Variational Approximations for Frequentist and Bayesian Inference

Approssimazioni Variazionali per Inferenza Frequentista e Bayesiana

Luca Maestrini and Matt P. Wand

Abstract Variational approximations are a flexible instrument for deterministic approximate inference in complex statistical models. We illustrate the concept of variational approximation from both frequentist and Bayesian perspectives, providing methodological examples that take advantage of the classical concepts of exponential families.

Sommario *Le approssimazioni variazionali sono uno strumento flessibile per l'inferenza su modelli statistici complessi. Viene illustrato il concetto di approssimazione variazionale da punti di vista frequentista e Bayesiano, proponendo esempi metodologici che fanno leva sulla classica teoria delle famiglie esponenziali.*

Key words: Bayesian inference, frequentist inference, Gaussian variational approximation, variational message passing.

1 Introduction

Variational approximations is a class of techniques for deterministic approximations which is now part of mainstream computer science and machine learning methodology. Applications cover a wide area of elaborate problems such as those arising in speech recognition, graphical models, document retrieval or genetic linkage analysis [2]. These methods are also widening their presence in statistics as a response to the increasing complexity of models in modern statistical applications [4].

We describe the concept of variational approximation referring to a Bayesian model. In keeping with the statistics literature on variational approximations, let p

Luca Maestrini
Department of Statistical Sciences, University of Padova, Italy
e-mail: luca.maestrini@phd.unipd.it

Matt P. Wand
School of Mathematical and Physical Sciences, University of Technology Sydney, Australia

be the generic symbol for a density function, denote with $\mathbf{y}$ the observed data, $\theta \in \Theta$ the parameters to be inferred and let q be an arbitrary density function over Θ.

The logarithm of the marginal likelihood satisfies

$$\begin{aligned}\log p(\mathbf{y}) &= \int q(\theta) \log\left\{\frac{p(\mathbf{y},\theta)}{q(\theta)}\right\} d\theta + \int q(\theta) \log\left\{\frac{q(\theta)}{p(\theta|\mathbf{y})}\right\} d\theta \\ &\geq \int q(\theta) \log\left\{\frac{p(\mathbf{y},\theta)}{q(\theta)}\right\} d\theta, \end{aligned} \tag{1}$$

giving a lower bound $\underline{p}(\mathbf{y};q)$ on the marginal likelihood such that

$$p(\mathbf{y}) \geq \underline{p}(\mathbf{y};q) \equiv \exp \int q(\theta) \log\left\{\frac{p(\mathbf{y},\theta)}{q(\theta)}\right\} d\theta. \tag{2}$$

Maximization of $\underline{p}(\mathbf{y};q)$ is equivalent to minimization of the Kullback–Leibler divergence between $q(\theta)$ and $p(\theta|\mathbf{y})$. The key idea of variational approximations is to approximate the posterior density $p(\theta|\mathbf{y})$, or the likelihood function itself in the frequentist case, by a $q(\theta)$ for which $\underline{p}(\mathbf{y};q)$ is more tractable than $p(\mathbf{y})$ and obtain approximate estimates through lower bound maximization. Tractability is achieved by restricting $q(\theta)$ to a more manageable class of densities. Common restrictions for the approximating density are:

a. $q(\theta)$ is a member of a parametric family of density functions;
b. $q(\theta)$ factorizes into $\prod_{i=1}^{M} q_i(\theta_i)$, for some partition $\{\theta_1, \ldots, \theta_M\}$ of θ.

We apply restrictions (a) and (b) to describe frequentist and Bayesian methodologies respectively which are known as Gaussian variational approximation and variational message passing.

2 Gaussian Variational Approximation

Frequentist models that stand to benefit from variational approximations are those for which the likelihood specification involves conditioning on a vector of latent variables $\mathbf{u}$. Given a log-likelihood of the model parameter vector θ

$$\ell(\theta) \equiv \log p(\mathbf{y};\theta) = \log \int p(\mathbf{y}|\mathbf{u};\theta)\, p(\mathbf{u};\theta)\, d\mathbf{u},$$

interest is in obtaining $\hat{\theta} = \underset{\theta}{\text{argmax}}\, \ell(\theta)$, maximum likelihood estimate of θ.

