Quasi Monte-Carlo Inference

for Log Gaussian Cox Processes

Kui Tang, Jessica Forde, Liam Paninski

Columbia University

May 9, 2013

Problem

- Goal: model an underlying intensity function based on temporal point data
  - Sparse and noisy signal
  - Examples: Neural spike trains, defaults, mine disasters [**?**] (1-dimensional), spatial data (2+ dimensional)

Model

- Log Gaussian Cox Process (LGCP) [**?**]: Poisson process driven by exponentiated hidden Gaussian process.
  - $g(t) \sim \mathcal{GP}(m(t), k(t, t'))$
  - $\lambda(t) = e^{g(t)}$
  - $x \sim \mathcal{IP}(\lambda(t))$
- X a set of N observed events (points) in time window $\mathcal{T} = [a, b]$; likelihood is

$$p(X|g) = \exp\left\{-\int_{\mathcal{T}} \lambda(\mathbf{x})dx\right\} \prod_{n=1}^{N} \lambda(x_n) \tag{1}$$

- Posterior of latent functions is

$$p(g|X) = \frac{\exp\left\{-\int_{\mathcal{T}} \lambda(x)dx\right\} \prod_{n=1}^{N} \lambda(x_n)\mathcal{GP}(g|m(t), k(t, t'))}{\int \exp\left\{-\int_{\mathcal{T}} \lambda(x)dx\right\} \prod_{n=1}^{N} \lambda(x_n)\mathcal{GP}(g|m(t), k(t, t'))dg} \tag{2}$$

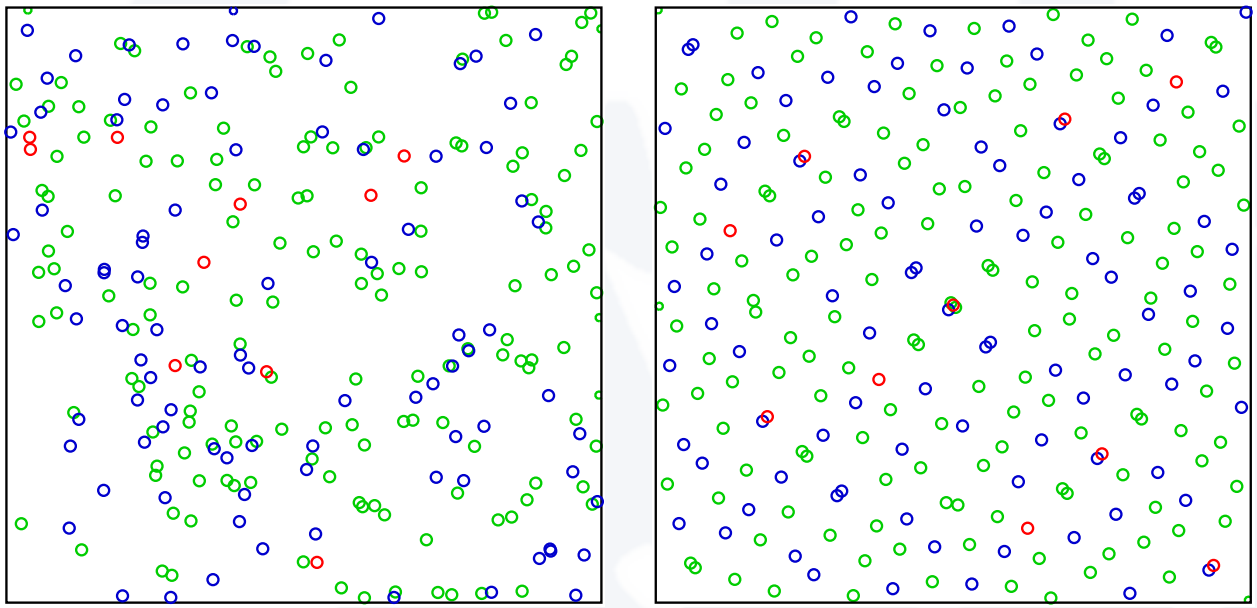
- **Doubly intractable**: g is infinite dimensional and evaluated at uncountably many points in $\mathcal{T}$.

