Discussion of “Representative points for Small and Big Data Problems”

(with thanks to Won Chang, Ben Lee, and Jaewoo Park)

Quasi Monte Carlo Transition Workshop

SAMSI, May 2018

Murali Haran

Department of Statistics, Penn State University

A few computational challenges

- Maximize (minimize) expensive or intractable likelihood (objective) function for data $\mathbf{X}$ and parameter θ,

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \mathcal{L}(\theta; \mathbf{X}), \text{ or } \hat{\beta} = \arg\min_{\beta} f(\beta; \mathbf{X})$$

- Bayesian inference, with prior θ

$$\pi(\theta|\mathbf{X}) \propto \mathcal{L}(\theta; \mathbf{X}) p(\theta).$$

- Approximating normalizing constants
- Notation: number of data points is n (as $\mathbf{X} = (X_1, \ldots, X_n)$), dimension of θ is d, dimension of each X is p

Big data and small data problems

These challenges (previous slide) can arise in different settings

- *Big data setting*: n is large, making $\mathcal{L}(\theta; \mathbf{X})$ expensive to evaluate due to matrix computations.
  - High-dimensional regression (e.g. song release prediction, Mak and Joseph, 2017)
  - Models for high-dimensional spatial data
  - High-dimensional output of a computer model
- *Small data setting*: each "data point" is expensive to obtain
  Statistical model = deterministic model + error model
  - deterministic model = climate model, engineering model
  - Very slow to run at each input (θ)
  - Studying deterministic model as we vary input similar to likelihood or objective function that is expensive

A general strategy

Work with surrogate: replace $\mathcal{L}(\theta; \mathbf{X})$ with $\mathcal{L}(\cdot)$.

- Evaluate $\mathcal{L}(\theta; \mathbf{X})$ on a relatively small set of θ values. Fit a Gaussian process (GP) approximation to these sample to obtain $\mathcal{L}_{GP}(\theta; \mathbf{X})$, treated as a surrogate.
- Literature starting with Sacks et al. (1989) and GP-based emulation-calibration (Kennedy and O'Hagan, 2001)
- Can do
  - optimization with $\mathcal{L}_{GP}(\theta; \mathbf{X})$
  - Bayesian inference based on $\pi(\theta|\mathbf{X}) \propto \mathcal{L}_{GP}(\theta; \mathbf{X})p(\theta)$

Challenges posed by GP approximations

- Gaussian processes use dependence to pick up non-linear relationships between input and output: remarkably flexible "non-parametric" framework and hence very widely applicable

(1) However, if input dimension (dimension of θ) is large
  - Expensive/impossible to fill up space with slow model, resulting in poor prediction

(2) If possible to obtain lots of runs (model is not too expensive), can fill up space but...
  - Expensive to fit GP to large n number of data points (model runs): order n^3 cost of evaluating $\mathcal{L}(\theta; \mathbf{X})$ for each θ

Working Group IV's solutions

Solutions:

(1) Kang and Huang (2018): Reduce dimension of input (θ) to θ^* using convex combination of kernels, $\mathcal{L}_{GP}(\theta^*; \mathbf{X})$

(2) Mak and Joseph (2018): Reduce the number of data points $\mathcal{L}(\theta; \mathbf{X})$ via clever design of "support points". Reduction from $\mathbf{X}$ to $\mathbf{X}^*$ to obtain surrogate $\mathcal{L}_{GP}(\theta; \mathbf{X}^*)$. Easier to evaluate

Active data reduction (Mak and Joseph, 2018): reduce the number of data points from n to much smaller number n'

Statistics literature on these problems

- There is a (large) body of work on dimension reduction
- Input space: Dimension reduction in regression by D.R. Cook, Bing Li, F. Chiaromonte, others
  - Finding central mean subspace (Cook and Li, 2002, 2004)
  - Lots of theoretical work, and lots of applications
- Also, literature separation between environmental/spatial and engineering folks?
  - composite likelihood (Vecchia, 1988) (no reduction)
  - reduced-rank approaches...

