infty - \int_b^\infty (y-1)(s+r)^{y-2} f(s) ds \\ &= -(b+r)^{y-1} f(b) - (y-1)I_{y-2}.\end{aligned}$$

Collecting the terms shows the desired recursion on the integral  $I_y$ .  $\square$

For  $b = -r$ , we have a simplified recursive formula for  $I_y$ :

$$I_y = (m - \sigma^2 + r)I_{y-1} + \sigma^2(y-1)I_{y-2}.$$

Lemma 4.1 uses a two-term linear recurrence relation for  $I_y$ 's. The coefficients of  $I_{y-1}$  and  $I_{y-2}$  are raised by power when expanding  $I_y$  in terms of  $I_0$  and  $I_1$ , and thus the computation of  $I_y$  is susceptible to the evaluation errors of  $I_0$  and  $I_1$  for large  $y$ . This motivates a reciprocal recursive scheme by introducing a ratio sequence  $\{L_y\}_y$  defined by  $L_y = \frac{yI_{y-1}}{I_y}$ , for  $r = 0$  or  $b = -r$ , in order to restore the numerical stability. Note that  $L_y$  also admits a recursive scheme  $L_y = \frac{y}{(m-\sigma^2+r)+\sigma^2 L_{y-1}}$ , and further  $I_y$  can be recovered from  $\{L_y\}$  by  $\ln I_y = \ln y! + \ln I_0 - \sum_{i=1}^y L_i$ .

**Table 1.** Three schemes for evaluating  $I_0$ ,  $I_1$  and  $I_2$ , with  $c_1 = m - \sigma^2 + b + 2r$  and  $c_2 = m - \sigma^2 + r$ .

Scheme	Formulae	$\eta$
1	$I_0 = \frac{1}{2}(1 - \operatorname{erf}(\eta)), \quad I_1 = \sqrt{\frac{\sigma^2}{2\pi}}e^{-\eta^2} + \frac{c_2}{2}(1 - \operatorname{erf}(\eta))$ $I_2 = \sqrt{\frac{\sigma^2}{2\pi}}c_1e^{-\eta^2} + \frac{c_2 + \sigma^2}{2}(1 - \operatorname{erf}(\eta))$	$(-\infty, 5)$
2	$I_0 = \frac{1}{2}\operatorname{erfc}(\eta), \quad I_1 = \sqrt{\frac{\sigma^2}{2\pi}}e^{-\eta^2} + \frac{c_2}{2}\operatorname{erfc}(\eta)$ $I_2 = \sqrt{\frac{\sigma^2}{2\pi}}c_1e^{-\eta^2} + \frac{c_2 + \sigma^2}{2}\operatorname{erfc}(\eta)$	$[5, 26]$
3	$\tilde{I}_0 = 1, \quad \tilde{I}_1 = \sqrt{\frac{2\sigma^2}{\pi}}\frac{1}{\operatorname{erfcx}(\eta)} + c_2$ $\tilde{I}_2 = \sqrt{\frac{2\sigma^2}{\pi}}\frac{c_1}{\operatorname{erfcx}(\eta)} + c_2^2 + \sigma^2$	$(26, \infty)$

We can compute  $\bar{s}$  and  $C_s$  directly from  $L_y$ . The identities follow from straightforward computation.

**Theorem 4.2.** *If  $r = 0$  or  $b = -r$ , the ratios for calculating  $\bar{s}$  and  $C_s$  are given by*

$$\frac{I_{y+1}}{I_y} = (m - \sigma^2 + r) + \sigma^2 L_y \quad \text{and} \quad \frac{I_{y+2}}{I_y} = e^{\ln(y+1) + \ln(y+2) - \ln L_{y+1} - \ln L_{y+2}}.$$

Last, we discuss the computation of the first three integrals  $I_0$ ,  $I_1$  and  $I_2$ , which are needed for the recursion. We employ three different forms according to the integration range with respect to the auxiliary variable

$$\eta = \frac{\sigma^2 - m + b}{\sqrt{2\sigma^2}}.$$

The formulas are listed in table 1, where  $\operatorname{erf}$  and  $\operatorname{erfc}$  denote the error function and complementary error function, respectively, and  $\operatorname{erfcx}(\eta) = e^{\eta^2}(1 - \operatorname{erf}(\eta))$ . Since the value of  $1 - \operatorname{erf}(\eta)$  is vanishingly small for large  $\eta$  value, we use Scheme 2 to avoid underflow. Scheme 3 is useful when the  $\eta$  value is large, since both  $1 - \operatorname{erf}(\eta)$  and  $\operatorname{erfc}(\eta)$  suffer from numerical underflow. Note that when  $\eta$  is small, Scheme 3 is not as accurate as Scheme 2, so we use Scheme 2 in the intermediate range. In our experiments, we use Scheme 1 for  $\eta \in (-\infty, 5)$ , Scheme 2 for  $\eta \in [5, 26]$  and Scheme 3 for  $\eta \in (26, \infty)$ . To use Scheme 3, we construct  $\tilde{I}_i = \frac{I_i}{I_0}$ ,  $i = 0, 1, 2$ , and  $\tilde{L}_y = \frac{L_{y-1}}{L_y}$ ,  $y \in \mathbb{N}_+$ . Then similar identities for computing  $\bar{s}$  and  $C_s$  hold, i.e.  $\bar{s} = \frac{\tilde{I}_{y+1}}{\tilde{I}_y} - r$  and  $C_s = \frac{\tilde{I}_{y+2}}{\tilde{I}_y} - (\frac{\tilde{I}_{y+1}}{\tilde{I}_y})^2$ , with  $\frac{\tilde{I}_{y+1}}{\tilde{I}_y} = (m - \sigma^2 + r) + \sigma^2 \tilde{L}_y$  and  $\frac{\tilde{I}_{y+2}}{\tilde{I}_y} = e^{\ln(y+1) + \ln(y+2) - \ln \tilde{L}_{y+1} - \ln \tilde{L}_{y+2}}$ .

#### 4.2. Laplace potential

Now we derive the formulas for evaluating the 1D integrals for the Laplace potential  $t(x) = \frac{\alpha}{2}e^{-\alpha|\ell^t x|}$ . For any fixed  $\ell \in \mathbb{R}^n$ , we divide the whole space  $\mathbb{R}^n$  into two disjoint half-spaces  $V_+$  and  $V_-$ , i.e.  $\mathbb{R}^n = V_+ \cup V_-$ , with  $V_+ = \{x|\ell^t x > 0\}$  and  $V_- = \{x|\ell^t x \leq 0\}$ . Then we split the Laplace potential  $t(x)$  into

$$t(x) = \frac{\alpha}{2} e^{-\alpha \ell x} \mathbf{1}_{V_+}(x) + \frac{\alpha}{2} e^{\alpha \ell x} \mathbf{1}_{V_-}(x).$$

The integrals involving  $t(x)\mathcal{N}(s|\mu, \sigma^2)$  (slightly abusing  $\mu$ ) can be divided into two parts:

$$\begin{aligned} \int_{\mathbb{R}_+} \frac{\alpha}{2} s^i e^{-\alpha s} \mathcal{N}(s|\mu, \sigma^2) ds &= \frac{\alpha}{2} e^{\frac{\alpha^2 \sigma^2}{2}} e^{-\alpha \mu} \underbrace{\int_{\mathbb{R}_+} s^i \mathcal{N}(s|\mu - \alpha \sigma^2, \sigma^2) ds}_{:=I_i^+}, \\ \int_{\mathbb{R}_-} \frac{\alpha}{2} s^i e^{\alpha s} \mathcal{N}(s|\mu, \sigma^2) ds &= \frac{\alpha}{2} e^{\frac{\alpha^2 \sigma^2}{2}} e^{\alpha \mu} \underbrace{\int_{\mathbb{R}_-} s^i \mathcal{N}(s|\mu + \alpha \sigma^2, \sigma^2) ds}_{:=I_i^-}. \end{aligned}$$

