High Performance Bayesian Applications in Medical, Economics and Climate Sciences

Ahmed Almulihi, Dr. Elise de Doncker

Department of Computer Science

Abstract

We apply high performance numerical integration to problems in Bayesian statistics. These are applied to data arising in the analysis of problems in such areas as medical statistics (birth weight data, heart transplant data, photocarcinogen data, radio therapy data), climate (global weather, tornado data), and general statistics applications (multivariate logistic distribution, multivariate normal, nonlinear regression).

We compute Bayesian moment integrals using the ParInt integration software that runs efficiently on computer clusters. We compare our results to those in the literature and show excellent performance with respect to accuracy and execution time.

Introduction

Bayesian analysis is a statistical method that uses probabilities to measure uncertainty about unknown data in order to draw a proper inference. The analysis of complex statistical models using Bayesian inference has become more prominent in recent years. This is partly due to the increasing availability of computing power. Another factor is the discovery of a number of simulation techniques that made Bayesian inference possible for complex models. However, many of these techniques are computationally demanding. This research attempts to enhance Bayesian applications performance by using the ParInt software package.

Problem

Bayesian models analysis often requires the evaluation of complicated multidimensional integrals. For higher-dimensional non-linear models, the only practical methods for analysis are based on stochastic simulation techniques such as Monte Carlo (MC), Quasi-Monte Carlo (QMC) and Markov chain Monte Carlo (MCMC) techniques. These are known to be computationally intensive, with some analyses requiring days of CPU time on powerful computers.

The ParInt software package has been developed at Western Michigan University. The system is designed to solve integration problems numerically via high performance computing. ParInt uses multiple solution methods including adaptive domain partitioning, QMC and MC. It has been implemented on a variety of platforms.

The package is written in the C programming language and it runs on UNIX platforms, using the Message Passing Interface (MPI). The user specifies the problem using C, C++ or Fortran functions in order to be evaluated.

For more Information about ParInt and to download the package source code, visit the ParInt research group website at:

https://cs.wmich.edu/parint/

To learn more about the computation cluster at the department of computer science, visit the High Performance Computational Science Laboratory website at:

https://cs.wmich.edu/~hpcs

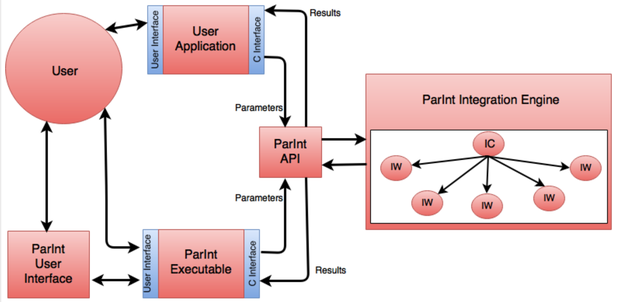

Figure 1: ParInt Architecture [2].

Methods

The purpose of ParInt is to approximate multivariate integrals of the form:

$$I = \int_D f(x)\,dx$$

The output contains an approximation Q and an absolute error estimate E_a that satisfying: $|I-Q| \le E_a \le \max\{\varepsilon_a, \varepsilon_r |I|\}$ where ε_a and ε_r are the absolute and the relative tolerance error, respectively.

ParInt applies an adaptive region subdivision algorithm (Figure 2) that divides the domain D into subregions (Figure 3) for evaluation.

```
Evaluate given domain and initialize results
Initialize priority queue with initial region
while (region evaluation limit not reached & Estimated error too
large)
    Retrieve region from priority queue
    Split region
    Evaluate new subregions and update results
    Insert new subregions into priority queue
```

Figure 2: Adaptive Integration Algorithm [2].

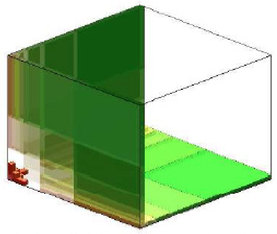

Figure 3: Subdivision of integration domain (D) [3].

We performed the calculations using ParInt on the computation cluster (Thor) at the Center for High Performance Computing and Big Data (WMU) (Figure 4). Thor has 22 computation nodes that provide high performance parallel computing.

Figure 4: A view of Thor at the Department of Computer Science.

Application Data

We selected several applications [4] from various areas to be implemented in C and evaluated using ParInt:

Medical Data

- Birth weight Data: predicting low birth weight, linked to several causes.
- Heart transplant data: survival of patients on the waiting list for the Stanford heart transplant program.
- Photocarcinogen data: mice with tumors survival time.
- Cross-classification in health-care data of schizophrenic patients.

Climate Data

- Global weather: an estimate of the joint probability density function (PDF) for uncertain climate system properties.
- Tornado data: forecasting tornado intensity.

Economics Data

- Various examples from Econometrics [5,6]

General Statistics

- Bayesian analysis of a linear model with simulated data.
- multivariate logistic distribution.
- multivariate normal.
- nonlinear regression.

