Non-standard approaches to nonparametric Bayes

Jeff Miller

Joint work with David Dunson

Harvard University
Department of Biostatistics

Bayesian semi- and nonparametric modelling session
CMStatistics, Dec 11, 2016

Motivation

- In Bayesian nonparametrics, the standard approach is to
  - specify a complete probabilistic model,
  - perform fully Bayesian inference, and
  - use algorithms that are exactly correct (possibly up to MCMC error).
- However, this can take a lot of time, both in terms of computation and implementation.
- Compromising on these standard assumptions can allow for methods that are faster and easier to use, and behave similarly.
- Specifically, there can be significant advantages to using
  - partially specified models,
  - a combination of frequentist and Bayesian inference, and
  - analytical approximations to BNP models.

Outline

1. Brief survey of some interesting non-standard approaches
   - Analytical approximations to BNP models
   - Hybrid Bayesian-frequentist methods
   - Partially specified models

2. Nonparametric Laplace approximation
   - Bayesian model checking with a nonparametric alternative
   - Group comparisons (Two-sample tests)
   - Regression with unknown error distributions

Outline

1. Brief survey of some interesting non-standard approaches
   - Analytical approximations to BNP models
   - Hybrid Bayesian-frequentist methods
   - Partially specified models
2. Nonparametric Laplace approximation
   - Bayesian model checking with a nonparametric alternative
   - Group comparisons (Two-sample tests)
   - Regression with unknown error distributions

Analytical approximations to BNP models

- Often, we are only interested in one part of a model, and the rest is just necessary to adequately fit the data.
- Idea: Find analytical approximations for dealing with the parts we don't care about.
- Shift the burden from computation to analysis (i.e., invest more time on derivation & justification instead of running MCMC forever).

Analytical approximations to BNP models

- Example: Recent work by Matt Taddy
  - Take some functional of interest $\beta(P)$, e.g., least squares linear fit.
  - Consider the posterior of $\beta(P)$ when $P \sim \mathrm{DP}(\alpha, H)$ with $\alpha \to 0$.
  - Take a first-order Taylor approximation $\widetilde{\beta}(P) \approx \beta(P)$, and obtain analytical expressions for posterior moments of $\widetilde{\beta}(P)$.
- Example: C-posterior — analytical approximation to marginal likelihood under nonparametrically coarsened models.
- Example: Nonparametric Laplace approximation (this talk) — approximation to marginal likelihood under a nonparametric sieve.

Hybrid Bayesian-frequentist methods

- There's some great stuff outside Bayesian statistics.
- Fast non-Bayesian algorithms for key problems
  - multivariate density estimation (GMRA, IFGT, Dual trees)
  - nearest neighbor search (random k-d tree, k-means tree, LSH)
  - clustering (CLIQUE, BIRCH, DBSCAN)
  - property testing with sublinear time algorithms
  - nonparametric regression, classification, dimensionality reduction, stochastic optimization, convex optimization, ensemble methods, randomized algorithms, etc., etc., etc.
- Can we combine these with Bayes, to obtain fast semi-Bayesian nonparametric methods?

Hybrid Bayesian-frequentist methods

- Example: For conditional density estimation, Petralia, Vogelstein, and Dunson (2013) use a frequentist method to choose a multiscale partition, and combine this with a Bayesian model.
- Example: Nonparametric Laplace approximation (this talk) employs a frequentist density estimate to approximate a nonparametric Bayesian marginal likelihood.

Partial models and generalized posteriors

- Fully Bayesian inference involves specifying a complete model.
- When little prior knowledge is available about a certain part of the model, it is common to use BNP for this part.
- In some cases, another option is to use a partially specified model in which this part is not modeled at all. This results in a loss of information, but often the loss is minimal.
- This is an old idea, but not many Bayesians seem to use it.
  "We're Bayesians — we don't want to lose any information!"