By the change of variable  $t = \frac{s - \mu \pm \alpha \sigma^2}{\sigma}$  for  $I_i^\pm$  respectively, we have

$$\begin{aligned} I_i^+ &= \frac{e^{-\alpha \mu}}{\sqrt{2\pi}} \int_{-\frac{\mu}{\sigma} + \alpha \sigma}^{+\infty} (\sigma t + \mu - \alpha \sigma^2)^i e^{-\frac{t^2}{2}} dt, \\ I_i^- &= \frac{(-1)^i e^{\alpha \mu}}{\sqrt{2\pi}} \int_{\frac{\mu}{\sigma} + \alpha \sigma}^{+\infty} (\sigma t - \mu - \alpha \sigma^2)^i e^{-\frac{t^2}{2}} dt. \end{aligned}$$

These integrals can be expressed using the cumulative distribution function  $\Phi$  of the standard Gaussian distribution. We shall view  $I_i^\pm$  as functions of  $\mu$  and let  $I_i = I_i^+(\mu) + (-1)^i I_i^+(-\mu)$ . Then we have

$$\bar{s} = \frac{I_1}{I_0} \quad \text{and} \quad C_s = \frac{I_2}{I_0} - \left(\frac{I_1}{I_0}\right)^2.$$

To avoid the potential underflow of direct evaluation of  $\Phi$ , we use the following well known (divergent) asymptotic expansion [1, item 7.1.23]

$$\begin{aligned} 1 - \Phi(\eta) &= \int_{\eta}^{\infty} e^{-\frac{t^2}{2}} dt = e^{-\frac{\eta^2}{2}} \left( \eta^{-1} + \sum_{k=1}^{\infty} \frac{(-1)^k (2k-1)!}{2^k (k-1)!} \eta^{-(2k+1)} \right) \\ &= \mathcal{N}(\eta|0, 1) \eta^{-1} \underbrace{\sum_{n=0}^{\infty} (-1)^n (2n-1)!! \eta^{-2n}}_{:=g(\eta)}. \end{aligned}$$

This formula follows by integration by parts, and allows accurate evaluation for large positive  $\eta$ . It was shown in [18] that the error of evaluating  $1 - \Phi(\eta)$  with a truncation of the asymptotic expansion is less than  $10^{-11}$  for  $\eta > 5$  with more than 8 terms in the summation of  $g(\eta)$ . For  $\eta \leq 5$ ,  $1 - \Phi(\eta)$  can be accurately evaluated directly. Then we introduce a ratio

$$\beta = \frac{I_0^+(-|\mu|)}{I_0^+(|\mu|)} = e^{2\alpha|\mu|} \frac{(\alpha\sigma^2 - |\mu|)g(\alpha\sigma + \frac{|\mu|}{\sigma})}{(\alpha\sigma^2 + |\mu|)g(\alpha\sigma - \frac{|\mu|}{\sigma})}.$$

With the ratio  $\beta$ , the two fractions  $\frac{I_1}{I_0}$  and  $\frac{I_2}{I_0}$  can be evaluated by

$$\begin{aligned}\frac{I_1}{I_0} &= \mu + \alpha\sigma^2 \operatorname{sgn}(\mu) \left(1 - \frac{2}{1+\beta}\right), \\ \frac{I_2}{I_0} &= -\frac{2\alpha\sigma^3}{\sqrt{2\pi}} e^{-\left(\frac{\mu^2}{2\sigma^2} + \frac{\alpha^2\sigma^2}{2}\right)} I_0^{-1} + (\sigma^2 + \alpha^2\sigma^4 - \mu^2) + 2\mu \frac{I_1}{I_0}.\end{aligned}$$

To avoid potential numerical instability of the first term in  $\frac{I_2}{I_0}$ , we use the identity

$$-\frac{2\alpha\sigma^3}{\sqrt{2\pi}} e^{-\left(\frac{\mu^2}{2\sigma^2} + \frac{\alpha^2\sigma^2}{2}\right)} I_0^{-1} = \frac{-2\alpha\sigma^2(-|\mu| + \alpha\sigma^2)}{g\left(-\frac{|\mu|}{\sigma} + \alpha\sigma\right)(1+\beta)}.$$

To avoid potential numerical instability of the term  $\sigma^2 + \alpha^2\sigma^4$ , we use the exp-log trick

$$\sigma^2 + \alpha^2\sigma^4 = \exp\left(-2\log\frac{1}{\alpha\sigma^2} + \log\left(1 + \frac{1}{\alpha^2\sigma^2}\right)\right),$$

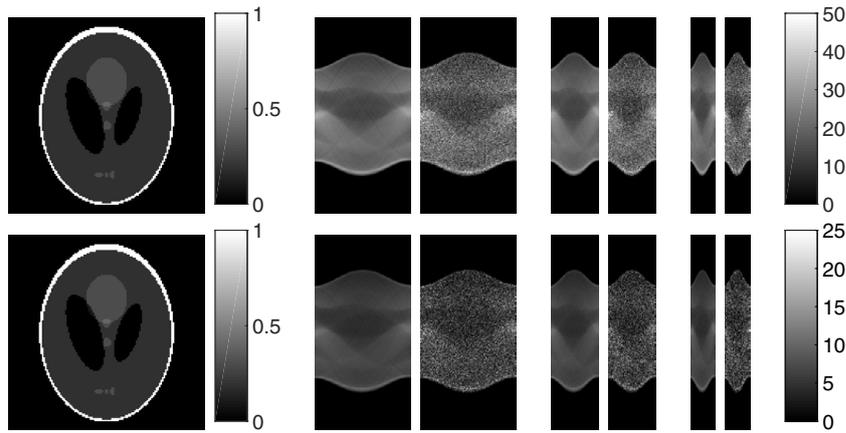
where  $\log\left(1 + \frac{1}{\alpha^2\sigma^2}\right)$  is evaluated by the MATLAB builtin function `log1p`. Thus,  $\bar{s}$  and  $C_s$  can be evaluated by  $\bar{s} = \frac{I_1}{I_0}$  and  $C_s = -\frac{2\alpha\sigma^2(-|\mu| + \alpha\sigma^2)}{g\left(-\frac{|\mu|}{\sigma} + \alpha\sigma\right)(1+\beta)} + \exp\left[-2\log\frac{1}{\alpha\sigma^2} + \log\left(1 + \frac{1}{\alpha^2\sigma^2}\right)\right] - \left(\mu - \frac{I_1}{I_0}\right)^2$ .

## 5. Numerical experiments and discussions

Now we numerically illustrate one EP algorithm on realistic images. In the implementation, we employ the natural parameter parameterization, i.e. algorithm 1, which appears to be numerically more robust. We measure the accuracy of a reconstruction  $x^*$  relative to the ground truth  $x^\dagger$  by the standard  $L^2$ -error  $\|x^* - x^\dagger\|_2$ , the structural similarity (SSIM) index (by MATLAB built-in `ssim`), and peak signal-to-noise ratio (PSNR) (by MATLAB built-in `psnr` with peak value 1 for Shepp–Logan and PET phantom, and 5 for IRT phantom). For comparison, we also present MAP, computed by a limited-memory BFGS algorithm [31] with constraint  $C_1$ . The hyperparameter  $\alpha$  in the prior  $p(x)$  is determined in a trial-and-error manner. Unless otherwise stated, the EP algorithm is run for four sweeps through the sites.

