Variational Gaussian Approximation for Poisson Data

Chen Zhang

Department of Computer Science
University College London

Joint work with Bangti Jin, Simon R. Arridge and Kazufumi Ito

Computational Strategies for Large-scale Statistical Data Analysis Workshop
ICMS, Edinburgh, 3 July

1. Introduction

2. Bayesian formulation

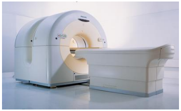
3. Theoretical analysis

4. Numerical algorithm

5. Hierarchical Bayes for hyperparameter

6. Numerical experiments

From ECT to Poisson models

Patient →  → Measurement

Photon numbers x → $g(x)$ → Detected numbers y

Probabilistic models

$$p(y_i|x) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \quad \lambda_i = g_i(x)$$

- Transmission tomography: $g_i(x) = b_i e^{-[Ax]_i} + r_i$
- Emission tomography: $g_i(x) = [Ax]_i + r_i$

Poisson regression: a simplified model

Poisson intensity (simplified version)

- $\lambda_i = e^{(a_i, x)}, \; i = 1, \ldots, n$

Unknown

- $x = [x_1, x_2, \ldots, x_m]^t \in \mathbb{R}^m$

Known

- $A = [a_i^t]_{i=1}^n \in \mathbb{R}^{n \times m}$
- $y = [y_1, y_2, \ldots, y_n]^t \in \mathbb{R}^n$

Likelihood function

$$p(y|x) = \prod_{i=1}^{n} p(y_i|x) = \exp[(Ax, y) - (e^{Ax}, 1_n) - (\ln(y!), 1_n)]$$

Bayesian formulation

Gaussian prior assumption on x

$$p(x) = \mathcal{N}(x; \mu_0, C_0).$$

Posterior distribution by Bayes' formula

$$p(x|y) = \frac{1}{Z}\exp[(Ax, y) - (e^{Ax}, 1_n) - (\ln(y!), 1_n) - \tfrac{1}{2}(x - \mu_0)^{\mathsf{T}} C_0^{-1}(x - \mu_0)],$$

where $Z = Z(y) = \int_{\mathbb{R}^m} p(x, y)\mathrm{d}x$ is the normalising constant and makes the posterior distribution intractable!

Variational inference: a quick review

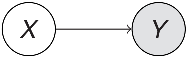


Figure: The hidden variable X and the observable variable Y

By solving a variational problem

$$q(x) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(x) || p(x|y))$$

we find

tractable $\quad q(x) \approx p(x|y) \quad$ intractable

KL divergence

$$KL(q(x)||p(x)) = \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x \tag{1}$$

A probabilistic metric

- $\geqslant 0$ (by Jensen's inequality)
- $\equiv 0$ if and only if $q(x) = p(x)$ almost everywhere

Z in $p(x|y)$ is unknown

$$q^*(x) = \arg\min_{q(x)\in\mathcal{Q}} \mathrm{KL}(q(x)||p(x|y)), \tag{2}$$

still intractable!

ELBO

Key observation

$$\underbrace{\log Z}_{\text{fixed!}} = \int q(x) \log \frac{p(x,y)}{q(x)} \mathrm{d}x + \underbrace{\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x}_{\text{KL, } \geqslant 0}$$

Evidence Lower BOund (ELBO)

$$F(q(x), p(x,y)) = \int q(x) \log \frac{p(x,y)}{q(x)} \mathrm{d}x$$

Equivalent problem

$$\arg\min_{q(x)\in\mathcal{Q}} \mathrm{KL}(q(x)||p(x|y)) = \arg\max_{q(x)\in\mathcal{Q}} F(q(x), p(x,y))$$

finally tractable!