Results

We consider two examples from Evans and Swartz (1995)[1] to illustrate our results.

Example 1: Linear model

This model represents simulated data and it is specified as:
$y = X\beta + \sigma z$ Where $y \in \mathbb{R}^{45}, X \in \mathbb{R}^{45\times 9}, \beta \in \mathbb{R}^{9}, \sigma \in (0,\infty), z \in \mathbb{R}^{45}$
The required integrals that are of the form:

$$I(m) = \int_{R^*} m(\theta) f(\theta)\, d\theta$$

where $m(\theta)$ are moment functions, $\theta \in \mathbb{R}^{10}$
and the integration variables $\theta_i = \beta_i$ for i = 1,...,9 and $\theta_{10} = \log\sigma$

Figure 5 shows the performance of the integral computation using ParInt, compared to Evans and Swartz results [1].

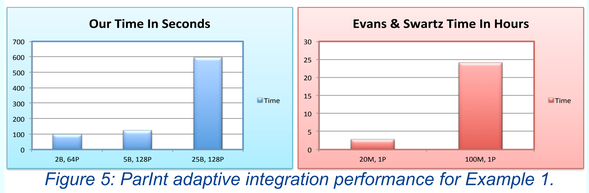

Figure 5: ParInt adaptive integration performance for Example 1.

Our results show good accuracy and a large reduction of the computation time reported by Evans & Swartz [1]. Figure 6 lists our results compared to to the exact results. Observe that the absolute error estimates drop considerably with the higher number of samples used (Figure 7).

w	*Results*	*Exact*
θ_1	2.0428	2.043
θ_2	0.09473	0.095
θ_4	0.01801	0.018
θ_{10}	-0.07290	-0.073

Figure 6: Example 1 results.

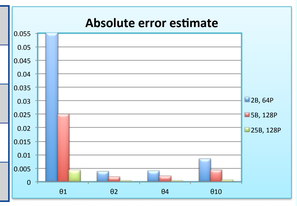

Figure 7: Absolute error estimate.

Example 2: Contingency table

This example analyzes a cross-classification of 132 long-term schizophrenic patients in a table with three row categories concerning the frequency of hospital visits and three columns categories with the length of stay. The cell probabilities are in the form:
$p_{ij} = \theta a_i(1)\beta_j(1) + (1-\theta)\, a_i(2)\,\beta_j(2)$, and the likelihood function is:

$\prod_{i=1}^{3}\prod_{j=1}^{3} p_{ij}^{f_{ij}}$ where f_{ij} is the count in the (i, j) cell.

Figure 8 shows the performance of the integration using ParInt compared to Evans & Swartz results [1].

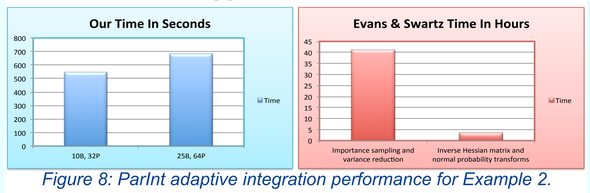

Figure 8: ParInt adaptive integration performance for Example 2.

Also for this problem the results are accurate compared to the exact results (Figure 9). Running this problem with a high number of function evaluations on 64 processes gives a very good absolute error estimate (Figure 10).

w	*Results*	*Exact*
θ	0.4217	0.422
θ^2	0.1808	0.181

Figure 9: Example 2 results.

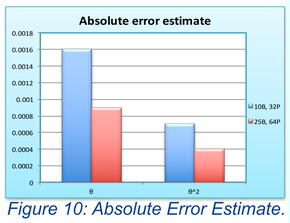

Figure 10: Absolute Error Estimate.

Conclusion & Future work

- Bayesian models require solving high-dimensional multivariate integration problems.
- We implemented a number of Bayesian problems using ParInt.
- ParInt computations deliver accurate and efficient results. Running the applications on Thor reduces the execution time considerably.
- We plan on adding tools for enhancing Bayesian inference solutions in ParInt.
- We will continue to investigate the role of High Performance Computing in analyzing complicated Bayesian models.

References

[1] M. Evans and T. Swartz, "*Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems*" Statistical Science, vol. 10, pp. 254–272, 1995.
[2] E. de Doncker, R. Zanny, and K. Karlis. "*Integrand and performance analysis with ParInt and ParVis.*" in Concurrency & Computation: Practice & Experience (2000).
[3] S. Li, Online Support For Multivariate Integration. A Ph.D Dissertation, WMU, 2005
[4] A. Genz and R. Kass, "BAYESPACK: A collection of numerical integration software for Bayesian analysis," 1997, software available from website at http://www.sci.wsu.edu/math/faculty/genz/homepage.
[5] Kloek, T., and van Dijk, H. K. (1978), "*Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo*"Econometrica, 46, 1-19.
[6] Naylor, J. C. and Smith, A. F. M. (1988), "Econometric Illustrations of Novel Numerical Integration Strategies for Bayesian Inference", J. Econometrics, 38, pp. 103-125.