### 5.1. Simulated data with two benchmark images

First, we take simulated data: the ground-truth images are Shepp–Logan and PET [15] phantoms of size  $128 \times 128$ ; The map  $A$  is a discrete Radon transform, formed using MATLAB built-in function `radon` with 185 projections per angle and three different angle settings, i.e.  $[0 : 2 : 179]$ ,  $[0 : 4 : 179]$  and  $[0 : 8 : 179]$ , and accordingly, the matrix  $A$  is of size  $A \in \mathbb{R}^{16650 \times 16384}$ ,  $A \in \mathbb{R}^{8325 \times 16384}$  and  $A \in \mathbb{R}^{4255 \times 16384}$ . For each image, we consider two count levels: the moderate count case is obtained from  $A$ , and the low count case from  $A/3$  (so that the measured counts are mostly below 10). The original image, sinogram and observed Poisson data are shown in figures 1 and 2 for Shepp–Logan and PET, respectively. The numerical results are summarized in tables 2 and 3, figures 3, 4, 5 and 6. The EP mean is mostly comparable with MAP in all three metrics for both moderate count and low count cases, and the reconstruction quality improves steadily as the number of projection angles increases. Interestingly, the shape of the EP variance resembles closely the outer boundary of the phantom, whereas within the boundary, there is little difference in the magnitudes. This might indicate that the algorithm is rather certain in the cold regions where the error is close to zero



**Figure 1.** The exact image, sinograms and data with three different  $\alpha$ s for Shepp–Logan phantom. The top and bottom rows refer to the moderate count and low count cases, respectively.

**Table 2.** Comparisons between EP mean and MAP for the Shepp–Logan phantom. The top and bottom blocks refer to the moderate count and low count cases, respectively.

Angle	[0:2:179]		[0:4:179]		[0:8:179]	
$\alpha$	6		4		3	
Method	EP	MAP	EP	MAP	EP	MAP
$L^2$ error	5.32	5.36	5.64	5.67	6.09	6.11
SSIM	0.74	0.78	0.70	0.75	0.67	0.72
PSNR	18.58	18.53	17.97	17.93	17.29	17.27
CPU time (s)	80187.88	124.44	46031.95	55.55	29274.16	27.23
$\alpha$	1.3		2		1	
Method	EP	MAP	EP	MAP	EP	MAP
$L^2$ error	4.07	4.09	6.15	6.24	6.14	6.19
SSIM	0.57	0.79	0.51	0.72	0.48	0.70
PSNR	19.50	19.47	17.53	17.42	17.18	17.15
CPU time (s)	82125.92	42.25	47110.50	29.69	29756.10	15.20

and more uncertain about the region where the error is potentially larger. It is observed that the computational complexity of the EP grows with the amount of the data. This is attributed to the following fact: the number of sweeps is fixed at four, and the complexity increases with the number of projection angles. Since the computing time is presented only for one reconstruction at each case, these numbers should be viewed as a representative instead of an absolute measure for algorithmic performance. Roughly, EP is about two orders of magnitude more expensive than the MAP approach (computed by limited memory BFGS [31]).

The Poisson model is especially useful for low count data, where a naive Gaussian approximation can fail to give reasonable reconstructions. The EP results for the low-count case are shown in figures 4 and 6. Just as expected, the reconstruction accuracy deteriorates as the count level decreases. Nonetheless, the EP means remain largely comparable with MAP results both qualitatively and quantitatively. Note that for the PET image, the reconstruction

**Table 3.** The comparisons between EP mean and MAP for the PET phantom. The top and bottom blocks refer to the moderate and low count cases, respectively.

Angle	[0:2:179]		[0:4:179]		[0:8:179]	
$\alpha$	1.6		1.4		1.2	
Method	EP	MAP	EP	MAP	EP	MAP
$L^2$ error	7.37	7.45	8.55	8.64	8.81	8.87
SSIM	0.72	0.81	0.61	0.75	0.57	0.70
PSNR	19.82	19.79	18.42	18.35	17.35	17.28
CPU time (s)	91263.00	110.05	53863.77	78.69	31537.05	28.20
$\alpha$	1.2		$9 \times 10^{-1}$		$7.5 \times 10^{-1}$	
Method	EP	MAP	EP	MAP	EP	MAP
$L^2$ error	8.96	9.04	9.30	9.35	10.13	10.17
SSIM	0.55	0.72	0.49	0.67	0.43	0.62
PSNR	17.66	17.61	16.93	16.89	15.84	15.81
CPU time (s)	82542.76	52.97	47263.64	32.43	29737.91	18.01

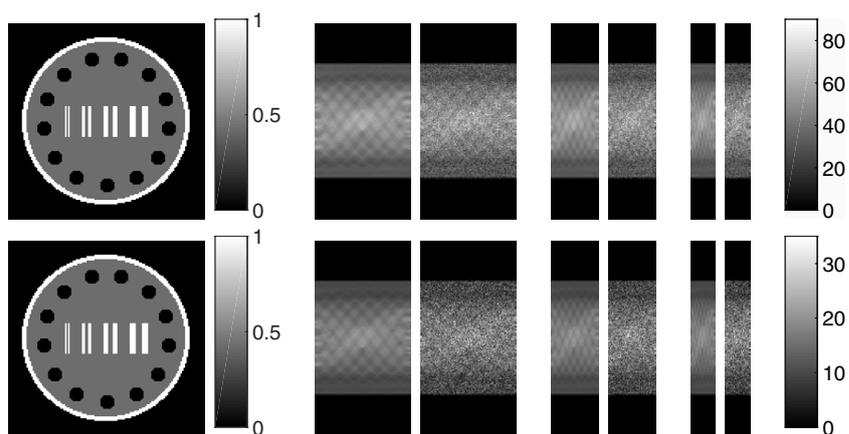
accuracy for both EP and MAP suffers significantly in that the fine details such as vertical bars in the true image disappear, especially when the number of projection angles is small. The computing times for the moderate count and low count cases are nearly the same; see tables 2 and 3. Thus, EP is still feasible for the low-count case.

To further illustrate the approximation, we plot in figure 7 the cross-sections and 95% highest posterior density (HDP) region, which is estimated from the EP covariance. The EP mean is close to MAP, and thus also suffers slightly from a reduced magnitude, as is typical of the total variation penalty in variational regularization [10]. This also concurs with the error metrics in tables 2 and 3. The thrust of EP is that it can also provide uncertainty estimates via covariance, which is unavailable from MAP. In sharp contrast, the popular Laplace approximation (see appendix B) can fail to yield a reasonable approximation for nonsmooth priors such as anisotropic total variation, whereas MCMC tends to be prohibitively expensive for large images, though being asymptotically exact; see appendix C for further numerical results. So overall, EP represents a computationally feasible approach to deliver uncertainty estimates for these benchmark images with Poisson data.