ELBO: as a regularisation

- ELBO

$$
\begin{aligned}
F(q(x), p(x,y)) &= \int q(x) \log \frac{p(x,y)}{q(x)} \mathrm{d}x \\
&= \int q(x) \log \frac{p(y|x)p(x)}{q(x)} \mathrm{d}x \\
&= \underbrace{\int q(x) \log p(y|x) \mathrm{d}x}_{\text{model fitting}} - \underbrace{\int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x}_{\text{prior penalty}}
\end{aligned}
$$

- Tikhonov regularisation

$$
F(x) = \underbrace{\phi(f(x), y)}_{\text{original functional}} + \underbrace{\alpha\psi(x)}_{\text{regulariser}}
$$

Explicit formula of ELBO

from variation to optimisation

$$\begin{aligned} F(q(x),p(x,y)) &= \underbrace{(y, A\bar{x}) - (1_n, e^{A\bar{x}+\frac{1}{2}\mathrm{diag}(ACA^{\mathsf{T}})}) - (1_n, \ln(y!))}_{\text{model fitting}} \\ &\quad - \frac{1}{2}\underbrace{(\bar{x}-\mu_0)^{\mathsf{T}} C_0^{-1}(\bar{x}-\mu_0)}_{\text{weighted distance } ||\bar{x}-\mu_0||^2_{C_0}} \\ &\quad - \frac{1}{2}\underbrace{[\mathrm{tr}(C_0^{-1}C) - \ln|C| + \ln|C_0| - m]}_{\text{Bregman divergence } D(C,C_0)} =: F(\bar{x}, C). \end{aligned} \tag{3}$$

Theoretical properties

existence and uniqueness

Theorem

The lower bound $F(\bar{x}, C)$ is strictly joint-concave with respect to $\bar{x} \in \mathbb{R}^m$ and $C \in \mathcal{S}_m^+$.

Theorem

For any A, y, μ_0 and C_0, there exists a unique pair of $(\bar{x}, C)$ solving the optimisation problem

$$\max F(\bar{x}, C) \tag{4}$$

Optimality system

$$\max_{\bar{x},C} F(\bar{x},C) \tag{5}$$

whose optimality conditions are

$$\frac{\partial F}{\partial \bar{x}} = 0 \quad \text{and} \quad \frac{\partial F}{\partial C} = 0. \tag{6}$$

Theorem

The gradients of $F(\bar{x},C)$ with respect to $\bar{x}$ and C are respectively given by

$$\frac{\partial F}{\partial \bar{x}} = A^t y - A^t e^{A\bar{x}+\frac{1}{2}diag(ACA^t)} - C_0^{-1}(\bar{x}-\mu_0),$$
$$\frac{\partial F}{\partial C} = \tfrac{1}{2}[-A^t diag(e^{A\bar{x}+\frac{1}{2}diag(ACA^t)})A - C_0^{-1} + C^{-1}].$$

An alternating optimisation scheme

Optimality system

The necessary and sufficient optimality system is given by

$$A^t y - A^t e^{A\bar{x} + \frac{1}{2}\mathrm{diag}(ACA^t)} - C_0^{-1}(\bar{x} - \mu_0) = 0 \tag{7}$$

$$\frac{1}{2}[-A^t \mathrm{diag}(e^{A\bar{x} + \frac{1}{2}\mathrm{diag}(ACA^t)})A - C_0^{-1} + C^{-1}] = 0 \tag{8}$$

To solve the optimal system, we designed an alternating direction algorithm based on Equation 7 and 8 seperately.

x Step: Newton method

Consider $-\frac{\partial F}{\partial \bar{x}}$

$$\mathbf{G}(\bar{x}) = A^t e^{A\bar{x}+\frac{1}{2}\mathrm{diag}(ACA^t)} + C_0^{-1}(\bar{x}-\mu_0) - A^t y.$$

Uniform invertibility

$$\partial\mathbf{G}(\bar{x}) = A^t \mathrm{diag}(e^{A\bar{x}+\frac{1}{2}\mathrm{diag}(ACA^t)})A + C_0^{-1} \geqslant C_0^{-1},$$

Newton update scheme

$$\partial\mathbf{G}(\bar{x}^k)\delta\bar{x} = -\mathbf{G}(\bar{x}^k), \qquad \bar{x}^{k+1} = \bar{x}^k + \delta\bar{x}. \tag{9}$$

Globally convergent!