Partial models and generalized posteriors

- The Neyman–Scott problem is a simple but really nice example:
- Suppose $X_i, Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ indep. for $i = 1, \ldots, n$, and we want to infer σ^2, but the distribution of the μ's is completely unknown.
- Problem: Prior on the μ's does not go away — using the wrong prior leads to inconsistency.
- Full BNP approach: put a prior on the distribution of the μ's, e.g., use a Dirichlet process mixture and do inference with usual algorithms.
- Partial model approach: Let $Z_i = X_i - Y_i \sim \mathcal{N}(0, 2\sigma^2)$ and use $p(\sigma^2 | z_1, \ldots, z_n)$ to infer σ^2. Way easier than full BNP!
- Partial model gives consistent and correctly calibrated Bayesian posterior on σ^2 — just slightly less concentrated.

Partial models and generalized posteriors

- There are a variety of ways to obtain a generalized likelihood.
  - conditional likelihood, partial likelihood, pseudo-likelihood, composite likelihood, restricted/marginal likelihood, rank likelihood, etc.
- Generalized posterior $\propto$ generalized likelihood $\times$ prior.
- Generalized posteriors can have advantages over the standard posterior in terms of computation and robustness.
  - Doksum & Lo (1990) — Using $p(\theta \mid \mathrm{median}(x_{1:n}))$ fixes the Diaconis & Freedman (1986) inconsistency issue.
  - Raftery, Madigan, & Volinsky (1996)
  - Hoff (2007)
  - Liu, Bayarri, & Berger (2009)
  - Pauli, Racugno, & Ventura (2011)
  - Lewis, MacEachern, & Lee (2014)
- Main issue is ensuring correct calibration of generalized posteriors.
- In recent work, we have developed Bernstein–Von Mises results for generalized posteriors, to facilitate correct calibration.

Outline

1. Brief survey of some interesting non-standard approaches
   - Analytical approximations to BNP models
   - Hybrid Bayesian-frequentist methods
   - Partially specified models

2. Nonparametric Laplace approximation
   - Bayesian model checking with a nonparametric alternative
   - Group comparisons (Two-sample tests)
   - Regression with unknown error distributions

Nonparametric Laplace approximation — Motivation

- A common usage of BNP models is as a prior on an unknown "nuisance" distribution.
- Examples
  - Regression with unknown error distribution(s).
  - Many parameters with a common unknown distribution (e.g., Neyman–Scott problem).
  - Nonparametric alternative for Bayesian model checking.
  - Comparing groups for equality of distribution (two-sample testing).
- In such cases, we don't care about the unknown distribution itself.
- Using something like a DPM for this is slow and tedious.
- The DP is (often) inapplicable if the data is continuous.

Nonparametric Laplace approximation — Motivation

- It would be nice to be able to integrate out the unknown distribution and have an analytical expression for the resulting marginal likelihood.
- Polya trees (Lavine, 1992) are often used for this reason.
  - Berger & Guglielmi (2001) — Bayesian model checking
  - Hanson and Johnson (2002) — nonparametric regression error
  - Holmes, Caron, Griffin, & Stephens (2015) — two-sample testing
- However, Polya trees strongly depend on a rather arbitrary choice of partition sequence (especially in multiple dimensions). Mixtures of Polya trees are better, but require additional computation. Polya trees also tend to generate spiky distributions.
- We are working on a new approach, with the aim of developing a nonparametric analogue of the Laplace approximation.

Nonparametric Laplace approximation

- Recall the Laplace approximation to the marginal likelihood:

$$m(x_{1:n}) = \int \Big(\prod_{i=1}^{n} p(x_i|\theta)\Big)\pi(\theta)d\theta \approx \Big(\prod_{i=1}^{n} p(x_i|\hat{\theta})\Big)\Big(\frac{2\pi}{n}\Big)^{D/2} \frac{\pi(\hat{\theta})}{|H(\hat{\theta})|^{1/2}}$$

  where $\hat{\theta}$ is the MLE and D is the dimensionality of θ.
- When θ is infinite-dimensional, this is clearly inapplicable. However, by using a sieve (i.e., let model complexity grow with n), perhaps we can mimic the infinite-dimensional case.
- Thus, to obtain an infinite-dimensional analogue, we consider a sieve of continuous coarsenings of $\mathrm{DP}(\alpha, H)$, leading to:

$$m(x_{1:n}) \approx \Big(\prod_{i=1}^{n} \hat{f}(x_i)\Big)\Big(\frac{2\pi}{n+\alpha}\Big)^{\widetilde{D}/2} C_n(\hat{f}, \alpha, H)$$

  where $\hat{f}(x)$ is a nonparametric density estimate and $\widetilde{D}$ is the "effective dimensionality."