In practice, $\ell(\theta)$ may not be available in closed form because of analytically intractable integration. In such circumstances and depending on the forms of $p(\mathbf{y}|\mathbf{u};\theta)$ and $p(\mathbf{u};\theta)$, variational approximations can provide a more amenable approximation. However, nontrivial frequentist examples where an explicit solution arises by applying a product density methodology as in (b) are not known.

Suppose instead to restrict q to a parametric family of densities $\{q(\mathbf{u};\xi) : \xi \in \boldsymbol{\Xi}\}$, similarly to (a). Then, similarly to (1) we can define the log-likelihood lower bound

$$\underline{\ell}(\theta,\xi;q) \equiv \int q(\mathbf{u};\xi) \log\left\{\frac{p(\mathbf{y},\mathbf{u};\theta)}{q(\mathbf{u};\xi)}\right\} d\mathbf{u}$$

and a new maximization problem

$$\left(\underline{\hat{\theta}},\underline{\hat{\xi}}\right) = \underset{\theta,\xi}{\operatorname{argmax}} \underline{\ell}(\theta,\xi;q),$$

with $\underline{\hat{\theta}}$ variational approximation to the maximum likelihood estimator. Furthermore, standard error estimates can be obtained by plugging in $\underline{\hat{\theta}}$ for θ and $\underline{\hat{\xi}}$ for ξ in the variational approximate Fisher information matrix arising from replacement of $\ell(\theta)$ by $\underline{\ell}(\theta,\xi;q)$.

In *Gaussian variational approximations* (GVA), $q(\mathbf{u};\xi)$ is assumed to be a multivariate normal density [5]. We investigate the application of GVA to generalized linear mixed models (GLMMs) for semiparametric regression.

Consider GLMMs within one-parameter exponential family

$$\mathbf{y}|\mathbf{u} \sim \exp\left\{\mathbf{y}^{\mathrm{T}}(\mathbf{X}\beta+\mathbf{Z}\mathbf{u}) - \mathbf{1}^{\mathrm{T}} b(\mathbf{X}\beta+\mathbf{Z}\mathbf{u}) + \mathbf{1}^{\mathrm{T}} c(\mathbf{y})\right\}, \quad \mathbf{u} \sim N(\mathbf{0},\mathbf{G})$$

where $\mathbf{X}$ and $\mathbf{Z}$ are general design matrices. Matrix $\mathbf{G}$ models random effect covariance, while $\mathbf{Z}$ can include, for instance, spline basis functions. The functions b and c characterize members of the exponential family.

Setting $q(\mathbf{u};\xi)$ to be the $N(\mu,\boldsymbol{\Lambda})$ we can derive variational lower bound

$$\begin{aligned}\underline{\ell}(\beta,\mathbf{G},\mu,\boldsymbol{\Lambda}) = \tfrac{n}{2} + \mathbf{y}^{\mathrm{T}}(\mathbf{X}\beta+\mathbf{Z}\mu) - \mathbf{1}^{\mathrm{T}} B\left(\mathbf{X}\beta+\mathbf{Z}\mu, \operatorname{dg}\left(\mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^{\mathrm{T}}\right)\right) + \mathbf{1}^{\mathrm{T}} c(\mathbf{y}) \\ -\tfrac{1}{2}\left\{\mu^{\mathrm{T}}\mathbf{G}^{-1}\mu + \operatorname{tr}\left(\mathbf{G}^{-1}\boldsymbol{\Lambda}\right)\right\} + \tfrac{1}{2}\log\left|\mathbf{G}^{-1}\boldsymbol{\Lambda}\right|, \qquad (3)\end{aligned}$$

where n is the number of rows in $\mathbf{y}$, $B\left(\mu,\sigma^2\right) \equiv \int_{-\infty}^{\infty} b(\mu+\sigma x)\,\phi(x)\,dx$, $\phi(x)$ is the $N(0,1)$ density function and, for a square matrix $\mathbf{A}$, $\operatorname{dg}(\mathbf{A})$ is the column vector containing the diagonal entries of $\mathbf{A}$. For vector arguments, function B is applied in element-wise fashion. Inference and prediction on nonparametric, additive or general semiparametric models in GLMM form follow directly from the lower bound optimization.

