Scalable maximum likelihood estimation for Gaussian processes

Michael Stein[1], Jie Chen[2] and Mihai Anitescu [2]

University of Chicago, ANL and ANL

2011 DOE Applied Mathematics Program Meeting

[1]Supported by U.S. Department of Energy Grant No. DE-SC0002557.
[2]Supported by the U.S. Department of Energy through Contract No. DE-AC02-06CH11357.

Gaussian processes

Gaussian process models are a fundamental tool in spatial statistics and statistical analysis of computer experiments.
Observational and computer-generated datasets often have spatial and temporal aspects and the resulting datasets can be enormous.

- MODIS measures light intensity in 36 spectral bands at ≈ 60 million locations daily.
- Computer models (climate models, high fidelity models for advanced nuclear reactors) can produce even larger datasets.

Gaussian process models can be used to

- describe/characterize the fluctuations in these processes
- predict unobserved values of the process and provide prediction uncertainties
- can serve as a building block for more complex models (e.g., for land use classification based on LANDSAT data).

A process Z on a domain D is called Gaussian if, for every $x_1, \ldots, x_n \in D$, $W = (Z(x_1), \ldots, Z(x_n))$ follows a Gaussian, or multivariate normal distribution: its probability density is of the form

$$p(w) = \frac{1}{(2\pi)^{n/2}|K|^{1/2}} \exp\left\{-\frac{1}{2}(w-\mu)K^{-1}(w-\mu)\right\}$$

for some mean vector μ and positive definite covariance matrix K, in which case, say $W \sim N(\mu, K)$.
Thus, the process Z is determined by

- its mean function $\mu(x) = EZ(x)$
- its covariance function $K(x, y) = \mathrm{cov}\{Z(x), Z(y)\}$

for all $x, y \in D$.
In practice, μ and/or K will be partially unknown and need to be estimated from observations.
This talk focuses on:

- estimation of K when specified up to some finite-dimensional θ.
- Assume $\mu = 0$ for simplicity.

Maximum likelihood estimation

A standard and generally effective way of estimating unknown parameters in a statistical model is via maximum likelihood:

Likelihood function: given θ, suppose $p_\theta(w)$ is the joint density of the observations. Given observations w, the likelihood function is just $p_\theta(w)$ viewed as a function of θ.

Maximum likelihood estimate (mle): $\hat{\theta}$ is called an mle of θ if it maximizes $p_\theta(w)$ over all possible θ.
Standard asymptotic theory says that when the data are highly informative about θ, one often has

$$\hat{\theta} \approx N(\theta, \mathcal{I}(\theta)^{-1}),$$

for $\mathcal{I}(\theta)$ the Fisher information matrix, defined as

- the covariance matrix of the *score function:* $\frac{\partial}{\partial\theta} \log p_\theta(w)$.
- Often cannot improve nonnegligibly on mle.

Computation

Exact computation of the likelihood for n irregularly sited observations generally requires $O(n^3)$ computation and $O(n^2)$ memory.

Options for large n:

- Use model that reduces computation and/or storage.
- Use approximate methods.
- Both.

Goal of project is to be able to use these models on petascale data.

- Even for terascale ($n \approx 10^{12}$) data, probably need single-pass methods if want to fit global model.

What would be lost by fitting a bunch of local models?

Models that reduce computation

There are many models that can reduce computations even for irregularly sited observations:

- Compactly supported covariance functions
- Reduced rank covariance functions
- Markov models

Each has their strengths and weaknesses, but all can lead to making unnatural assumptions about the process.
I would rather approximate the likelihood than use what I consider a less appropriate model.

> *An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.*

John Tukey

Approximate computation

Spectral methods (Whittle likelihood).

- For gridded data and stationary processes, can use fft, so
  - fast
  - storage needs linear in number of observations
  - considerable theory showing it is statistically efficient.
- Can use for nonstationary and/or ungridded data, but not as effective/efficient.

Various forms of composite likelihood:

$$\prod_j p_\theta(s_j \mid c_j)$$

for subsets s_j and c_j of the observations.

- Combine local and sparse subsets of data (Caragea and Smith, 2007).
- $s_1, \ldots, s_m$ partition of data; $c_j \subset (s_1, \ldots, s_{j-1})$ (Vecchia 1988; Stein, Chi and Welty 2004).
  - Best approach in terms of accuracy per flop?