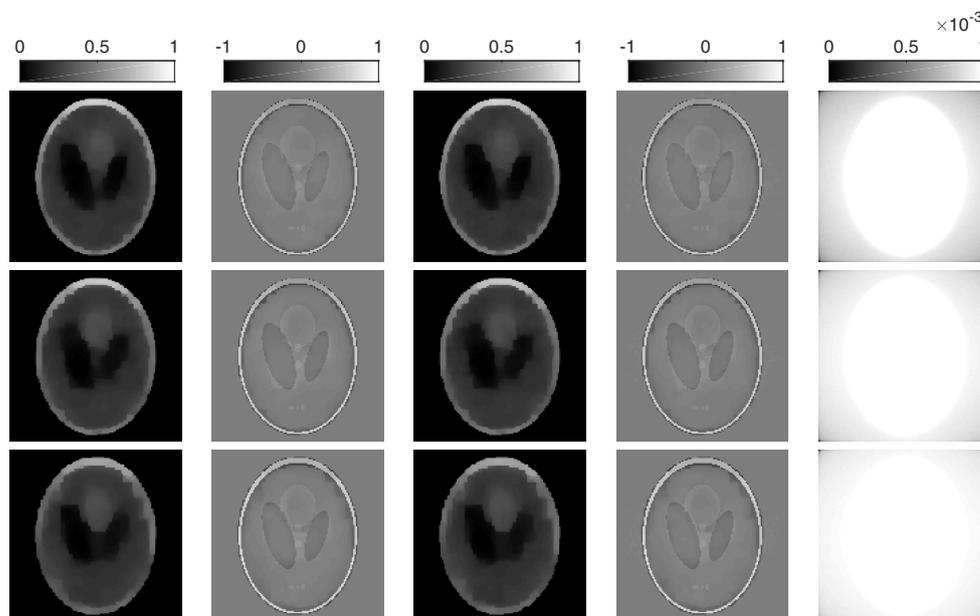
## 5.2. Convergence of the EP algorithm

Next, we present an experimental evaluation of the convergence of the EP algorithm, which is a long outstanding theoretical issue, on the following experimental setup: Shepp–Logan phantom and Radon matrix  $A \in \mathbb{R}^{4255 \times 16384}$  (i.e. 185 projections per angle and [0 : 8 : 179], moderate count case). We denote the mean and covariance after  $k$  outer iterations (i.e. sweeps through all the sites) by  $\mu^k$  and  $C^k$ , respectively, and the converged iterate tuple by  $(\mu^*, C^*)$ . The EP mean  $\mu^k$  converges rapidly, and visually it reaches convergence after five iterations since thereafter the cross-sections graphically overlap with each other; see figure 8. Thus, in the numerical experiments, we have fixed the number of outer iterations to four, and the complexity of the reconstruction algorithm is of order  $O(mn^2)$ . Figure 9 shows the errors of the iterate tuple  $(\mu^k, C^k)$  with respect to  $(\mu^*, C^*)$ , where the errors

$$\delta\mu = \mu^k - \mu^* \quad \text{and} \quad \delta C = C^k - C^*$$



**Figure 2.** The exact image, sinograms and observed data with three different  $A$ 's for the PET phantom. The top and bottom rows refer to the moderate count and low count cases, respectively.

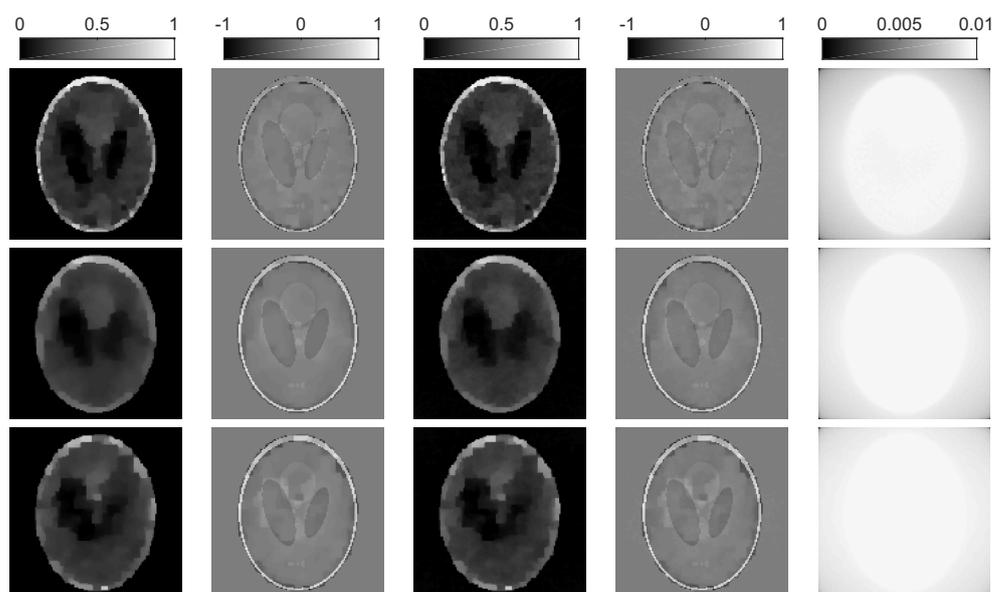


**Figure 3.** MAP versus EP with anisotropic TV prior for the Shepp-Logan phantom, moderate count case. Rows from top to bottom: [0:2:179], [0:4:179] and [0:8:179]. Columns from left to right: MAP, MAP error, EP mean, EP error and EP variance.

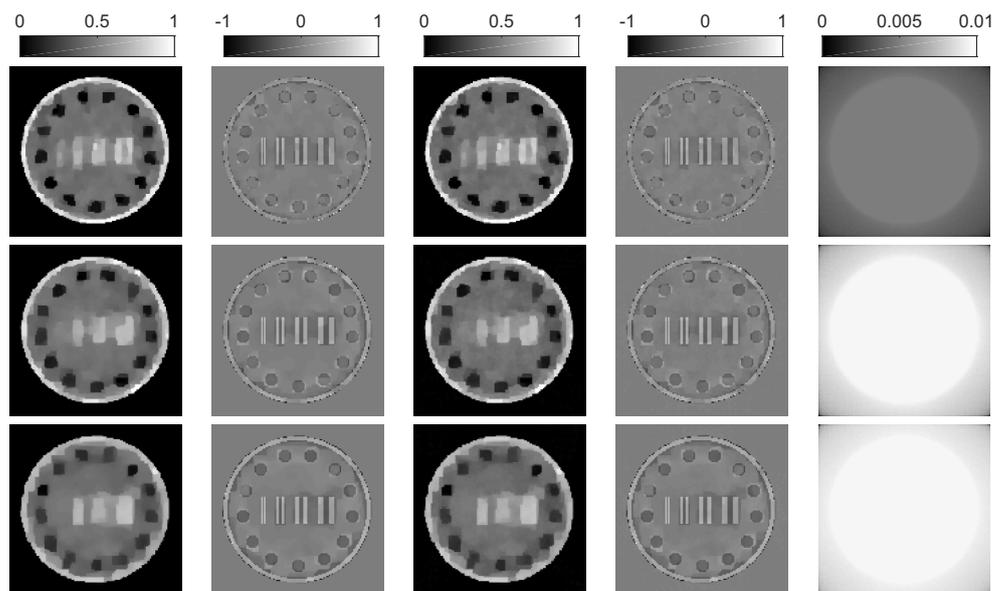
are measured by the  $L^2$ -norm and spectral norm, respectively. This phenomenon is also observed for all other experiments, although not presented. Hence, both mean and covariance converge rapidly, showing the steady and fast convergence of EP.