3 Variational Message Passing

In Bayesian inference, a mean field variational approximation $q^*(\theta)$ is the maximizer of expression (2) subject to a product density restriction as in (b).

It can be shown that the optimal q-density functions satisfy

$$q^*(\theta_i) \propto E_{q(\theta\backslash\theta_i)}\{p(\theta_i|\mathbf{y},\theta\backslash\theta_i)\}, \quad 1 \le i \le M, \qquad (4)$$

where $\theta \backslash \theta_i$ denotes the entries of θ with θ_i omitted. Expression (4) gives rise to an iterative scheme for obtaining the parameters of the optimal density functions $q^*(\theta_i)$ which is known as mean field variational Bayes. *Variational message passing* (VMP) arrives at the same approximation via message passing on an appropriate factor graph. Among the several variants of VMP in the literature, we consider the factor graph fragment approach introduced in [8] and based on [3], whose major advantage is that calculations only need to be done once for a certain distribution family and can be easily adapted to accommodate more complex model structures. The use of conjugates exponential families streamlines the algebraic and computational effort in deriving messages between factor graph components at the base of VMP algorithms. A listing of such a procedure can be found in Sect. 2.5 of [8].

This framework gives rise to a class of VMP algorithms to approximate fitting and inference for a wide range of common and non-standard likelihoods.

4 Illustrations

The next two illustrative examples are applications in frequentist and Bayesian settings that witness the flexibility of variational approximations. The former provides approximate estimates for a simulated Poisson spline regression model, the latter concerns a skew t response regression model on real data.

4.1 GVA for Poisson Spline Regression

We simulate 500 observations from a Poisson process as a function of a Uniform$(0,1)$ covariate and estimate a Poisson semiparametric regression model with canonical link using O'Sullivan penalized splines [6] on 50 interior knots via GVA.

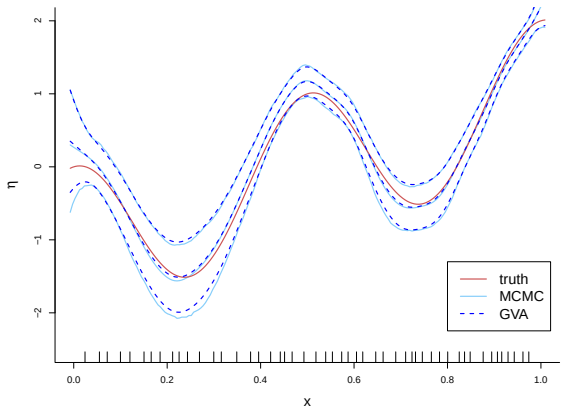

Fig. 1 Data generating process for the Poisson semiparametric regression. 95% MCMC credible intervals and GVA confidence bands are compared to the true η function that generates data according to a Poisson(e^{η}) distribution. Knot positions are also displayed.

Let $\hat{\boldsymbol{\Lambda}}_{\text{GVA}}$ be the estimate of $\boldsymbol{\Lambda}$ obtained maximizing the Gaussian variational lower bound (3) adapted to the current model. Given $\text{H}_{\mu\mu}\underline{\ell}$ Hessian matrix of $\underline{\ell}(\beta, \mathbf{G}, \mu, \boldsymbol{\Lambda})$ with respect to μ, one can prove that $\hat{\boldsymbol{\Lambda}}_{\text{GVA}} = \left(-\text{H}_{\mu\mu}\underline{\ell}\right)^{-1}$. We use this result and the estimates obtained through the lower bound optimization to derive the plot in Fig. 1, which concerns the true generating process. For a rough performance evaluation we plot GVA results as in (2) with those from Markov chain Monte Carlo (MCMC) samples obtained using the `R` package `rstan` [7] setting priors $N\left(0, 10^5\right)$ on β and Half-Cauchy$\left(10^5\right)$ on the scale parameter appearing from defining $\mathbf{G} \equiv \sigma^2\mathbf{I}$. GVA seems to adequately approximate the MCMC process prediction and credible intervals.

4.2 VMP for Skew t Regression

We illustrate the parameter estimation of a skew t regression model via VMP.