C Step: Fixed point method

Based on $\left(\frac{\partial F}{\partial C} = 0\right)$

$$C^{-1} = A^{\mathsf{T}}\mathrm{diag}\left(e^{A\bar{x}+\frac{1}{2}\mathrm{diag}\left(ACA^{\mathsf{T}}\right)}\right)A + C_0^{-1}$$

we iterate

$$C^{k+1} = \left(C_0^{-1} + A^t D^k A\right)^{-1}, \quad \text{with} \quad D^k = \mathrm{diag}\left(e^{A\bar{x}+\frac{1}{2}\mathrm{diag}\left(AC^k A^t\right)}\right)$$

Uniformly bounded sequence $\{C^k\}_{k=0}^{\infty}$

$$\lambda_{\max}(C^k) = v_*^t C^k v_* \leqslant v_*^t C_0 v_* \leqslant \sup_{v\in\mathbb{R}^m} v^t C_0 v = \lambda_{\max}(C_0)$$

Sub-sequentially convergent![1]

[1] Another interesting 'monotone' type of convergence is also discussed in our paper

Computational complexity reduction

Structural assumptions

- C – k sparsity
  - Banded matrix with band width k or
  - At most k non-zero elements each row
- A – r sparsity
  - Low rank approximation $A_r \approx A$ $(r \ll m \wedge n)$

Table: Computational cost comparisons

Operation	General case	Structural assumptions
x step	$\mathcal{O}(m^3 + m^2 n)$	$\mathcal{O}(m^2 + kmn)$
C step	$\mathcal{O}(m^3 + m^2 n)$	$\mathcal{O}(r^2 n + r^2 m + kmn)$

Algorithm 1 Variational Gaussian Approximation Algorithm

```
Input: $(A, y)$, specify the prior $(\mu_0, C_0)$, and the maximum number $K$ of iterations
Initialize $\bar{x} = \bar{x}^1$ and $C = C^1$;
SVD: $(U, \Sigma, V) = \mathrm{rSVD}(A)$;
for $k = 1, 2, \ldots, K$ do
    Update the mean $\bar{x}^{k+1}$ by Newton method;
    Update the covariance $C^{k+1}$ by fixed point method;
    Check the stopping criterion.
end for
Output: $(\bar{x}, C)$
```

Hyperparamter choice

In the Gaussian prior $p(x)$, $C_0 = \alpha^{-1}\bar{C}_0$.

$$\alpha(\bar{x} - \mu_0)^\mathsf{T}\bar{C}_0^{-1}(\bar{x} - \mu_0) = \alpha||L(\bar{x} - \mu_0)||^2,$$

where $\bar{C}_0^{-1} = L^t L$.

- $\bar{C}_0$ encodes smoothness into prior (interactive strucutre)
- α determines the strength of the interaction

How to determine α?

Hierarchical model and joint ELBO

Hyperprior distribuion

- $p(\alpha|a,b) = \mathrm{Gamma}(\alpha|a,b)$
- Noninformative settings: $a \approx 1$ and $b \approx 0$

Joint lower bound

$$
\begin{aligned}
F(\bar{x}, C, \alpha) = (y, A\bar{x}) - (1_n, e^{A\bar{x} + \frac{1}{2}\mathrm{diag}(ACA^t)}) - \tfrac{\alpha}{2}(\bar{x} - \mu_0)^t \bar{C}_0^{-1}(\bar{x} - \mu_0) \\
- \tfrac{\alpha}{2}\mathrm{tr}(\bar{C}_0^{-1} C) + \tfrac{1}{2}\ln|C| + \tfrac{m}{2}\ln\alpha - \tfrac{1}{2}\ln|\bar{C}_0| \\
+ (a-1)\ln\alpha - \alpha b + \tfrac{m}{2} - (1_n, \ln(y!)) + \ln\frac{b^a}{\Gamma(a)}.
\end{aligned}
$$

EM algorithm for joint ELBO optimisation

- E-step: fix α, and maximize $F(\bar{x}, C, \alpha)$ by Algorithm 1.
- M-step: fix $(\bar{x}, C)$ and update α by

$$\alpha = \frac{m + 2(a-1)}{(\bar{x}_\alpha - \mu_0)^t \bar{C}_0^{-1} (\bar{x}_\alpha - \mu_0) + \operatorname{tr}(\bar{C}_0^{-1} C_\alpha) + 2b}. \tag{10}$$

