BAYESCOMP 2018

26-28 March 2018, Barcelona, Spain.

Approximate Likelihood functions

Nancy Reid

March 27, 2018

University of Toronto

Table of contents

Introduction

Models and likelihood

- Model for the probability distribution of y — given x
- Density $f(y \mid x)$ — with respect to, e.g., Lebesgue measure
- Parameters for the density $f(y \mid x; \theta)$, — $\theta = (\theta_1, \dots, \theta_d)$
- Data $y = (y_1, \dots, y_n)$ — often independent

- Likelihood function $L(\theta; y) \propto f(y; \theta)$ — $(y_1, \dots, y_n)$
- log-likelihood function $\ell(\theta; y) = \log L(\theta; y)$

- intractable log-likelihood function, e.g.

$$\ell(\theta; y) = \log \int f(y \mid z; \theta) f(z) dz$$

$y \in \mathbb{R}^k, z \in \mathbb{R}^q$

- z is a set of latent variables — possibly $f(z; \tau)$

Intractable log-likelihood functions

- latent Gaussian model

Rue et al. 2017

$$f(y \mid z; \theta) = \prod_{i \in \mathcal{I}} f(y_i \mid z_i; \theta),$$

$$z \sim N\{\mu(\tau), \Sigma(\tau)\},$$

$$L(\theta; y) = \int f(y \mid z; \theta) f(z; \tau) \pi(\tau) dz d\tau, \quad \pi(\theta \mid y) \propto L(\theta; y)\pi(\theta)$$

- latent variable model

Verbeke & Molenberghs, 2017

$$f(y_i \mid z_i; \theta) = g_i(z_i; \theta), i = 1, \ldots, n,$$

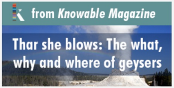
$$\ell(\theta; y) = \sum_{i=1}^{n} \log \int g_i(z_i; \theta) dQ(z_i) dz_i$$

$\ell = \log L$

Q nonparametric

Annual Review of Statistics and Its Application

View full Current Table of Contents



About This Journal

The *Annual Review of Statistics and Its Application* informs statisticians, and users of statistics about major methodological advances and the computational tools that allow for their implementation.

The *Annual Review of Statistics and Its Application* debuted in the 2016 Release of the Journal Citation Report (JCR) with an Impact Factor of **3.045**.

View more information, including the complete list of journal Impact Factors and category rankings here.

... intractable log-likelihood functions

- generalized linear mixed models — Ogden, 2017
- model defined by differential equations — Papavasiliou & Taylor 2016
- exponential random graph models — e.g. $\exp\{\Sigma\theta_{jk}y_jy_k - \log Z(\theta)\}$
- multivariate extreme value processes

  $F(y;\theta) = \exp\{-V(y_1,\dots,y_d;\theta)\}$

  Davison & Huser, 2015
- Gaussian processes on a lattice — Stroud et al., 2017

$$-\frac{1}{2}\{\log|\Sigma(\theta)| + y^{\mathrm{T}}\Sigma^{-1}(\theta)y\}$$

$y = \{y(s_1),\dots,y(s_n)\}$

Approximate likelihood functions

- Laplace approximation to integrals — PQL, INLA
- misspecified
  - composite likelihood — ignore some dependencies
  - indirect likelihood — often based on normal approximation
- variational approximation — K-L lower bound
- simulation — ABC, MCMCML
- change the inference
  - indirect likelihood — several moments
  - quasi-likelihood — first two moments
  - estimating equations

Inference based on likelihood functions

- maximum likelihood estimator $\hat{\theta} = \arg\sup_\theta \log L(\theta; y)$

- score function $s(\theta) = \partial \log L(\theta; y)/\partial\theta$

- observed Fisher information $j(\hat{\theta}) = -\left.\dfrac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\right|_{\hat{\theta}}$

- Wald statistic $Q(\theta) = (\hat{\theta} - \theta)^{\mathrm{T}} j(\hat{\theta})(\hat{\theta} - \theta)$

- likelihood ratio statistic $w(\theta) = 2\{\log L(\hat{\theta}) - \log L(\theta)\}$

$\dot{\sim}\ \chi^2_d$

- in high dimensions might use instead $\log L(\theta; y) - P_\lambda(||\theta||)$