Related Work

- Metropolis Adjusted Langevin Algorithm (MALA) [**?**] on (3), 1998 [**?**]
- Laplace approximation, 2008 [**?**]:
  - ☺: $O(cT \log T)$ MAP estimation; $O(N^3)$ model selection ($N \ll T$).
  - ☹: Equally-spaced, one-dimensional process. MAP only.
- Auxiliary MCMC, 2009 [**?**]:
  - ☺: Asymptotically exact. No discretization. Arbitrary spatial dimensions.
  - ☹: Expensive $O\left(C(N+M)^3\right)$ posterior estimation; C MCMC iterations, M auxiliary events.
- MALA and Iterated Nested Laplace Approximations (INLA) [**?**] for 2d LGCP [**?**]:
  - While INLA is faster, MALA is more exact

Quasi-Monte Carlo

- **MC**: N points $\{\mathbf{x}_n\}$ drawn uniformly from $[0,1]^D$, $V[g(X)] < \infty$
  - $\left| E[g(X)] - \frac{1}{N}\sum_{n=1}^{N} g(\mathbf{x}_n) \right| = O(1/\sqrt{N})$
- **Quasi MC**: Choose $\{\mathbf{x}_n\}$ from $[0,1]^D$ s.t. the fraction of points enclosed in any sub-rectangle $\approx$ volume of the rectangle
  - $\{\mathbf{x}_n\}$: **low-discrepancy set**
  - $\left| E[g(X)] - \frac{1}{N}\sum_{n=1}^{N} g(\mathbf{x}_n) \right| = O\left(N^{-1}(\log N)^D\right)$



Evaluating the Integral

- We approximate $-\int_{\mathcal{T}} \lambda(\mathbf{x})dx$ with Simpson's rule, using T linearly spaced time points in the time window $\mathcal{T}$
- Also have N original observations $\mathbf{t}_{(T+1):(T+N)}$.
- Evaluate $\mathbf{g} \sim \mathcal{N}(0, K)$ where $g_i = g(t_i)$ and $K_{ij} = k(t_i, t_j)$
- The evidence, $P(X)$, in (2) is approximately

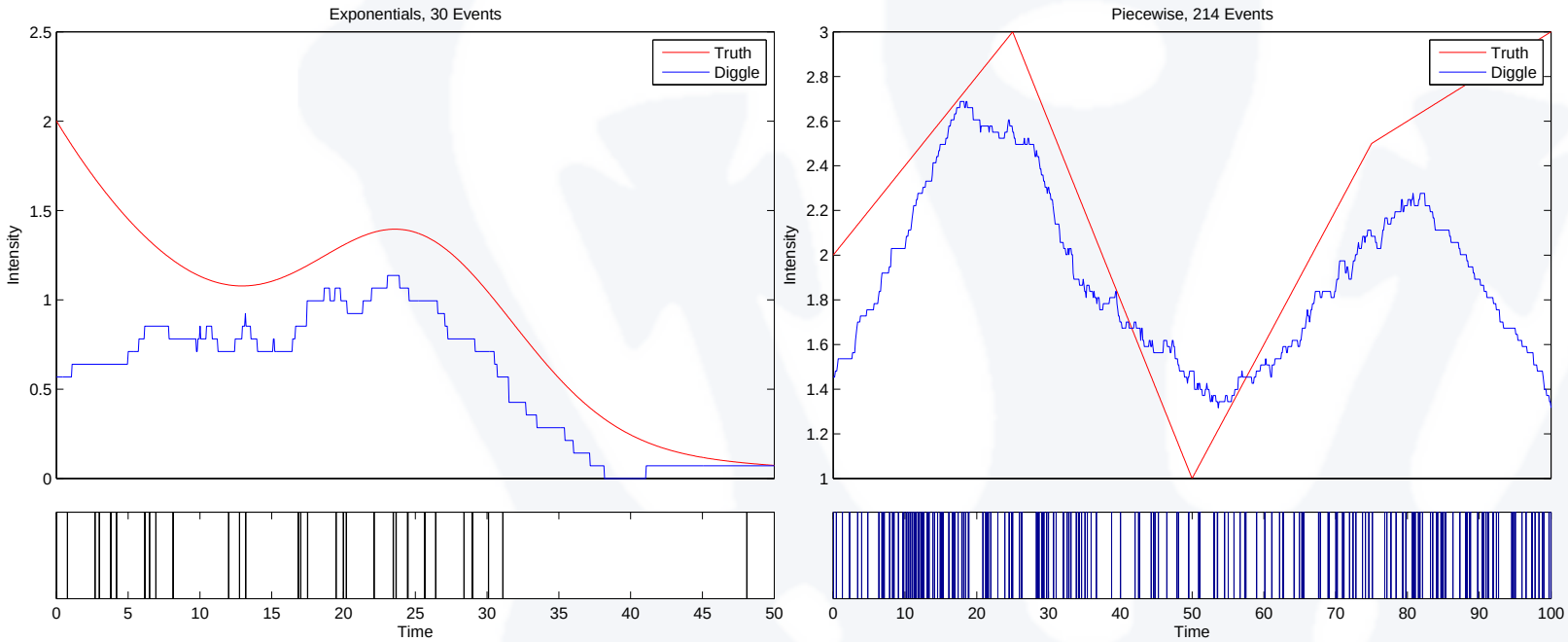
$$\frac{1}{\sqrt{(2\pi)^D |K|}} \int_{\mathbb{R}^{N+T}} \exp\left\{ Z(\mathbf{t}_{1:T}) \sum_{i=1}^{T} e^{g_i} \right\} \prod_{i=T+1}^{N+T} e^{g_i} \exp\left\{ \mathbf{g}^\top K^{-1} \mathbf{g} \right\} d\mathbf{g} \tag{3}$$