An extension of a balancing principle in Tikhonov regularisation

$$E_{q(x)}[\log p(x)] = \alpha[(\bar{x}_\alpha - \mu_0)^t \bar{C}_0^{-1} (\bar{x}_\alpha - \mu_0) + \operatorname{tr}(\bar{C}_0^{-1} C_\alpha)],$$

Algorithm 2 Hierarchical variational Gaussian approximation

```
Input $(A, y)$, and initialize $\alpha^1$
for $k = 1, 2, \ldots$ do
    E-step: Update $(\bar{x}^k, C^k)$ by Algorithm 1:
        $(\bar{x}^k, C^k) = \arg \max_{(\bar{x}, C) \in \mathbb{R}^m \times \mathcal{S}_m^+} F_{\alpha^k}(\bar{x}, C);$
    M-step: Update $\alpha$ by (10).
    Check the stopping criterion;
end for
Output: $(\bar{x}, C)$
```

Monotonic convergence

Theorem

For any initial guess $\alpha^1 > 0$, the sequence $\{\alpha^k\}$ generated by Algorithm 2 is monotonically convergent to some $\alpha^* \geqslant 0$, and if the limit $\alpha^* > 0$, then it satisfies the fixed point equation (10)*.*

Remarks

- The uniqueness of the solution α^* to (10) is generally not ensured.
- In practice, it seems to have only two fixed points: one is in the neighborhood of $+\infty$, which is uninteresting, and the other is the desired one.

`Phillips` test

an example from package `Regutools`[2]

Fredholm integral Eq Galerkin discretisation linear system

$$\int K(s,t)f(t)\mathrm{d}t = g(s) \quad \longrightarrow \quad Ax = b$$

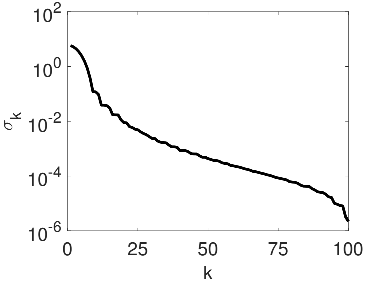
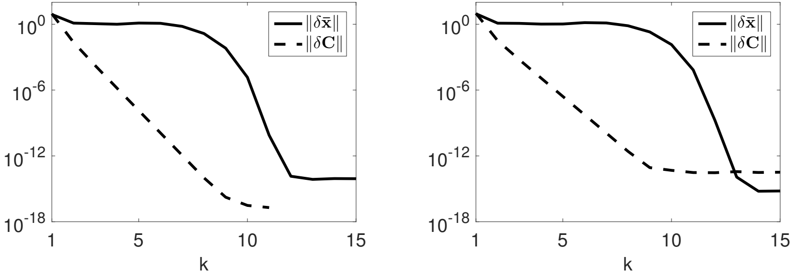


Figure: Ill-posedness reflexed by singular value decay of A

[2] www.imm.dtu.dk/ pcha/Regutools/

Empirical Inner Convergence



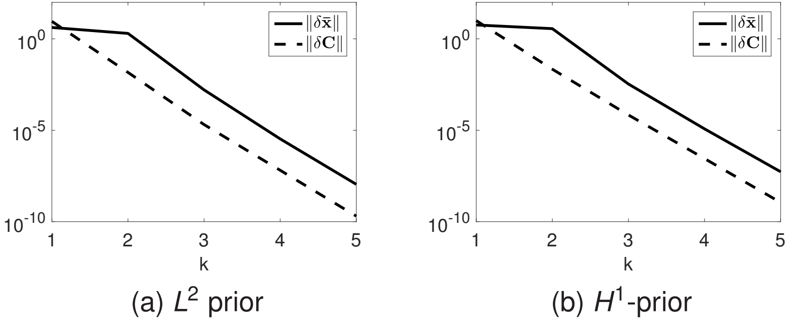
(a) L^2-prior

(b) H^1-prior

Figure: The convergence of the inner iterations of Algorithm 1 for `phillips`.