Nonparametric Laplace approximation

- In more detail,

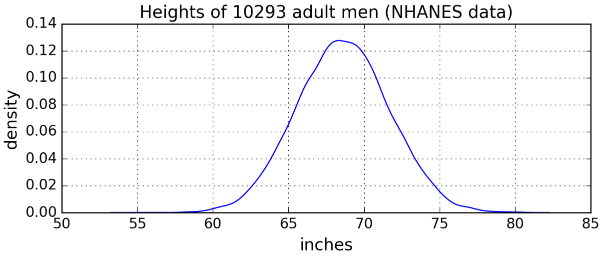
$$m(x_{1:n}) \approx \widetilde{m}_{\mathsf{NPL}}(x_{1:n}) = \Big(\prod_{i=1}^{n} \hat{f}(x_i)\Big)\Big(\frac{2\pi}{n+\alpha}\Big)^{\widetilde{D}/2} C_n(\hat{f},\alpha,H)$$

$$C_n(\hat{f},\alpha,H) = \frac{\Gamma(\alpha)(n+\alpha)^n}{\Gamma(n+\alpha)e^n}\prod_{i=1}^{n}\left(\frac{\big(w\hat{f}(x_i)(n+\alpha)/e\big)^{wh(x_i)}}{\sqrt{w\hat{f}(x_i)}\,\Gamma(wh(x_i))}\right)^{D_i}$$

  - $\hat{f}(x)$ is a nonparametric density estimate,
  - $\widetilde{D} = \sum_{i=1}^{n} D_i$ and $D_i = 1/(wn\hat{f}(x_i))$,
  - w is a complexity parameter of the density estimate (e.g., bandwidth),
  - α is the concentration parameter, and
  - h is the density of the base distribution H.
- Given a nonparametric density estimate $\hat{f}$, this is easy to compute.
- It applies in the multivariate case.

Example 1: Bayesian model checking

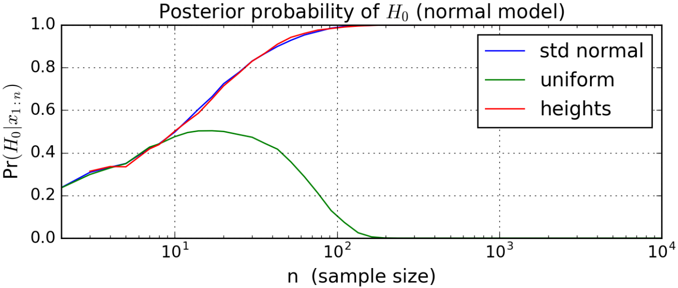
- Are human heights normally distributed? (Well, obviously not, since height is nonnegative. But how good is the normal model for my set of n datapoints?)



- H_0: Normal model, H_1: Nonparametric alternative.
- $\Pr(\mathrm{H}_0 | x_{1:n}) = \left(1 + \frac{p(x_{1:n}|\mathrm{H}_1)p(\mathrm{H}_1)}{p(x_{1:n}|\mathrm{H}_0)p(\mathrm{H}_0)}\right)^{-1}$
- $p(x_{1:n}|\mathrm{H}_0) =$ Normal–NormalGamma marginal likelihood
- $p(x_{1:n}|\mathrm{H}_1) \approx \widetilde{m}_{\mathsf{NPL}}(x_{1:n}) =$ NP Laplace approx

Example 1: Bayesian model checking

Results as the sample size n increases, for three different datasets:



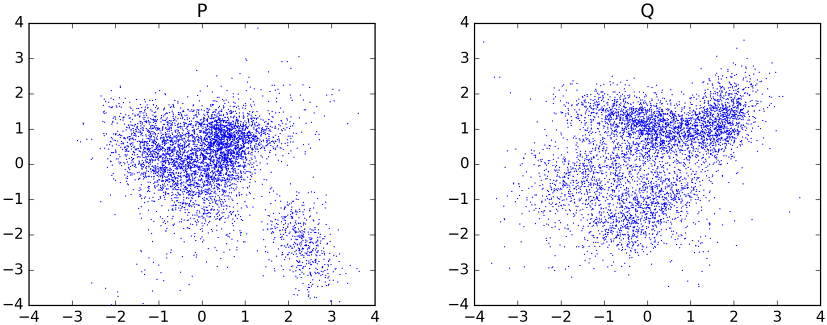
- std normal: Data is $x_1, \ldots, x_n$ i.i.d. $\sim \mathcal{N}(0,1)$.
- uniform: Data is $x_1, \ldots, x_n$ i.i.d. $\sim \mathrm{Uniform}(-1,1)$.
- heights: $x_1, \ldots, x_n$ are the heights of adult men (NHANES data). The normal model appears to be adequate *for the given sample size*.
- (Curves shown are averaged over multiple permutations of the data.)

Example 2: Two-sample testing (Group comparison)

- Do groups A and B have the same distribution? This question is ubiquitous in scientific and industrial applications, e.g.,
  - Does the treatment have any effect?
  - Does knocking out gene G affect disease D?
  - Does using material M affect product quality?
- Assume $X_1, \ldots, X_n | P$ i.i.d. $\sim P$ and $Y_1, \ldots, Y_m | Q$ i.i.d. $\sim Q$.
- $\mathrm{H}_0 : P = Q$, $\mathrm{H}_1 : P \neq Q$
- BNP approach: Put nonparametric priors on P and Q.
- We can approximate a nonparametric marginal likelihood using NPL.
- $p(x_{1:n}, y_{1:m} | \mathrm{H}_0) \approx \widetilde{m}_{\mathsf{NPL}}(x_{1:n}, y_{1:m})$
- $p(x_{1:n}, y_{1:m} | \mathrm{H}_1) = p(x_{1:n} | \mathrm{H}_1) p(y_{1:m} | \mathrm{H}_1) \approx \widetilde{m}_{\mathsf{NPL}}(x_{1:n}) \widetilde{m}_{\mathsf{NPL}}(y_{1:m})$

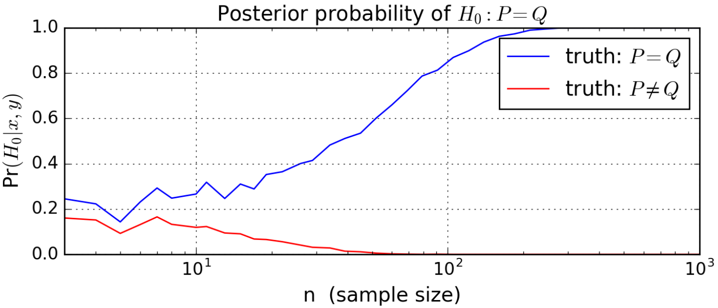
Example 2: Two-sample testing (Group comparison)

Simulated data from two randomly-chosen normal mixtures, P and Q



Example 2: Two-sample testing (Group comparison)

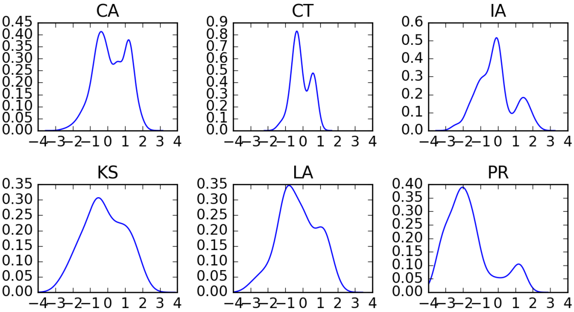
Results as the sample size n increases (averaged over multiple runs):



- When the truth is $P = Q$, we observe $\widetilde{\Pr}(P = Q|x_{1:n}, y_{1:m}) \to 1$.
- When the truth is $P \neq Q$, we observe $\widetilde{\Pr}(P = Q|x_{1:n}, y_{1:m}) \to 0$.
- The NPL approach seems to be working as expected.