Iterative solution of linear equations

Kriging and computing quadratic form in likelihood requires solving $Kx = y$ for covariance matrix K.

Iterative methods effective if

- multiplying vector by K is fast
- condition number (ratio of largest to smallest eigenvalue) $\kappa(K)$ of K is small.

Under fixed domain asymptotics, generally $\kappa(K) \to \infty$ as sample size increases.

- Increase can be rapid if process smooth.

With evenly spaced observations in time, could prewhiten data.

- If truth is random walk, then first differences iid and $\kappa(K) = 1$ for differenced data.

Can we do something similar for more general models, irregular observations or spatial data?

Results from Stein, Chen and Anitescu (under review at SIMAX):

Z on real line with spectral density f satisfying

$$f(\omega)\omega^2 \text{ bounded away from 0 and } \infty \text{ as } \omega \to \infty.$$

- Condition says process is not too different from Brownian motion.

Observations arbitrarily located on some fixed, bounded interval.

Let L be filter matrix for normalized first differences.

Theorem: There exists $C_f < \infty$ such that, for *any* set of observations of Z in $[0, 1]$,

$$\kappa(LKL') \leq C_f. \qquad (*)$$

Note: C_f is independent of the sample size n.

Can add row to L to get matrix $\tilde{L}$ such that

- $\tilde{L}$ is full rank
- $(*)$ holds for $\tilde{L}$.

Analogous result for processes not too different than integrated Brownian motion:

Z on real line with spectral density f satisfying

$$f(\omega)\omega^4 \text{ bounded away from 0 and } \infty \text{ as } \omega \to \infty.$$

Now let L be filter matrix for normalized second differences.

Theorem: There exists $C_f < \infty$ such that for *any* set of observations in $[0, 1]$

$$\kappa(LKL') \leq C_f. \qquad (*)$$

C_f is independent of the sample size n.

Can add two rows to L and get full rank $\tilde{L}$ for which $(*)$ holds.

Proofs make use of results on equivalence of Gaussian measures(!)

More than one dimension?

- Now much harder to handle irregular observation locations.

Suppose Z is a process on $\mathbb{R}^d$

- observed on grid $\frac{1}{n}(j_1, \ldots, j_d)$ for $0 \leq j_k \leq n$ and $1 \leq k \leq d$
- with spectral density f satisfying

$$f(\omega) \asymp (1 + |\omega|)^{-4p}$$

  for some positive integer $p > \frac{1}{4}d$.

If apply discrete Laplacian p times to the observations, then the condition numbers of the resulting covariance matrices are bounded in n.

- Number of observations reduced from $(n+1)^d$ to $(n+1-2p)^d$.
- Should be possible to augment filter matrix as in one dimension, but not so clear how.

If spectral densities aren't like $|\omega|^{-2p}$ (for $d = 1$) or $|\omega|^{-4p}$ ($d > 1$) for an integer p, then

- shouldn't expect simple filters to yield bounded condition numbers
- but empirical results show can still improve conditioning a lot.

Iterative methods can work well on matrices with a few extreme eigenvalues, which preconditioning can produce quite broadly.

Can we use these iterative methods to help with likelihood computations?

- Likelihood requires $\log |K|$.

Solve score equations instead?

$$\frac{1}{2} z' K(\theta)^{-1} \frac{\partial K(\theta)}{\partial \theta_i} K(\theta)^{-1} z - \frac{1}{2} \mathrm{tr}\left\{ K(\theta)^{-1} \frac{\partial K(\theta)}{\partial \theta_i} \right\} = 0$$

First term requires only one solve.

Instead of log determinant, need

- for each component of θ,

$$\mathrm{tr}\left\{ K(\theta)^{-1} \frac{\partial K(\theta)}{\partial \theta_i} \right\},$$

  which requires n solves for exact calculation.

Approximate by the unbiased estimate

$$\frac{1}{N} \sum_{j=1}^{N} u_j^T K(\theta)^{-1} \frac{\partial K(\theta)}{\partial \theta_i} u_j,$$

where $u_j = (u_{j1}, \ldots, u_{jn})'$ is random vector with u_{jk}'s iid and $\Pr(u_{jk} = 1) = \Pr(u_{jk} = -1) = \frac{1}{2}$.

- If don't compute loglikelihood, what do you do if solution not unique?

How large does N (the number of u_j's) need to be to yield an accurate approximation to mle?

- If need N comparable to n, then this approach not attractive.