  $Z(\mathbf{t}_{1:T})$: constants from Simpson's rule
- To compute moments of $P(g|X)$, multiply powers of $\mathbf{g}$ by the integrand of (3)

Simulated Data

- $N = 30$ events from $\mathcal{T} = [0, 50]$ using an exponential basis function
- $N = 240$ events from $\mathcal{T} = [0, 100]$ using a piecewise basis function
- Goal: Evaluate (3) over a total of $N + T = 1001$ time intervals over $\mathcal{T}$
- As a baseline, we perform uniform kernel smoothing [**?**]

$$\tilde{\mu}(x) = \frac{N(x - t, x + t)}{2t} \tag{4}$$



Covariance Function

- To calculate $\mathbf{g}$, we draw M MC/QMC vectors, $\mathbf{u}_1, ..., \mathbf{u}_M$, and compute $\mathbf{n}_i = \Phi^{-1}(\mathbf{u}_i)$
- In this model, we assume a squared exponential covariance function

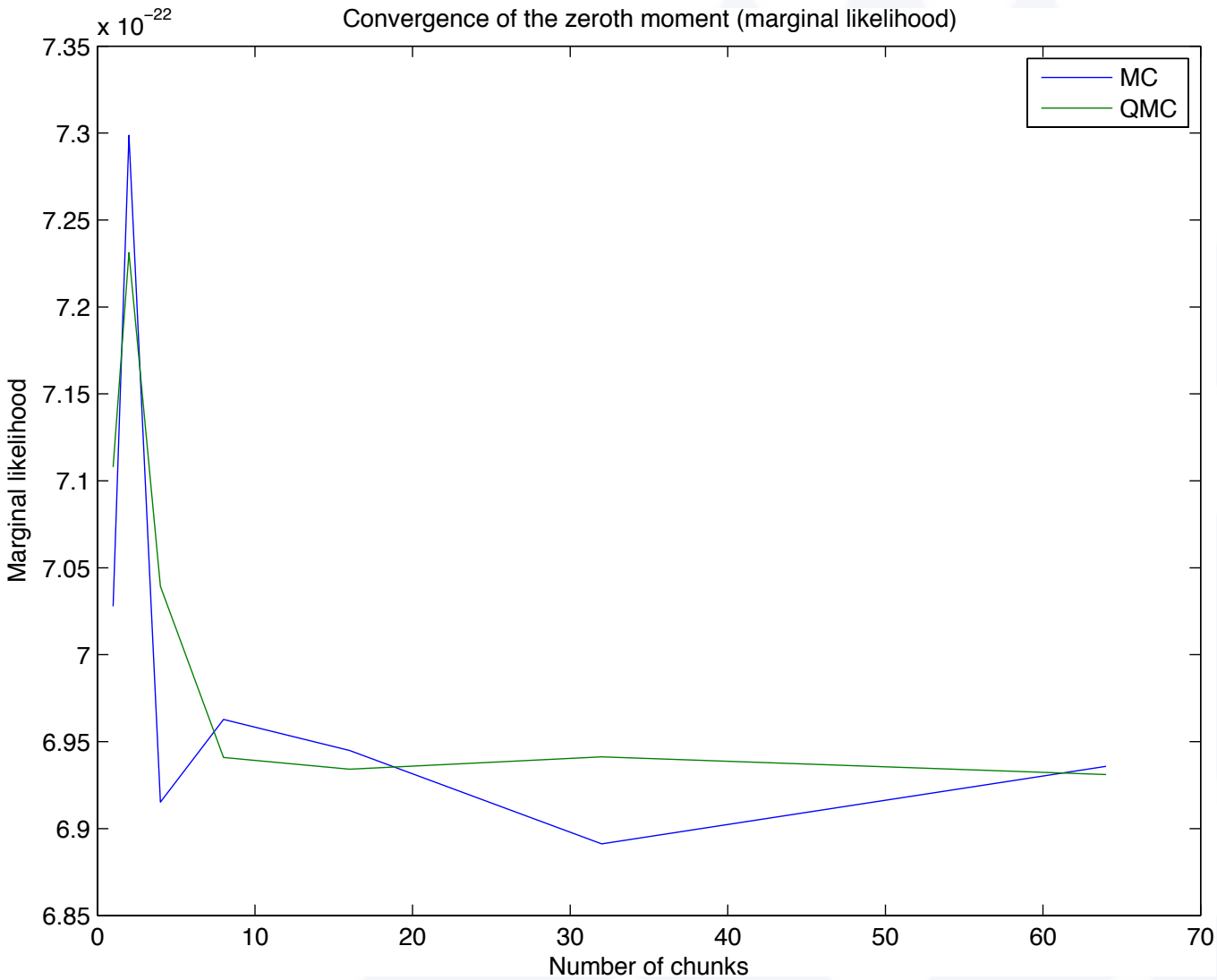
$$K(t,s) = k \exp\left(-\left(\frac{t-s}{\ell}\right)\right)^2 \tag{5}$$

- ℓ controls the strength of long-run correlations between timestamps
- K is often a rank-deficient matrix
- Algorithms relying on the precision, i.e. INLA, cannot directly use K
- Previous research into LGCP [**?**] [**?**] use other covariance functions
- As a result, we use a non-square Cholesky-like decomposition of K, L, to transform antithetic standard normal variables into $\mathbf{g}_i = L\mathbf{n}_i$

Variance and Dimensionality Reduction

- Because our goal is to converge quickly upon a mean intensity function, we want to reduce variance of the integrand
- We use antithetic variables such that for each $\mathbf{u}_i$ there exists $\mathbf{u}_j$ such that $\mathbf{u}_i = \mathbf{1} - \mathbf{u}_j$
- Because K is low rank, the length of $\mathbf{u}_i$ is the rank of K, thereby reducing the number of variables we need to simulate
- Samples of $\mathbf{g}$ remain $N + T$ dimensional since $\mathbf{g}_i = L\mathbf{n}_i$ and L is $(N + T) \times rank(K)$
- With $\mathbf{g}$, we can estimate the intensity using Laplace approximation, MC/QMC, and MALA
  - To combine MALA with QMC, we follow **?** by permuting blocks of QMC points