Composite likelihood

Inference based on mis-specified likelihood functions

- maximum likelihood estimator $\hat{\theta} = \arg\sup_\theta \log L(\theta; y)$
- score function $s(\theta) = \partial \log L(\theta; y)/\partial\theta$
- score covariance $J(\theta) = \mathrm{E}\{s(\theta)s^{\mathrm{T}}(\theta)\}$
- sensitivity $H(\theta) = \mathrm{E}\{-\dfrac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\}$
- Godambe information $G(\theta) = H(\theta)\{J(\theta)\}^{-1}H(\theta)$
- Wald statistic $Q(\theta) = (\hat{\theta} - \theta)^{\mathrm{T}} G(\hat{\theta})(\hat{\theta} - \theta)$
- likelihood ratio statistic $w(\theta) \dot{\sim} \sum \lambda_j \chi^2_{1j}$

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2018)
67, *Part* 3, *pp.* 575–598

Segmentation of sea current fields by cylindrical hidden Markov models: a composite likelihood approach

Monia Ranalli and Francesco Lagona,

University of Roma Tre, Italy

Marco Picone

Istituto Superiore per la Protezione e la Ricerca Ambientale, Rome, Italy

and Enrico Zambianchi

'Parthenope' University of Naples and Consorzio Nazionale Interuniversitario per le Scienze del Mare, Rome, Italy

[Received December 2016. Revised July 2017]

- Potts model on neighbourhood structure

$$p(\xi;\rho) = \exp\left\{(\rho/2)\sum_{i=1}^{n}\sum_{j:c_{ij}=1}\xi_i^{\mathrm{T}}\xi_j - \log W(\rho)\right\}$$

C adjacency matrix

- Hidden Markov random field

$$f(z \mid \xi;\theta) = \prod_{i=1}^{n} f(x_i, y_i \mid \xi;\theta) = \prod_{i=1}^{n}\prod_{k=1}^{K} f(x_i, y_i;\theta_k)^{\xi_{ik}}$$

cyclindrical densities on angle x, intensity y

- likelihood function

$$L(\theta,\rho) = \sum_{\xi} f(z \mid \xi;\theta)p(\xi;\rho)$$

- composite likelihood function $cL(\theta,\rho) =$

$$\prod_{A\in\mathcal{A}}\sum_{A} f(z_A, \xi_A;\theta,\rho)$$

- composite likelihood function

$$cL(\theta, \rho) = \prod_{A \in \mathcal{A}} \sum_{A} f(z_A, \xi_A; \theta, \rho)$$

- cover $\mathcal{A}$ of subsets of sites $1, \ldots, n$; e.g. all possible pairs
- and then discard pairs that are not in the neighbourhood structure specified in the Potts model

- estimation using the E-M algorithm
- variance estimation using the bootstrap

- inference tested in simulations
- CL inference typically consistent, not efficient

Laplace approximation

- latent Gaussian model

$$
\begin{aligned}
L(\theta; y) &= \int f(y \mid x; \theta_1) f(x; \theta_2) \pi(\theta_2) dx \\
&= \int \prod_{i \in \mathcal{I}} f(y_i \mid x_i; \theta_1) \varphi\{x_i; \mu_i(\theta_2), \Sigma_i(\theta_2)\} dx_i
\end{aligned}
$$

φ normal density

- Laplace approximation to posterior

$$\pi(\theta \mid y) \propto L(\theta; y)\pi(\theta) \quad \text{or} \quad \pi(\theta, x \mid y)$$

- generalized linear mixed models

Gaussian prior on fixed and random effects

- sparse Markov random field for x

plus some additional computational advances, Martins et al. 2013

- Laplace approximation to posterior

$$\pi(\theta \mid y) \propto L(\theta; y)\pi(\theta) \quad \text{or} \quad \pi(\theta, x \mid y)$$

- captures asymmetry

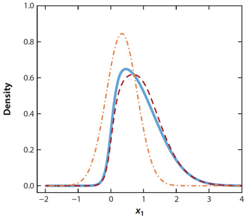


Figure 1

The true marginal (*solid blue line*), the Laplace approximation (*dashed red line*) and the Gaussian approximation (*dot-dashed orange line*).

- good support for spatial modelling

$n = 1$

- "our main concern is how we think about and specify priors "
- "often conceptually difficult to encode prior knowledge"

Simpson et al. 2017

Variational methods

- in a Bayesian context, want $f(\beta \mid y)$, use an approximation $q(\beta)$
- dependence of q on y suppressed

- choose $q(\beta)$ to be
  - simple to calculate
  - close to posterior

- simple to calculate
  - $q(\beta) = \prod q_j(\beta_j)$
  - simple parametric family

- close to posterior: miminize Kullback-Leibler divergence between $q(\cdot)$ and $f(\cdot \mid y)$