Open questions - I

- Reduced-rank approaches (active area) statistics):
  - kernel convolutions (Higdon, 1998)
  - predictive process (Banerjee et al., 2008)
  - random projections (Banerjee et al, 2012; Guan, Haran, 2018)
  - multi-resolution approaches (Katzsfuss, 2017)
- Data compression literature?
- How do the existing approaches compare to the proposed approaches from this group?
- Useful thought experiment, even without simulation study
  - computational costs? detailed complexity calculations?
  - approximation error?
  - ease of implementation? (should not be underestimated!)
  - theoretical guarantees?

A different kind of dimension-reduction problem

(Aside)

- In many problems the output of the model is very high-dimensional, that is, if **X** is *p*-dimensional in $\mathcal{L}(\theta; \mathbf{X})$, with *p* large
- Example: climate model output (SAMSI transition workshop next week)
- An approach: Principal components for fast Gaussian process emulation-calibration (e.g. Chang, Haran et al., 2014, 2016a, b; Higdon et al., 2008):
  - Treat multiple model runs as replicates and find principal components to obtain low-dimensional representation
  - Use GP to emulate just the principal components

Open questions - II

- Is it possible to handle higher dimensions than the examples shown in Kang and Huang? E.g. in climate science interested in 10-20 or even larger dimension of θ
- Are there connections between the dual optimization approach (Lulu Kang's talk) and other surrogate methods?
- Does active data reduction preserve dependence structure and other complexities in the data?
  - E.g. consider data compression work by Guinness and Hammerling (2018), specifically targeted at spatial data
- Active data reduction: How is GP fit quickly with new samples at each iteration? (important!)
- Any way to batch this instead of 1 point at a time?

Adaptive estimation of normalizing constants

- Idea: fit linear combination of normal basis functions using MCMC samples + unnormalized posterior evaluations
- Closed-form normalizing constant from approximation
- How does methodology work if (i) unnormalized posterior is expensive, (ii) sampling is expensive?
- Approximating covariance Σ: Fast? What is being assumed about Σ? Need some restrictions, but cannot be restrictive or it will not work well for complicated dependence in posterior
- Why refer to “rejected” samples from MCMC separately? Treat as Monte Carlo procedure regardless of whether MCMC was used (all “accepted”!)
- Work would benefit from challenging Bayes example!

A sense of scale (what is "big"?)

- Different ice sheet simulation models I work with
  - <1 to 20 seconds per run ("run" = one input (θ))
  - 2 to 10 minutes per run
  - 48 hours per run
- # evaluations (n) possible: hundreds to millions
- # of parameters (d) of interest varies between 4 and 16
- # dimensions of output (p) varies from 4 to $\approx$ 100,000
- Different computational methods for different settings
  - MCMC algorithms (fast model, many parameters)
  - Gaussian process emulation (slow model, few parameters)
  - Reduced-dimensional GP (slow model, few parameters, high-dimensional output), e.g. Chang, Haran et al. (2014)
  - Particle-based methods (moderately fast, many parameters): ongoing work with Ben Lee et al. (talk at SAMSI Transition Workshop next week)

Another problem that pushes the envelope

- Consider a problem where evaluating $\mathcal{L}(\theta; \mathbf{X})$ is expensive *and* θ is not low-dimensional
- Question: How well would the working group's methods adapt to this scenario?
- Example: Bayesian inference for doubly intractable distributions

Models with intractable normalizing functions

- Data: $\mathbf{x} \in \chi$, parameter: $\theta \in \Theta$
- Probability model: $h(\mathbf{x}|\theta)/Z(\theta)$
  where $Z(\theta) = \int_\chi h(\mathbf{x}|\theta) d\mathbf{x}$ is intractable
- Popular examples
  - Social network models: exponential random graph models (Robins et al., 2002; Hunter et al., 2008)
  - Models for lattice data (Besag, 1972, 1974)
  - Spatial point process models: interaction models Strauss (1975), Geyer (1999), Geyer and Møller (1994), Goldstein, Haran, Chiaromonte et al. (2015)