Example 3: Regression with unknown error distributions

- Consider HHS data on pneumonia treatment quality in US hospitals.
  - Covariate vector $x_{ij} \in \mathbb{R}^p$ for each hospital j in each state i.
  - y_{ij} = percent of patients given correct treatment (logit-transformed).
- Residuals from a pooled linear regression indicate non-normal errors:



- Following Rodriguez, Dunson, & Gelfand (2008), we model the error distribution for each state nonparametrically.

Example 3: Regression with unknown error distributions

- Model:

  $\beta \sim$ multivariate normal

  $f_1, \ldots, f_k \sim$ nonparametric prior on densities

  $p(y_{ij}|x_{ij}, \beta, f_i) = f_i(y_{ij} - \beta^{\mathtt{T}} x_{ij})$.

- Suppose we're interested in β, but not $f_1, \ldots, f_k$.
- We can use NPL to construct an approximate marginal likelihood:

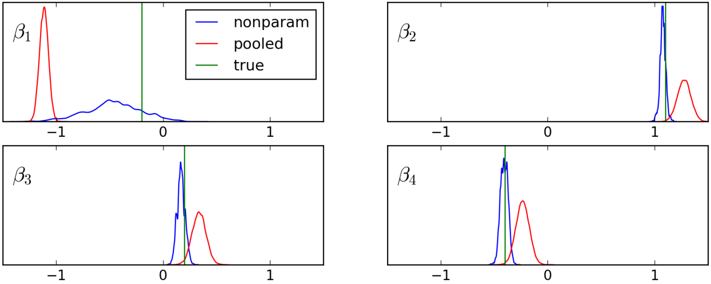
$$p(y|x, \beta) \approx \prod_{i=1}^{k} \widetilde{m}_{\mathsf{NPL}}(r_{i1}(\beta), \ldots, r_{in_i}(\beta))$$

  where $r_{ij}(\beta) = y_{ij} - \beta^{\mathtt{T}} x_{ij}$.

- We can then run Metropolis–Hastings to sample β.

Example 3: Regression with unknown error distributions

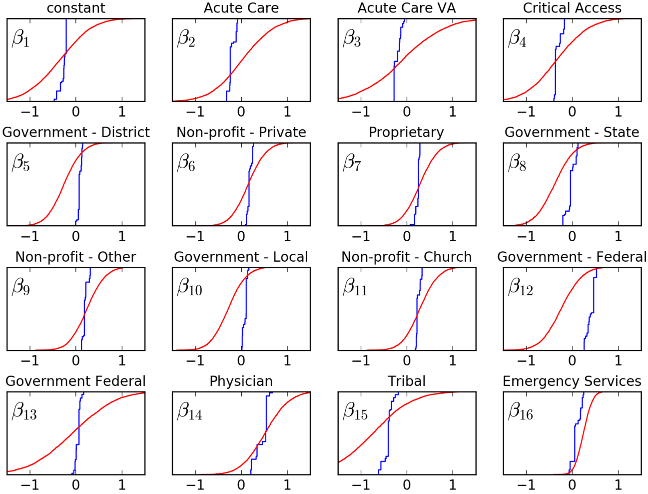
Posterior densities of the coefficients β_i, for simulated data



- As one might expect, a pooled linear regression model doesn't work well — the posterior on β is not concentrating at the true values.
- Meanwhile, the nonparametric Laplace (NPL) approach seems to work quite well — the true values are well-supported by the posterior.
- (A hierarchical normal model should be added to this comparison.)

Example 3: Regression with unknown error distributions

CDFs for results on hospital data (blue=nonparam, red=pooled):



Conclusion

- These preliminary results suggest that the nonparametric Laplace approximation idea is promising as a computationally-efficient alternative to a full Bayesian nonparametric marginal likelihood.
- More generally, non-standard approaches to BNP provide interesting opportunities for advances in terms of computation, ease-of-use, and robustness.

Non-standard approaches to nonparametric Bayes

Jeff Miller

Joint work with David Dunson

Harvard University
Department of Biostatistics

Bayesian semi- and nonparametric modelling session
CMStatistics, Dec 11, 2016