QMC converges faster than MC

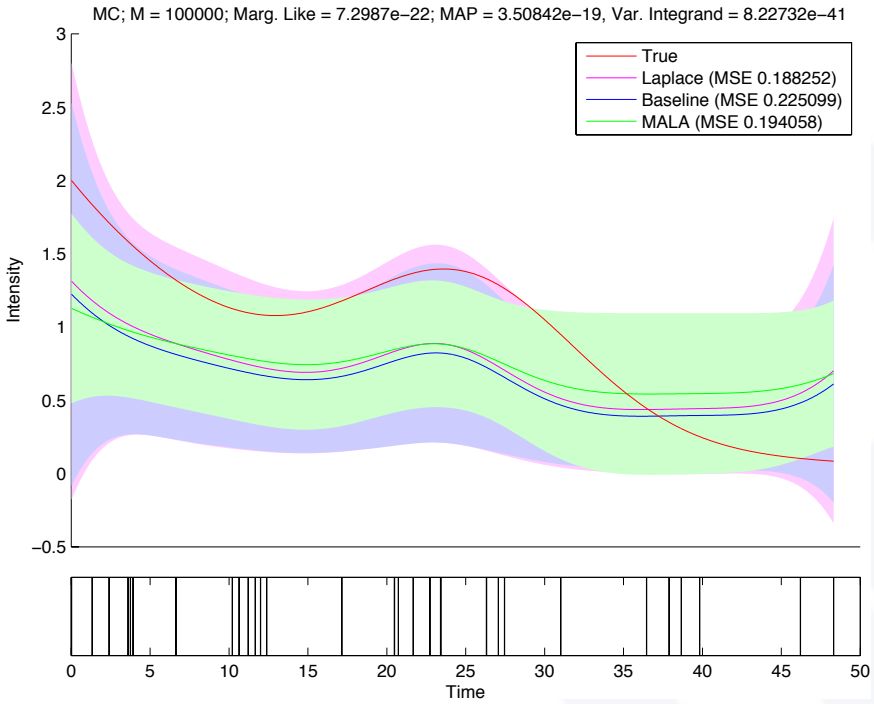


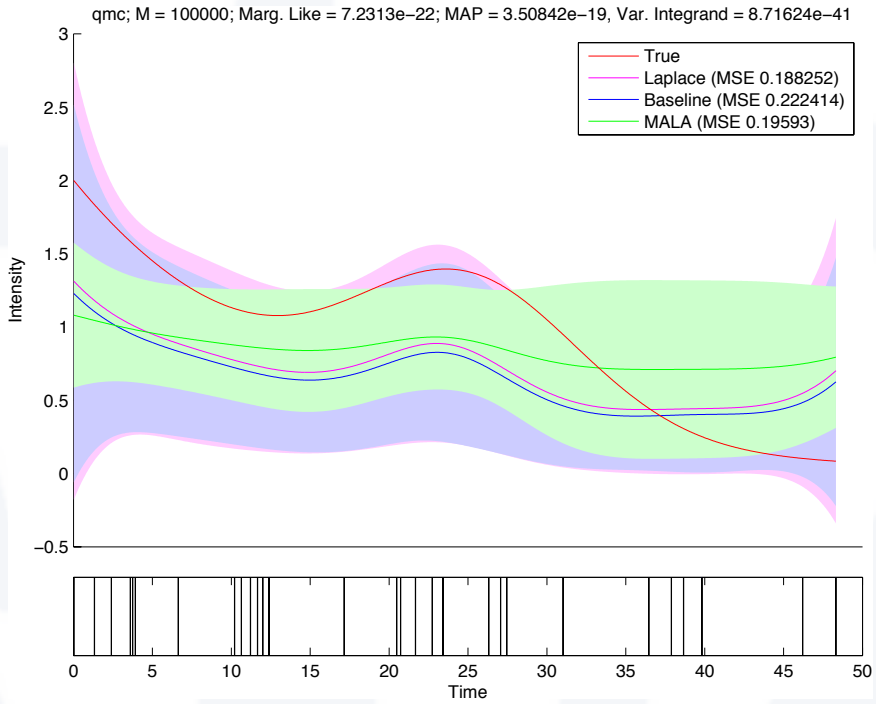
Variance Reduction Results

- Sample variance of marginal likelihood estimate across 24 runs
- QMC is Sobol points with random shift modulo 1