### 5.3. Real data

Last, we illustrate the inference procedure with a dataset taken from Michigan Image Reconstruction Toolbox<sup>2</sup>. The ground truth image is denoted by IRT. The map

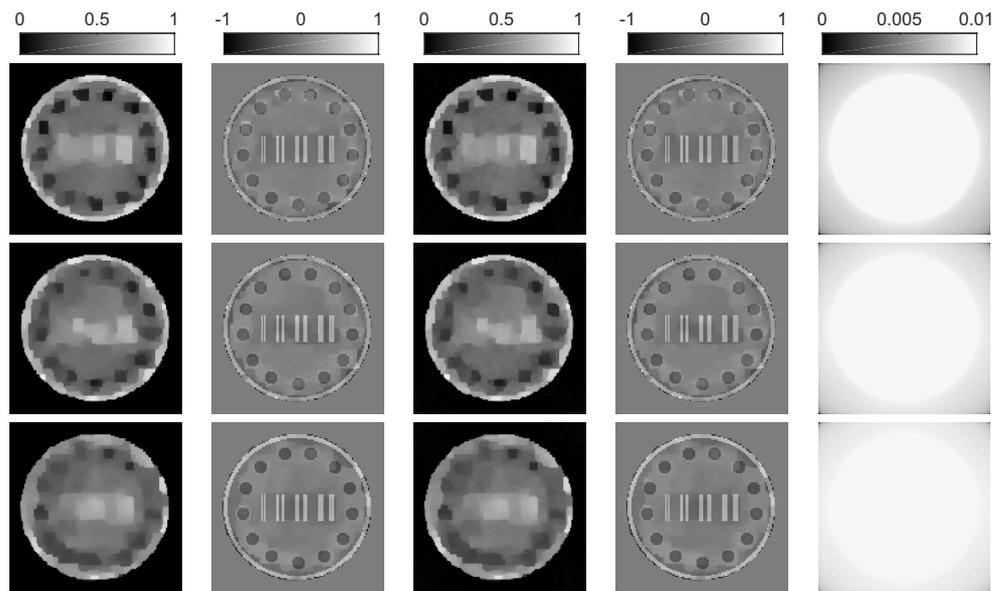


**Figure 4.** MAP versus EP with anisotropic TV prior for the Shepp–Logan phantom, low count case. Rows from top to bottom: [0:2:179], [0:4:179] and [0:8:179]. Columns from left to right: MAP, MAP error, EP mean, EP error and EP variance.

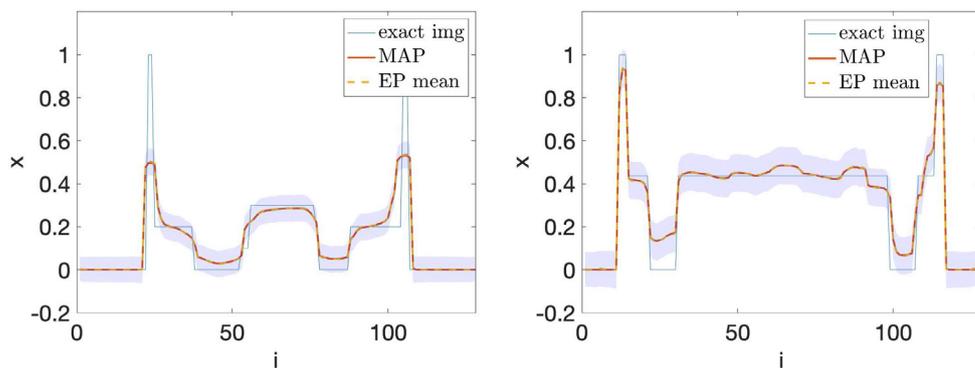


**Figure 5.** MAP versus EP with anisotropic TV prior for the PET phantom, moderate count case. Rows from top to bottom: [0:2:179], [0:4:179] and [0:8:179]. Columns from left to right: MAP, MAP error, EP mean, EP error and EP variance.

$A \in \mathbb{R}^{24960 \times 16384}$  is assembled by  $A = \text{diag}(c_i)G$ , where  $G$  is the system matrix and  $c_i$  is an attenuation vector, by setting the mask to all values being unity and other parameters to default. The exact image and data are shown in figure 10; see figure 11 for reconstructions, obtained with a regularization parameter  $\alpha = 0.4$ . The  $L^2$  error, SSIM and PSNR for EP and



**Figure 6.** MAP versus EP with anisotropic TV prior for the PET phantom, low count case. Rows from top to bottom: [0:2:179], [0:4:179] and [0:8:179]. Columns from left to right: MAP, MAP error, EP mean, EP error and EP variance.

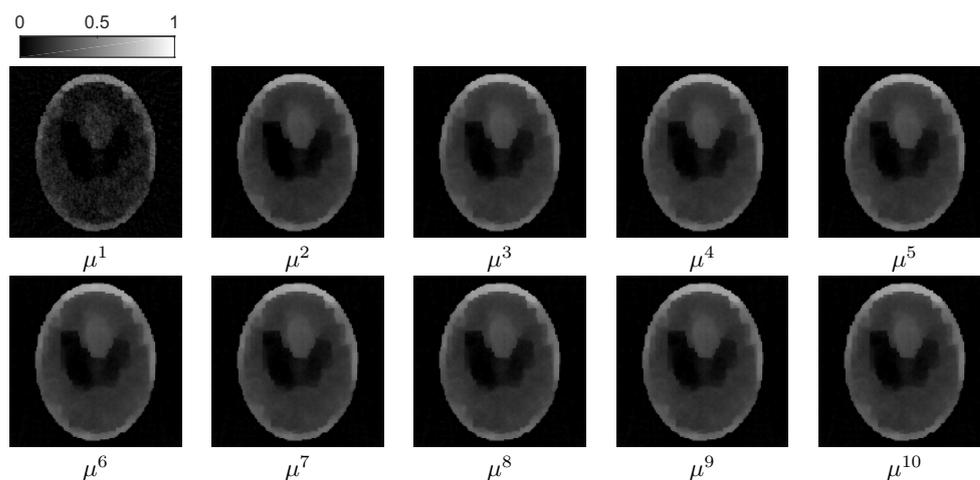


**Figure 7.** The 50th cross-sections of the two phantoms and 0.95-HPD regions, moderate count case. From left to right: Shepp–Logan phantom and PET phantom.

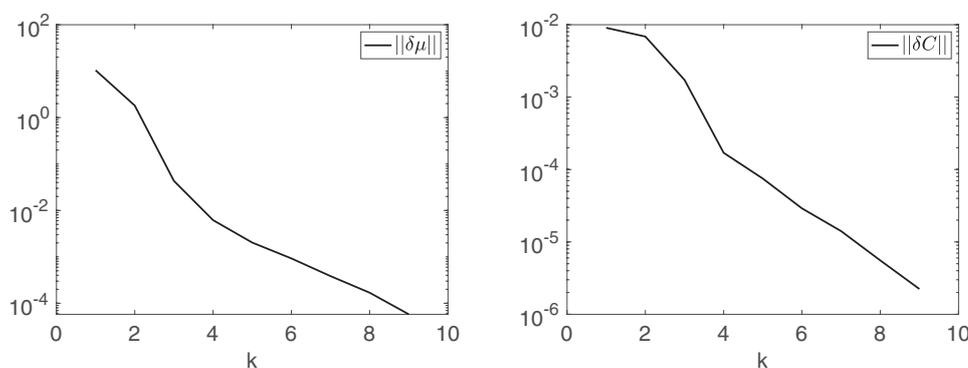
MAP are, respectively, 13.46 and 13.48, 0.62 and 0.83, and 25.66 and 25.64. Thus, the EP results and MAP are comparable, and the preceding observations remain valid.

These numerical results with different experimental settings show clearly that EP can provide comparable point estimates with MAP as well as uncertainty information by means of the variance estimate.

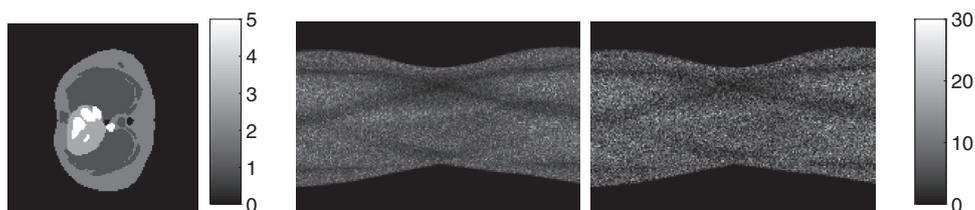
<sup>2</sup><https://web.eecs.umich.edu/~fessler/code/>, last accessed on July 30, 2018.