For z the vector of observations,

$$g_\theta(z) = \frac{1}{2} z' K(\theta)^{-1} \frac{\partial K(\theta)}{\partial \theta_i} K(\theta)^{-1} z - \frac{1}{2N} \sum_{j=1}^{N} u_j^T K(\theta)^{-1} \frac{\partial K(\theta)}{\partial \theta_i} u_j = 0$$

is a set of unbiased estimating equations for θ:

$$E_\theta g_\theta(z) = 0,$$

where E_θ means to average over z (given θ) and the u_j's.

Define the matrices

$$A = E_\theta \frac{\partial}{\partial \theta} g_\theta(z)$$

and

$$B = \mathrm{cov}\{g_\theta(z)\}.$$

"Standard theory" of estimating equations shows ($\hat{\theta}_N$ is solution of approximate score equations)

$$\hat{\theta}_N - \theta \approx N(0, A^{-1}BA^{-1}),$$

as n increases, although this result may not apply under fixed domain asymptotics.

The score equations yield optimal estimating equations with $A = B = \mathcal{I}(\theta)$, the Fisher information matrix. Here,

- $A = \mathcal{I}(\theta)$ as for score equations
- $B = \mathcal{I}(\theta) + \frac{1}{N} J(\theta) > \mathcal{I}(\theta)$.

Can prove $J(\theta) \leq \frac{\{\kappa(K)+1\}^2}{4\kappa} \mathcal{I}(\theta)$.

Estimating equations are asymptotically optimal if, as $n \to \infty$,

$$B\mathcal{I}(\theta)^{-1} = I + \frac{1}{N} J(\theta)\mathcal{I}(\theta)^{-1} \to I.$$

- If $\kappa(K)$ bounded in n, $N \to \infty$ suffices.
  - Don't need N comparable to n!
- Indeed, for bounded $\kappa(K)$, $N = 1$ yields equations with optimal rate as $n \to \infty$.
  - If that seems shocking, note method is exact with $N = 1$ if $K(\theta)$ diagonal.
- Filtering first to control condition number helps in two ways:
  - Reduces number of iterations needed in iterative solver.
  - Reduces need for large N.

The bound $J(\theta) \leq \frac{\{\kappa(K)+1\}^2}{4\kappa}\mathcal{I}(\theta)$ can be quite conservative.

- $J(\theta) \leq \mathcal{I}(\theta)$ if $K(\theta) = \theta_0 I + \theta_1 K_1$ for all K_1.

$J(\theta)$ can be further reduced by choosing u_j's not independent.

Example

Stationary Gaussian random field on $\mathbb{R}^2$ observed on $n \times n$ square grid (so $m = n^2$ observations) with spacing $\frac{100}{n}$.

Model for autocovariance function:

$$\sqrt{2} r_\theta(x) \mathcal{K}_1\left(\sqrt{2} r_\theta(x)\right), \quad \text{where} \quad r_\theta(x) = \sqrt{\frac{x_1^2}{\theta_1^2} + \frac{x_2^2}{\theta_2^2}},$$

$\mathcal{K}_1$ is modified Bessel function and $\theta = (7, 10)$ is truth.

Spectral density f satisfies $f(\omega) \asymp (1 + |\omega|)^{-4}$, so if apply Laplacian once, covariance matrices