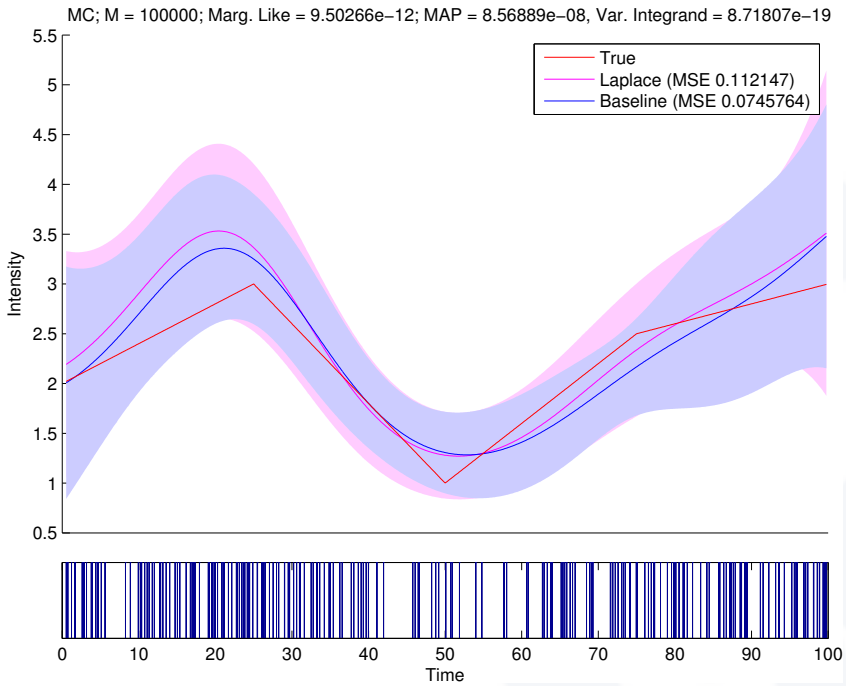
10^5 samples	1	2	4
MC Variance $\times 10^{-45}$	0.1016	0.0817	0.0301
QMC Variance $\times 10^{-45}$	0.0669	0.0514	0.0141
Ratio	1.5182	1.5901	2.1343

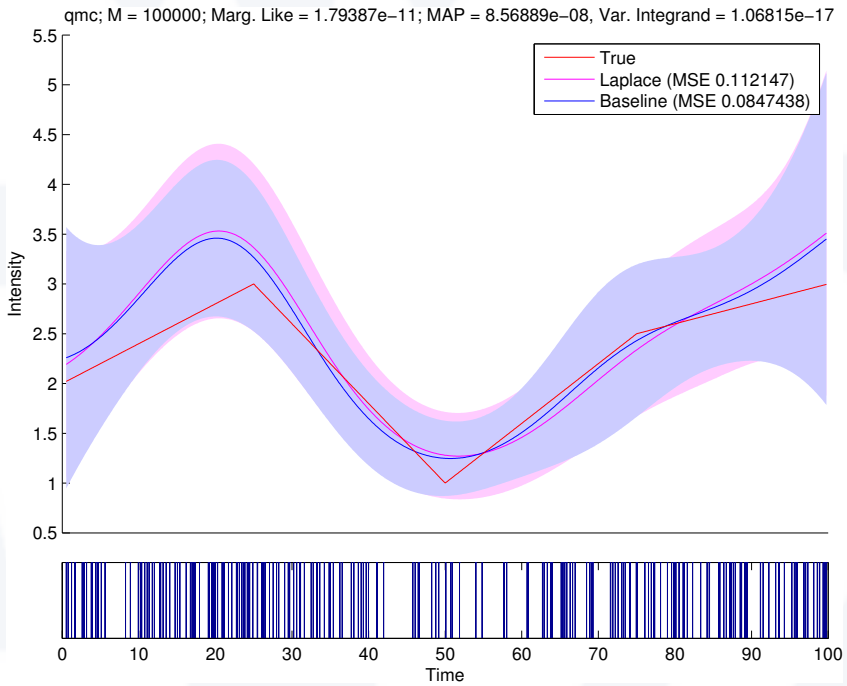
Exponential Results: MALA predicts small credible bands, Laplace large credible bands





Piecewise Results: Laplace suggests high variance





References I