- close to posterior: miminize Kullback-Leibler divergence

$$KL(q \,||\, f_{post}) = \int q(\beta) \log\{q(\beta)/f(\beta \mid y)\}d\beta$$

- equivalent to

$$\max_q \int q(\beta) \log\{f(y, \beta)/q(\beta)\}d\beta$$

- because

$$\log f(y; \theta) \geq \int q(\beta) \log\{f(y, \beta; \theta)/q(\beta)\}d\beta$$

- in a likelihood context

$$\log f(y; \theta) = \log \int f(y \mid \beta; \theta) f(\beta) d\beta$$

here β represent random effects u, or b, or ...

log-likelihood:

$$\ell(\beta,\Sigma) = \sum_{i=1}^{m}\Big(y_i^{\mathrm{T}}X_i\beta - \frac{1}{2}\log|\Sigma| + \log\int_{\mathbb{R}^k}\exp\{y_i^{\mathrm{T}}Z_iu_i - 1_i^{\mathrm{T}}b(X_i\beta+Z_iu_i) - \frac{1}{2}u_i^{\mathrm{T}}\Sigma^{-1}u_i\}du_i\Big)$$

variational approx:

$$\ell(\beta,\Sigma) \geq \sum_{i=1}^{m}\Big(y_i^{\mathrm{T}}X_i\beta - \frac{1}{2}\log|\Sigma|\Big) + \sum_{i=1}^{m}E_{u\sim N(\mu_i,\Lambda_i)}\Big(y_i^{\mathrm{T}}Z_iu - 1_i^{\mathrm{T}}b(X_i\beta+Z_iu) - \frac{1}{2}u^{\mathrm{T}}\Sigma^{-1}u - \log\{\phi_{\Lambda_i}(u-\mu_i)\}\Big)$$

simplifies to k one-dim. integrals

$$\ell(\beta, \Sigma) \geq \ell(\beta, \Sigma, \mu, \Lambda)$$

- variational estimate:

$$\ell(\tilde\beta, \tilde\Sigma, \tilde\mu, \tilde\Lambda) = \text{arg max}_{\beta,\Sigma,\mu,\Lambda} \ell(\tilde\beta, \tilde\Sigma, \tilde\mu, \tilde\Lambda)$$

- are $\tilde\beta, \tilde\Sigma$ consistent and/or asymptotically normal?

- Blei et al. (2017) §5.2 Theory
  - Bayesian linear model — You et al 2014ab
  - Poisson mixed effects model — Hall, Ormerod, Wand 2011; Hall et al. 2011
  - stochastic block-models — Celise et al. 2012, Bickel et al. 2013
  - mixtures of Gaussians — Wang & Titterington 2006
  - asymptotic variational posterior variance is "too small" — Mackay 2003

- McCormick & Westling (2017)
  - consistency and asymptotic normality, through profiling variational parameters

Variational inference

- VL: approx $L(\theta; y)$ by a simpler function of θ, e.g. $\prod q_j(\theta)$
- CL: approx $f(y; \theta)$ by a simpler function of y, e.g. $\prod_A f(y_j \in A; \theta)$

Some Links between Variational Approximation and Composite Likelihoods?

S. Robin

UMR 518 AgroParisTech / INRA Applied Math & Comput. Sc.

MSTGA, Paris, November 22-23, 2012

Inference and likelihood

- simplify the likelihood
  - composite likelihood
  - Laplace approximation to integrals
  - variational approximation

- simulate the likelihood
  - approximate Bayesian computation
  - Markov chain Monte Carlo ML

- change the mode of inference
  - indirect inference
  - quasi-likelihood

High-dimensional inference

High-dimensional inference?

- complex models, but $p < n$ fixed
- sometimes simplified, or dimension-reduced, by imposed sparsity

  e.g. INLA
- what about regularized approximate likelihoods

$$\ell(\theta) - P_\lambda(||\theta||)$$

- example: penalized composite likelihood for Ising model

  Xue et al. 2012
- inference after selection ...
- high-dimensional logistic regression   Sur & Candes 2018, Sur et al. 2017

$$\log(p_i/1 - p_i) = x_i^{\mathrm{T}}\beta, \quad y_i \sim \text{Bernoulli}(p_i)$$

- if the MLE exists, then

$$\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - a_*\beta) \quad \longrightarrow \quad 0$$

$$\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - a_*\beta)^2 \quad \longrightarrow \quad \sigma_*^2$$

- LRT for $H : \beta_j = 0$ has scaled χ^2

$$w(\beta_j) \xrightarrow{d} \frac{\kappa\sigma_*^2}{\lambda_*}\chi_1^2$$

- (α, σ, κ) characterized as the solution of three equations

Many other approximate likelihood functions

- synthetic likelihoods — e.g. Price et al. (2017)
- iterated filtering — Bretó 2018
- generalized profile likelihood — Severini 1988
- approximation based on case-control studies — Raftery et al.
- hybrid empirical likelihood and likelihood — Hjort et al.
- extended saddlepoint — Fasiolo et al.
- ... Ogden, 2017 & 2018
- leading to many approximate posteriors
  Ruli, 2016 & 2018

References i

Bickel, P., Choi, D., Chang, X. & Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.* **41**, 1933–1943.