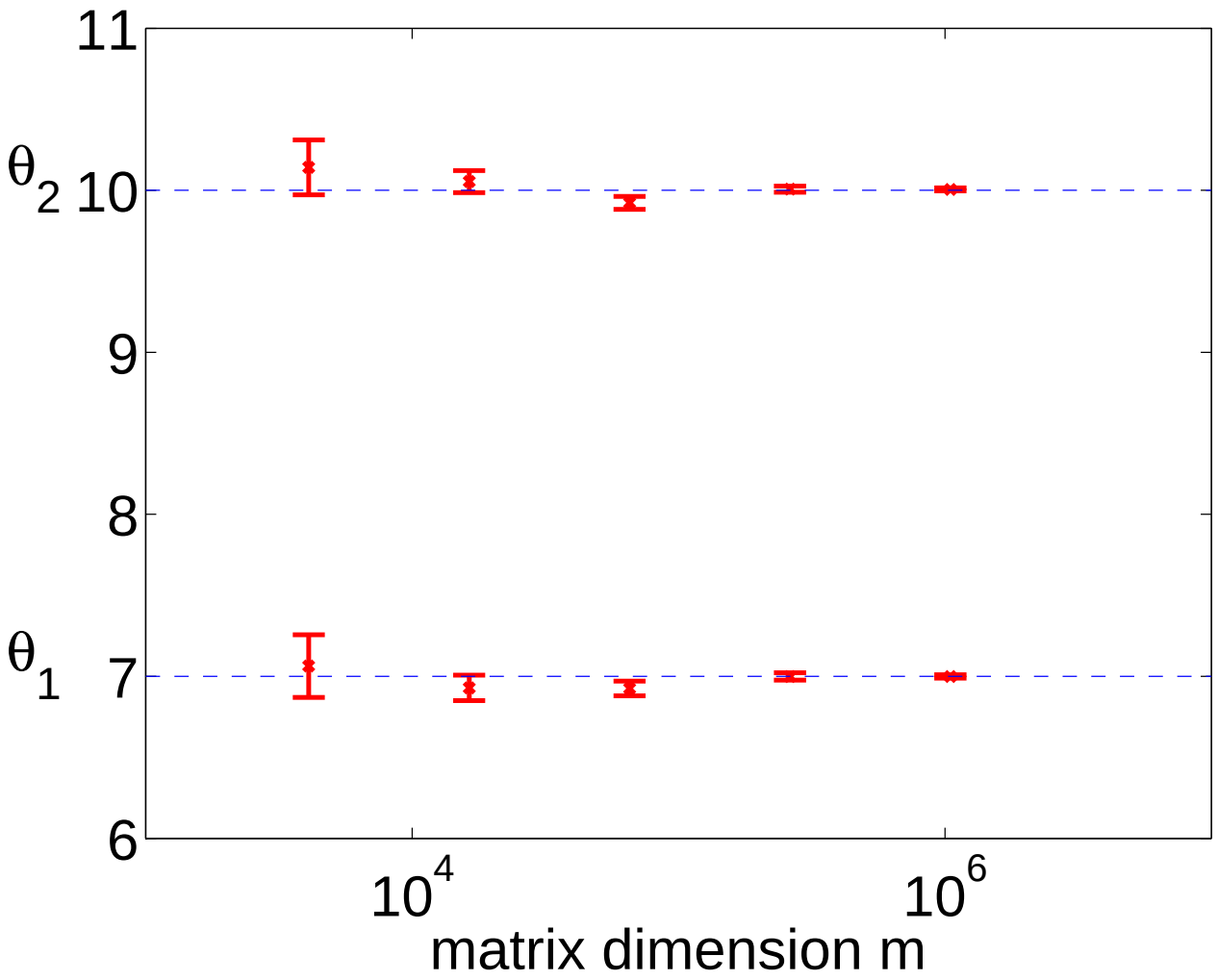
- have condition number bounded in m
- are BTTB, so multiplication is fast

and algorithm scales nearly linearly in m.

Use approximate score function with

- $N = 100$ for all m.
- Intervals indicate uncertainty due to using finite N.

As theory predicts, fixed N leads to decreasing uncertainty as m increases.


11
θ2 10
9
8
θ1 7
6
10^4
10^6
matrix dimension m

One-pass methods

Look at data block by block and summarize the information about $K(\theta)$ from that block so that don't have to go back to raw data again.

Simple example:

- Divide data into B blocks.
- Within each block, approximate the loglikelihood (or score) function, which is a *sufficient* statistic.
  - Mle of θ and observed information matrix an adequate approximation?
  - If not, store more complete representation of loglikelihood function. Adding loglikelihoods across blocks reduces storage with little loss of information?
- Save a few observations (or other summaries) from each block to recover information at larger scales.

When might this procedure do asymptotically as well as full likelihood?

- "Vecchia" version of this may be more efficient statistically but not as simple. Not so easy to parallelize.

For petascale data, probably need more than two "layers."

Further thoughts

Opportunities to combine approaches?

- Tapering (to induce sparseness) together with filtering (to improve conditioning).
- Using various methods for variance reduction in simulations in stochastic algorithms (doing calculation for many θ).

Massive datasets generally show clear nonstationarity and/or non-Gaussianity, so need for good statistical methods won't disappear with increasing sample sizes.

Some standard methods in numerical linear algebra (iterative approaches, multipole, multigrid) underutilized by statisticians?

Some ideas from probability and statistics (equivalence of Gaussian measures, asymptotics of kriging) underutilized by numerical analysts?