Bayesian inference

- Bayesian inference
  - Prior : $p(\theta)$
  - Posterior: $\pi(\theta|\mathbf{x}) \propto p(\theta)h(\mathbf{x}|\theta)/Z(\theta)$
- Acceptance ratio for Metropolis-Hastings algorithm

$$\frac{\pi(\theta'|\mathbf{x})q(\theta_n|\theta')}{\pi(\theta_n|\mathbf{x})q(\theta'|\theta_n)} = \frac{p(\theta')Z(\theta_n)h(\mathbf{x}|\theta')q(\theta_n|\theta')}{p(\theta_n)Z(\theta')h(\mathbf{x}|\theta_n)q(\theta'|\theta_n)}$$

Cannot evaluate because of $Z(\cdot)$

A function emulation approach

- Existing algorithms are all computationally very expensive (Park and Haran, 2018a)
  - Each iteration of algorithm involves an "inner sampler", a sampling algorithm for a high-dimensional auxiliary variable. Inner sampler is expensive (again, expensive $\mathcal{L}(\theta; \mathbf{X})$)
- Our function emulation approach (Park and Haran, 2018b)
  1. Approximate $Z(\theta)$ using importance sampling on some k design points, $\widehat{Z}_{IMP}(\theta_1), \dots, \widehat{Z}_{IMP}(\theta_k)$
  2. Use Gaussian process emulation approach on k points to interpolate this function at other values of θ, $\widehat{Z}_{GP}(\theta)$
  3. Run MCMC algorithm using $\widehat{Z}_{GP}(\theta)$ at each iteration
- We have theoretical justification as # design points (k) and # importance sampling draws increases

Results for an example

Emul_1, Emul_{10} are two versions of our algorithm

Double M-H is fastest of existing algorithms

Simulated social network (ERGM): 1400 nodes

θ_2	Mean	95%HPD	Time(hour)
Double M-H	1.77	(1.44, 2.12)	23.83
Emul_1	1.79	(1.45, 2.13)	0.45
Emul_{10}	**1.96**	(1.87, 2.05)	**1.39**

True $\theta_2 = 2$: Emul_{10} is accurate, others are not

Computational efficiency allows us to use longer chain (Emul_{10}). Corresponding DMH algorithm $\approx$ 10 days

Positives and limitations

- Our approach can provide accurate approximations for problems for which other methods are unfeasible
- Works well only for θ of dimension under 5. This still covers a huge number of interesting problems, but it would be nice to go beyond
  - higher-dimensions: unable to fill the space well enough to approximate the normalizing function well
- We require a good set of design points at the beginning. Hence, have to run another (expensive) algorithm before running this one. This is a major bottleneck
- *Interesting opportunities for (i) input-space dimension reduction, (ii) clever design strategies*

Discussion (of discussion)

- Congratulations to the speakers: they are tackling numerous very interesting and useful problems, broadly related to handling expensive likelihood/objective functions
- They offer creative solutions to challenging problems:
  - Clever design (support points)
  - New methods for dimension reduction of data
- Lots of existing work in dimension reduction, and in Gaussian process emulation-calibration literature that might be worth investigating
- Open problem when parameters are not low-dimensional and the objective function is expensive to evaluate

Selected references

- Higdon (1998) A process-convolution approach to modelling temperatures, *Env Ecol Stats*
- Park and Haran (2018a) Bayesian Inference in the Presence of Intractable Normalizing Functions (on arxiv.org) to appear *J of American Stat Assoc*
- Guan, Y. and Haran, M. (2018) “A Computationally Efficient Projection-Based Approach for Spatial Generalized Linear Mixed Models,” to appear in *J of Comp and Graph Stats*
- Chang, W., Haran, M., Applegate, P., and Pollard, D. (2016) “Calibrating an ice sheet model using high-dimensional binary spatial data,” *J of American Stat Assoc*, 111 (513), 57-72.
- Cook, R.D. and Li, B. (2002) “Dimension reduction for conditional mean in regression,” *Annals of Stats*