Blei, D.M., Kucukelbir, A. & McAuliffe, J.D. (2017). Variational inference: a review for statisticians. *J. Am. Statist. Assoc.* **112**, 859–877.

Bretó, C. (2018). Modeling and inference for infectious disease dynamics: a likelihood-based approach. *Statist. Sci.* **33**, 57–69.

Davison, A.C. & Huser, R.(2015). Statistics of Extremes *Annual Review of Statistics and its Application* **2**, 203–235.

Fasiolo, M., Wood, S.N., Hartig, F. & Bravington, M.V. (2017). An extended empirical saddlepoint approximation for intractable likelihoods. arXiv:1601.01849v5 [stat.ME]

Hall, P., Ormerod, J.T. & Wand, M.P. (2011). Theory of Gaussian variational approximation for a Poisson mixed model *Stat. Sinica* **21**, 369–389.

Hall, P., Pham, T., Wand, M.P. & Wang S.S.J. (2011). Asymptotic normality and valid inference for Gaussian variational approximation. *Ann. Statist.* **39** 2502–2532.

References ii

Hjort, N.L., McKeague, I.W. & van Keilegom, I. (2017). Hybrid combinations of parametric and empirical likelihoods. *Statist. Sinica*, to appear.

Martins, T.G., Simpson, D., Lindgren, F. & Rue, H. (2013). Bayesian computing with INLA: new features. *Comput. Stat. Data Anal.* **67**, 68–83.

Ogden, H. (2017). On asymptotic validity of naive inference with an approximate likelihood. *Biometrika* **104**, 153 – 164.

Ormerod, & Wand, M. (2010). Explaining variational approximations. *Am. Stat.* **64**, 140–153.

Ormerod, & Wand, M. (2012). Gaussian variational approximate inference... *J Comp Graph Statist***21**, 2–17.

Papavasiliou, A. & Taylor, K.B. (2016). Approximate likelihood construction for rough differential equations. arXiv:1612.02536 [math.ST]

Price, L.F., Drovandi, C.C., Lee, A. & Nott, D.J. (2017). Bayesian synthetic likelihood. *J. Comp. Graph. Statist.* doi:10.1080/10618600.2017.1302882.

Raftery, A.E., Niu, X., Hoff, P.D. & Yeung, K.Y. (2012). Fast inference for the latent space network model Using a case-control approximate likelihood. *J. Comput. Graph. Statist.* **21**, 901–919.

Ranalli, M., Lagona, F., Picone, M. & Zambianchi, E. (2018). Segmentation of sea current fields by cylindrical hidden Markov models: a composite likelihood approach. *J. R. Statist. Soc.* C **67**, 575–598.

Rue,H., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P. & Lindgren, F.K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and its Application* **4**, 395–421.

Ruli, E. & Ventura, L. (2016). Higher order Bayesian approximations for pseudo-posterior distributions. *Comm. Statist. – Sim. Comp.* **45**, 2863–2873.

Severini, T.A. (1998). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika* **85**, 507–522.

Simpson, D.P., Rue, H., Riebler, A., Martins, T.G. and Sørbye, S.H. (2017). Penalising model component complexity: a principled, practical approach to constructing priors (with discussion). *Statist. Sci.*,

References iv

Stroud, J.R., Stein, M. L. & Lysen, S. (2017). Bayesian and Maximum Likelihood Estimation for Gaussian Processes on an Incomplete Lattice. *J. Comp. Graph. Statist.* **26**, 108–120.

Sur, P. & Candés, E. (2018). A modern maximum-likelihood theory for high-dimensional logistic regression. arXiv:1803.06964 [math.ST]

Sur,P., Chen, Y. & Candés, E. (2017). The likelihood ratio test in high-dimensional logistic regression is asymptotically a *rescaled* chi-square. arXiv:1706.01191 [math.ST]

Titterington, D.M. (2006). Bayesian methods for neural networks ... *Statistical Science* **19**, 128–139.

Verbeke, G. & Molenberghs, G. (2017). Modelling Through Latent Variables. *Annual Review of Statistics and its Application* **4**, 267 – 282.

Westling, T. and McCormick, T.H. (2017). Consistency, calibration and efficiency of variational inference. arXiv:1510.08151v3 [stat.ME] 27 Jan 2017.

Thank You!