[1] $\delta\bar{x} = \bar{x}_{k+1} - \bar{x}_k$ and $\delta C = C_{k+1} - C_k$

Empirical Outer Convergence



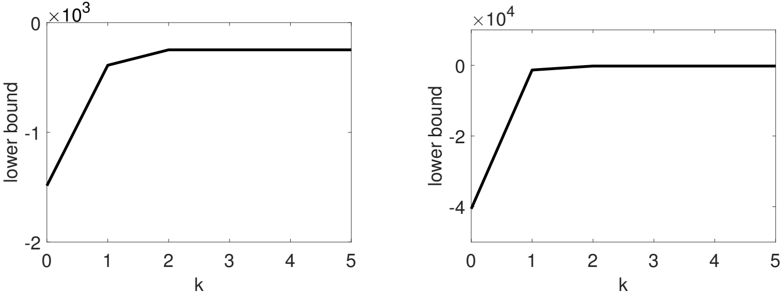
(a) L^2 prior

(b) H^1-prior

Figure: The convergence of outer iterations of Algorithm 1 for `phillips`.

[1] $\delta\bar{x} = \bar{x}_{k+1} - \bar{x}_k$ and $\delta C = C_{k+1} - C_k$

Empirical ELBO Convergence

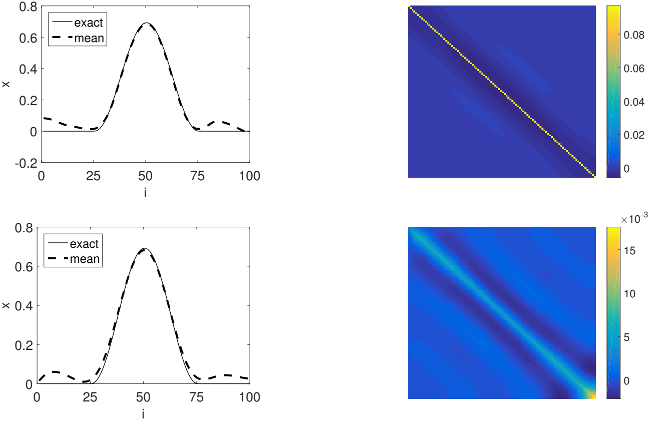


(a) L^2-prior

(b) H^1-prior

Figure: The convergence of the lower bound $F(\bar{\mathbf{x}}, C)$ for `phillips`.

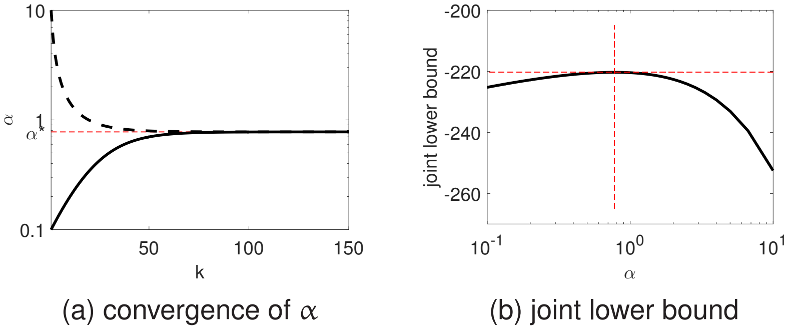
Singal reconstructions



(Upper) $C_0 = 1.00 \times 10^{-1}\bar{\mathbf{C}}_0$ (Lower) $C_0 = 2.5 \times 10^{-3}\bar{\mathbf{C}}_1$

Figure: The Gaussian approximation for `phillips`.

Hierarchical parameter convergence



(a) convergence of α

(b) joint lower bound

Figure: (a)The convergence of Algorithm 2 initialized with 0.1 and 10, both convergent to $\alpha^* = 0.7778$ (b) the joint lower bound versus α, for `phillips` with L^2-prior.