**Figure 8.** The convergence of the mean  $\mu^k$  by EP after  $k$  outer iterations for the Shepp-Logan phantom, moderate count case.



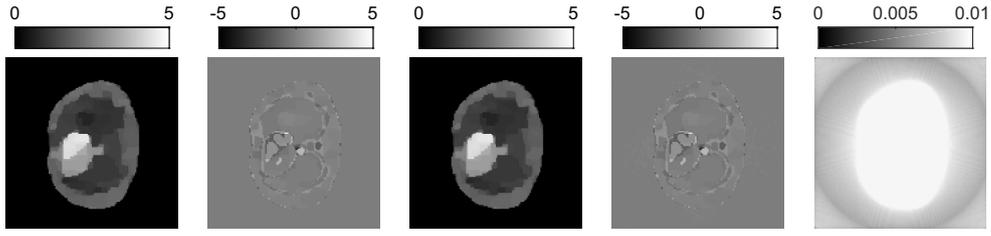
**Figure 9.** The convergence of the mean  $\mu$  and covariance  $C$  after each outer iteration, moderate count case.



**Figure 10.** The exact image, sinograms and observed data for IRT phantom.

## 6. Conclusion

In this work, we have developed inference procedures for the constrained Poisson likelihood arising in emission tomography. They are based on expectation propagation developed in the machine learning community. The detailed derivation of the algorithms, complexity and their



**Figure 11.** MAP versus EP with anisotropic TV prior for the IRT phantom. From left to right: MAP, MAP error, EP mean, EP error and EP variance.

stable implementation are given for a Laplace type prior. Extensive numerical experiments show that the EP algorithm (with natural parameters) converges rapidly and can deliver an approximate posterior distribution with the approximate mean comparable with MAP, together with uncertainty estimate, and can handle real images of medium size. Thus, the approach can be viewed as a feasible fast alternative to the general-purposed but expensive MCMC for rapid uncertainty quantification with Poisson data.

There are several avenues for future works. First, it is of enormous interest to analyze the convergence rate and accuracy of EP, and more general approximate inference techniques, e.g. variational Bayes, which have all achieved great practical successes but largely defied theoretical analysis. Second, it is important to further extend the flexibility of EP algorithms to more complex posterior distributions, e.g. lack of projection form. One notable example is isotropic total variation prior that appears frequently in practical imaging algorithms. This may require introducing an additional layer of approximation, e.g. in the spirit of iteratively reweighted least-squares or (quasi-)Monte Carlo computation of low-dimensional integrals. Third, many experimental studies show that EP converges very fast, with convergence reached within five sweeps for the Poisson model under considerations. However, the overall  $O(mn^2)$  computational complexity per sweep of all current implementations [20] is still very high, and not scalable well to large images that are required in many real world applications. Hence, there remains great demand to further accelerate the algorithms, e.g. via low-rank structure of the map  $A$  and diagonal dominance of the posterior covariance. Fourth and last, it is also important to derive rigorous error estimates for the quadrature rules developed in section 4.

## Appendix A. Parameterizing Gaussian distributions

For a Gaussian  $\mathcal{N}(x|\mu, C)$  with mean  $\mu \in \mathbb{R}^n$  and covariance  $C \in \mathcal{S}_+^n$ , the density  $\pi(x|\mu, C)$  is given by

$$\pi(x|\mu, C) = (2\pi)^{-\frac{n}{2}} |C|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)'C^{-1}(x-\mu)} = e^{\zeta+h'x-\frac{1}{2}x'\Lambda x},$$

where the parameters  $\Lambda \in \mathcal{S}_+^n$ ,  $h \in \mathbb{R}^n$  and  $\zeta \in \mathbb{R}$  are respectively given by

$$\Lambda = C^{-1}, \quad h = \Lambda\mu, \quad \text{and} \quad \zeta = -\frac{1}{2}(n \log 2\pi + \log |\Lambda| + \mu'\Lambda\mu).$$

Thus, the density  $\pi(x|\mu, C)$  is also uniquely defined by  $\Lambda$  and  $h$ . In the literature,  $\Lambda$  is often referred to as the precision matrix and  $h$  as the precision mean and the pair  $(h, \Lambda)$  is called the natural parameter of a Gaussian distribution.

It is easy to check that the product of  $k$  Gaussians  $\{\mathcal{N}(x|\mu_k, C_k)\}_{k=1}^m$  is also a Gaussian  $\mathcal{N}(x|\mu, C)$  after normalization, and the mean  $\mu$  and covariance  $C$  of the product are given by

$$\mu = C \sum_{k=1}^m C_k^{-1} \mu_k \quad \text{and} \quad C = \left( \sum_{k=1}^m C_k^{-1} \right)^{-1},$$

or equivalently

$$h = \sum_{k=1}^m h_k \quad \text{and} \quad \Lambda = \sum_{k=1}^m \Lambda_k.$$

## Appendix B. Laplace approximation

In the engineering community, one popular approach to approximate the posterior distribution  $p(x|y)$  is Laplace approximation [5, 43]. It constructs a Gaussian approximation by the second-order Taylor expansion of the negative log-posterior  $-\log p(x|y)$  around MAP  $\hat{x}$ . Upon ignoring the unimportant constant and smoothing the Laplace potential, the negative log-posterior  $J(x)$  is given by

$$J(x) = \sum_{i=1}^{m_1} (-y_i \log(a_i^t x + r_i) + a_i^t x + r_i) + \alpha \sum_{i=1}^{m_2} ((L_i^t x)^2 + \epsilon^2)^{1/2},$$

where  $\epsilon > 0$  is a small smoothing parameter to restore the differentiability. The gradient  $\nabla J(x)$  and Hessian  $\nabla^2 J(x)$  are given respectively by

$$\begin{aligned} \nabla J(x) &= \sum_{i=1}^{m_1} \left( -\frac{y_i}{a_i^t x + r_i} + 1 \right) a_i + \alpha \sum_{i=1}^{m_2} ((L_i^t x)^2 + \epsilon^2)^{-1/2} (L_i^t x) L_i, \\ \nabla^2 J(x) &= \sum_{i=1}^{m_1} \frac{y_i}{(a_i^t x + r_i)^2} a_i a_i^t + \alpha \epsilon^2 \sum_{i=1}^{m_2} ((L_i^t x)^2 + \epsilon^2)^{-3/2} L_i L_i^t. \end{aligned}$$

Since  $\nabla J(\hat{x}) = 0$ , the Taylor expansion reads

$$J(x) \approx J(\hat{x}) + \frac{1}{2} (x - \hat{x})^t \nabla^2 J(\hat{x}) (x - \hat{x}), \quad (\text{B.1})$$

and  $\nabla^2 J(\hat{x})$  approximates the precision matrix. When  $\epsilon \ll |L_i^t \hat{x}|$ , the second term in  $\nabla^2 J(x)$  can be negligible and thus the Hessian of the negative log-likelihood is dominating; whereas for  $\epsilon \gg |L_i^t \hat{x}|$ , the second term is dominating. In either case, the approximation is problematic. In practice, it is also popular to combine smoothing with an iterative weighted approximation (e.g. lagged diffusivity approximation [45]) by fixing  $((L_i^t x)^2 + \epsilon^2)^{1/2}$  in  $\nabla J(x)$  at  $((L_i^t \hat{x})^2 + \epsilon^2)^{1/2}$ , which leads to a modified Hessian:

$$\tilde{\nabla}^2 J(x) = \sum_{i=1}^{m_1} \frac{y_i}{(a_i^t x + r_i)^2} a_i a_i^t + \alpha \sum_{i=1}^{m_2} ((L_i^t \hat{x})^2 + \epsilon^2)^{-1/2} L_i L_i^t.$$

The Hessians  $\nabla^2 J(\hat{x})$  and  $\tilde{\nabla}^2 J(\hat{x})$  will be close to each other, if  $|L_i^t \hat{x}|$  are all small, which is expected to hold for truly sparse signals, i.e.  $L_i^t x \approx 0$  for  $i = 1, \dots, m_2$ . One undesirable feature of Laplace approximation is that the precision approximation depends crucially on the smoothing parameter  $\epsilon$ .