Consider the dataset examined in [1] with the linear model

$$y_i = \beta_0 + \beta_1 \text{CRSP}_i + \varepsilon_i, \quad \varepsilon_i \sim \text{Skew-t}\left(0, \sigma^2, \lambda, \nu\right), \quad 1 \leq i \leq 60$$

with λ parameter of symmetry and $\nu > 0$ degrees of freedom. The variables y_i and CRSP_i denote the Martin Marietta company excess rate and the return excess index for the New York Stock Exchange respectively. We adopt the skew t distribution described in [1] and write it in terms of standard normal and inverse χ^2 auxiliary variables to limit the complexity of algebraic derivations and numerical integration appearing in the derivation of a VMP algorithm. We choose a product density restriction on $q(\theta)$ which is a compromise between approximation performances and algebraic complexity. We approximate the parameter posterior densities with VMP

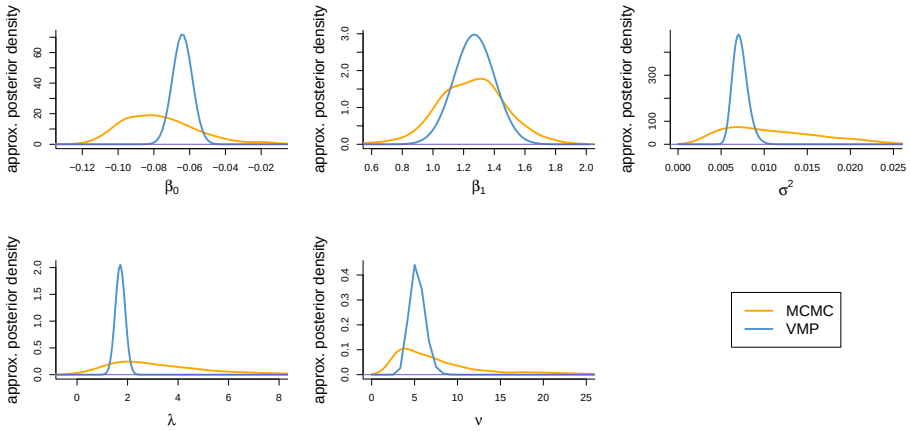

Fig. 2 Martin Marietta data: posterior density plots via MCMC and VMP.

and compare them to MCMC density estimation via `rstan`. The hyperparameters for β are fixed to $\mu_\beta = \mathbf{0}$ and $\boldsymbol{\Sigma}_\beta = 10^5\mathbf{I}$ over a prior $N\left(\mu_\beta, \boldsymbol{\Sigma}_\beta\right)$ while those on the shape parameters are Inverse-$\chi^2(0.01, 0.01)$ on the squared scale, $N\left(0, 10^5\right)$ on λ and $\Gamma(1, 0.01)$ on ν. Posterior density plots are shown in Fig. 2. VMP curves apparently underestimate the variance of MCMC posterior densities but locate around their modes.

Acknowledgements Luca Maestrini carried out his research activity during a visiting period at the School of Mathematical and Physical Sciences, University of Technology Sydney, Australia. We are grateful to Alessandra Salvan and Nicola Sartori for their comments on this research.

References

1. Azzalini, A., Capitanio, A.: Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. J. R. Stat. Soc. Ser. B. **65**, 367–389 (2003)
2. Jordan, M. I.: Graphical Models. Stat. Sci. **19**, 140–155 (2004)
3. Minka, T.: Divergence measures and message passing. Microsoft Res. Tech. Rep. Ser. **173**, 1–17 (2005)
4. Ormerod, J. T., Wand, M. P.: Explaining variational approximations. Am. Stat. **64**, 140–153 (2010)
5. Ormerod, J. T., Wand, M. P.: Gaussian variational approximate inference for generalized linear mixed models. J. Comput. Graph. Stat. **21**, 2–17 (2012)
6. O'Sullivan, F.: Nonparametric estimation of relative risk using splines and cross-validation. J. Sci. Stat. Comput. **9**, 531–542 (1988)
7. Stan Development Team: rStan: the R interface to Stan. R package version 2.17.3. http://mc-stan.org/ (2018)
8. Wand, M. P.: Fast approximate inference for arbitrarily large semiparametric regression models via message passing. J. Am. Stat. Assoc. **112**, 137–168 (2017)