Hierarchical reconstructions

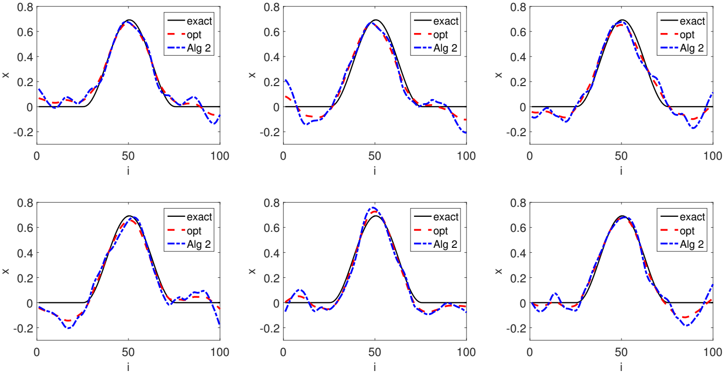
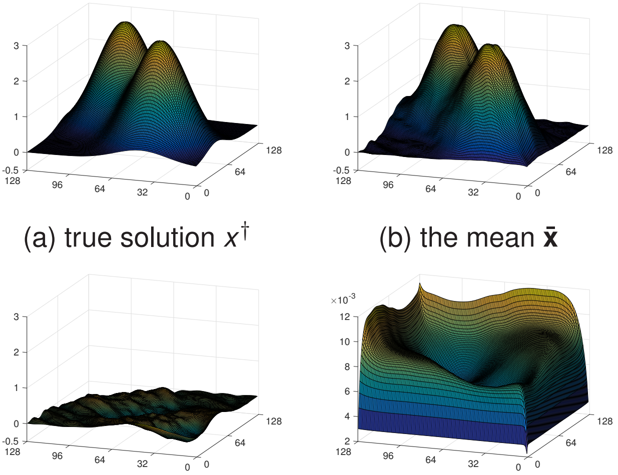


Figure: The mean $\bar{\mathbf{x}}$ of the Gaussian approximation by the hierarchical algorithm (Alg2) and the “optimal” solution (opt) for 6 realizations of Poisson data for `phillips` with the L^2-prior.

A large scale example of Gaussian deblurring



(a) true solution $x^\dagger$

(b) the mean $\bar{\mathbf{x}}$

(c) the error $x^\dagger - \bar{\mathbf{x}}$

(d) the variance $\mathrm{diag}(C)$

Main contributions

ELBO

- Explicit expression
- Existence and uniqueness

Numerical algorithm

- Alternating direction maximisation algorithm
- Convergence
- Computational complexity reduction strategies

Hyperparameter

- Discuss hierarchical Bayesian modelling
- Monotonical convergence

Main references

Inverse problems

- Ito, K. and Jin, B., 2015. Inverse problems: Tikhonov theory and algorithms.
- Stuart, A.M., 2010. Inverse problems: a Bayesian perspective. Acta Numerica, 19, pp.451-559.

Emission/Transmission tomography

- Erdogan, H. and Fessler, J.A., 2002, June. Monotonic algorithms for transmission tomography. In Biomedical Imaging, 2002. 5th IEEE EMBS International Summer School on (pp. 14-pp). IEEE.
- Yavuz, M. and Fessler, J.A., 1997, June. New statistical models for randoms-precorrected PET scans. In Biennial International Conference on Information Processing in Medical Imaging (pp. 190-203). Springer, Berlin, Heidelberg.

Main references

Bayesian inference

- Wainwright, M.J. and Jordan, M.I., 2008. Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 1(12), pp.1-305.
- Challis, E. and Barber, D., 2013. Gaussian kullback-leibler approximate inference. The Journal of Machine Learning Research, 14(1), pp.2239-2286.
- Blei, D.M., Kucukelbir, A. and McAuliffe, J.D., 2017. Variational inference: A review for statisticians. Journal of the American Statistical Association, 112(518), pp.859-877.

Our paper

- Arridge, S.R., Ito, K., Jin, B. and Zhang, C., 2018. Variational Gaussian approximation for Poisson data. Inverse Problems, 34(2), p.025005.[3]

[3] The work of mine is fully funded by UCL Department of Computer Science