## Appendix C. Comparison with MCMC and Laplace approximation

Numerically, the accuracy of EP has found to be excellent in several studies [19, 36], although there is still no rigorous justification. We provide an experimental evaluation of its accuracy with Markov chain Monte Carlo (MCMC) and Laplace approximation. The true posterior distribution  $p(x|y)$  can be explored by MCMC [32, 37]. However, usually a large number of samples are required to obtain reliable statistics. Thus, to obtain further insights, we consider a one-dimensional problem, i.e. a Fredholm integral equation of the first kind [35] over the interval  $[-6, 6]$  with the kernel  $K(s, t) = \phi(s - t)$  and exact solution  $x(t) = \phi(t)$ , where  $\phi(s) = 10 + 10 \cos \frac{\pi}{3} s \chi_{[-3,3]}$ . It is discretized by a standard piecewise constant Galerkin method, and the resulting problem is of size 100, i.e.  $x \in \mathbb{R}^{100}$  and  $A \in \mathbb{R}^{100 \times 100}$ . We implement a random walk Metropolis–Hastings sampler with Gaussian proposals, and optimize the step size so that the acceptance ratio is close to 0.23 in order to ensure good convergence [7]. The hyperparameter  $\alpha$  in the prior distribution is set to 1. The chain is run for a length of  $2 \times 10^7$ , and the last  $10^7$  samples are used for computing the mean and covariance.

To compare the Gaussian approximation by EP and MCMC results, we present the mean, MAP, covariance and 95% HPD region. Both approximations concentrate in the same region, and the shape and magnitude of 95% HPD/covariance are mostly comparable; see figures C1 and C2, showing the validity of EP. However, there are noticeable differences in the recovered mean: the EP mean is nearly piecewise constant, which differs from that by MCMC. So EP gives an intermediate approximation between the MAP and posterior mean. In comparison with MAP, EP provides not only a point estimate, but also the associated uncertainty, i.e. covariance. Interestingly, the covariance is clearly diagonal dominant, which suggests the use of a banded covariance or its Cholesky factor for potentially speeding up the algorithm.

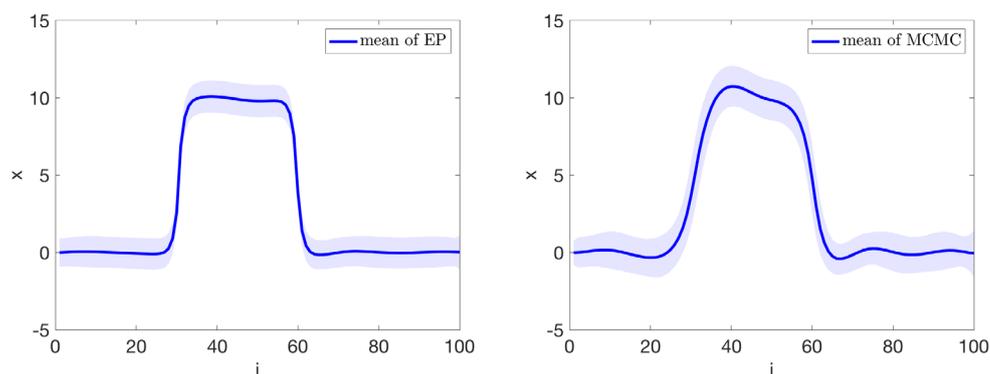
The Laplace approximation described in appendix B depends heavily on the smoothing parameter  $\epsilon$ , and clearly there is a tradeoff between accuracy of MAP and the variance approximation; see figure C3 for the numerical results corresponding to four different smooth parameters  $\epsilon$ , based on the approximation (B.1). This tradeoff is largely attributed to the nonsmooth Laplace type prior, which pose significant challenges for constructing the approximation. Thus, it is tricky to derive a reasonable approximation to the target posterior distribution. In contrast, the EP algorithm only involves integrals, which are more amenable to non-differentiability, and thus can handle nonsmooth priors naturally.

In passing, we note that the uncertainty estimate from the posterior probability distribution differs greatly from the concept of noise variance [16], which is mainly concerned with the *sensitivity* of the reconstruction with respect to the noise in the input data  $y$ . It is derived using chain rule and implicit function theorem, under the assumptions of good smoothness and local strong convexity of the associated functional [16]. In contrast, the uncertainty in the Bayesian framework as in this work originates from imprecise knowledge of the inverse solution encoded in the prior and the statistics of the data. Thus, the results of these two approaches are not directly comparable.

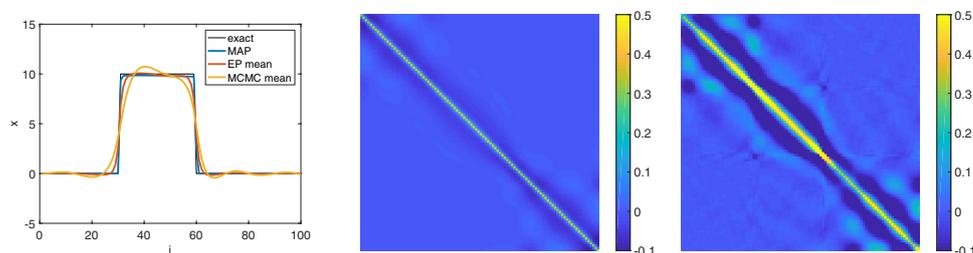
### ORCID iDs

Simon Arridge  <https://orcid.org/0000-0003-1292-0210>

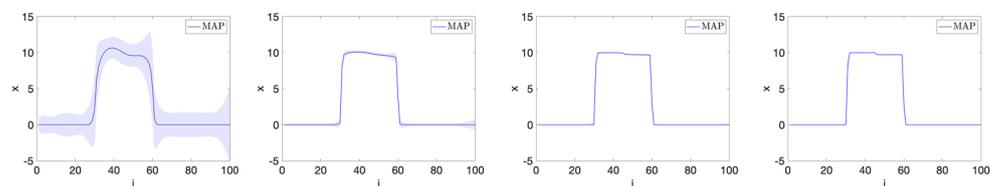
Bangti Jin  <https://orcid.org/0000-0002-3775-9155>



**Figure C1.** Comparisons of mean and 0.95 HPD between EP and MCMC for Phillips test. From left to right: EP mean with 0.95-HPD and MCMC mean with 0.95-HPD.



**Figure C2.** Comparisons of mean and covariance of EP and MCMC for Phillips test. From left to right: EP and MCMC means, EP covariance and MCMC covariance.



**Figure C3.** Laplace approximation with different smoothing  $\epsilon$ . From left to right:  $\epsilon = 5.18 \times 10^{-1}$ ,  $\epsilon = 5.18 \times 10^{-2}$ ,  $\epsilon = 5.18 \times 10^{-3}$  and  $\epsilon = 5.18 \times 10^{-4}$ .

## References

- [1] Abramowitz M and Stegun I A 1965 *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables* (Washington DC: National Institute of Standards and Technology)
- [2] Anscombe F J 1948 The transformation of Poisson, binomial and negative-binomial data *Biometrika* **35** 246–54
- [3] Arridge S R, Ito K, Jin B and Zhang C 2018 Variational Gaussian approximation for Poisson data *Inverse Problems* **34** 025005
- [4] Bertero M, Boccacci P, Desiderà G and Vicidomini G 2009 Image deblurring with Poisson data: from cells to galaxies *Inverse Problems* **25** 123006
- [5] Bishop C M 2006 *Pattern Recognition and Machine Learning* (New York: Springer)

- [6] Blei D M, Kucukelbir A and McAuliffe J D 2017 Variational inference: a review for statisticians *J. Am. Stat. Assoc.* **112** 859–77
- [7] Brooks S P and Gelman A 1998 General methods for monitoring convergence of iterative simulations *J. Comput. Graph. Stat.* **7** 434–55
- [8] Cameron A C and Trivedi P K 2013 *Regression Analysis of Count Data* 2nd edn (Cambridge: Cambridge University Press)
- [9] Challis E and Barber D 2013 Gaussian Kullback–Leibler approximate inference *J. Mach. Learn. Res.* **14** 2239–86
- [10] Chan T F and Shen J 2005 *Image Processing and Analysis* (Philadelphia, PA: SIAM) (Variational, PDE, Wavelet, and Stochastic Methods)
- [11] Cunningham J P, Hennig P and Lacoste-Julien S 2011 Gaussian probabilities and expectation propagation (arXiv:1111.6832)
- [12] De Pierro A R 1995 A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography *IEEE Trans. Med. Imaging* **14** 132–7
- [13] Dehaene G and Barthelmé S 2015 Bounding errors of expectation-propagation *Advances in Neural Information Processing Systems* vol 28, ed C Cortes *et al* pp 244–52
- [14] Dehaene G and Barthelmé S 2018 Expectation propagation in the large data limit *J. R. Stat. Soc. B* **80** 199–217
- [15] Ehrhardt M J, Thielemans K, Pizarro L, Atkinson D, Ourselin S, Hutton B F and Arridge S R 2015 Joint reconstruction of PET-MRI by exploiting structural similarity *Inverse Problems* **31** 015001
- [16] Fessler J A 1996 Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): applications to tomography *IEEE Trans. Imaging Proc.* **5** 493–506
- [17] Gal Y and Ghahramani Z 2016 Dropout as a Bayesian approximation: representing model uncertainty in deep learning *Proc. Int. Conf. on Machine Learning* ed M F Balcan and K Q Weinberger pp 1050–9
- [18] Gehre M 2013 Rapid uncertainty quantification for nonlinear inverse problems *PhD Thesis* University of Bremen Bremen
- [19] Gehre M and Jin B 2014 Expectation propagation for nonlinear inverse problems—with an application to electrical impedance tomography *J. Comput. Phys.* **259** 513–35
- [20] Gelman A, Vehtari A, Jylänki P, Robert C, Chopin N and Cunningham J P 2014 Expectation propagation as a way of life (arXiv:1412.4869)
- [21] Golub G H and Van Loan C F 2012 *Matrix Computations* 3rd edn (Baltimore, MD: John Hopkins University Press)
- [22] Hohage T and Werner F 2016 Inverse problems with Poisson data: statistical regularization theory, applications and algorithms *Inverse Problems* **32** 093001
- [23] Ito K and Jin B 2015 *Inverse Problems: Tikhonov Theory and Algorithms* (Hackensack, NJ: World Scientific)
- [24] Jin B 2012 A variational Bayesian method to inverse problems with impulsive noise *J. Comput. Phys.* **231** 423–35
- [25] Jin B and Zou J 2009 Augmented Tikhonov regularization *Inverse Problems* **25** 025001
- [26] Jordan M I, Ghahramani Z, Jaakkola T S and Saul L K 1999 An introduction to variational methods for graphical models *Mach. Learn.* **37** 183–233
- [27] Kaipio J and Somersalo E 2005 *Statistical and Computational Inverse Problems* (New York: Springer)
- [28] Ko Y J and Seeger M W 2016 Expectation propagation for rectified linear poisson regression *Asian Conf. on Machine Learning* vol PMLR 45, ed G Holmes and T-Y Liu pp 253–68
- [29] Kullback S and Leibler R A 1951 On information and sufficiency *Ann. Math. Stat.* **22** 79–86
- [30] Lim H, Dewaraja Y K and Fessler J A 2018 A PET reconstruction formulation that enforces non-negativity in projection space for bias reduction in Y-90 imaging *Phys. Med. Biol.* **63** 035042
- [31] Liu D C and Nocedal J 1989 On the limited memory bf GS method for large scale optimization *Math. Program B* **45** 503–28
- [32] Liu J S 2001 *Monte Carlo Strategies in Scientific Computing* (New York: Springer)
- [33] Minka T P 2001 A family of algorithms for approximate Bayesian inference *PhD Thesis* Massachusetts Institute of Technology Cambridge
- [34] Minka T P 2001 Expectation propagation for approximate Bayesian inference *Proc. 17th Conf. Uncertainty in Artificial Intelligence* ed J S Breese and D Koller pp 362–9
- [35] Phillips D L 1962 A technique for the numerical solution of certain integral equations of the first kind *J. Assoc. Comput. Mach.* **9** 84–97

- [36] Rasmussen C E 2004 Gaussian processes in machine learning *Advanced Lectures on Machine Learning (Lecture Note in Computer Science* vol 3176) ed O Bousquet *et al* (Heidelberg: Springer) pp 63–71
- [37] Robert C P and Casella G 2004 *Monte Carlo Statistical Methods* 2nd edn (New York: Springer)
- [38] Rudin L I, Osher S and Fatemi E 1992 Nonlinear total variation based noise removal algorithms *Physica D* **60** 259–68
- [39] Seeger M W 2008 Bayesian inference and optimal design for the sparse linear model *J. Mach. Learn. Res.* **9** 759–813
- [40] Shepp L A and Vardi Y 1982 Maximum likelihood reconstruction for emission tomography *IEEE Trans. Med. Imaging* **1** 113–22
- [41] Soththivirat S and Fessler J A 2002 Image recovery using partitioned-separable paraboloidal surrogate coordinate ascent algorithms *IEEE Trans. Imaging Proc.* 11 306–17
- [42] Stuart A M 2010 Inverse problems: a Bayesian perspective *Acta Numer.* **19** 451–559
- [43] Tierney L and Kadane J B 1986 Accurate approximations for posterior moments and marginal densities *J. Am. Stat. Assoc.* **81** 82–6
- [44] Vardi Y, Shepp L and Kaufman L 1985 A statistical model for positron emission tomography *J. Am. Stat. Assoc.* **80** 8–20
- [45] Vogel C R and Oman M E 1996 Iterative methods for total variation denoising *SIAM J. Sci. Comput.* **17** 227–38
- [46] Wang J and Zabaras N 2005 Hierarchical Bayesian models for inverse problems in heat conduction *Inverse Problems* **21** 183–206
- [47] Welling M and Teh Y W 2011 Bayesian learning via stochastic gradient Langevin dynamics *Proc. of the 28th Int. Conf. on Machine Learning* pp 681–8
- [48] Zhang C, Butepage J, Kjellstrom H and Mandt S 2019 Advances in variational inference *IEEE Trans. Pattern Anal. Mach. Intell.* in press