Project Ideas in Computational Statistics*

Generated with assistance from OpenAI's ChatGPT & Isaque Pim

June 2025

Overview

This document contains a curated list of project ideas suitable for a PhD-level Computational Statistics course. The topics are based on advanced sampling and inference techniques, including Markov Chain Monte Carlo (MCMC), Hamiltonian Monte Carlo (HMC), Simulated Annealing, Expectation-Maximization (EM), Sequential Monte Carlo (SMC), and related algorithms. Each project is designed to be computationally substantial and research-oriented, connecting with current research frontiers.

Project Ideas

1. Scalable Hamiltonian Monte Carlo for Big Data

Description: Develop or evaluate sub-sampling or mini-batch approaches to HMC that scale to massive datasets without sacrificing accuracy.

Challenges: Address noisy gradient estimation and control the bias-variance trade-off.

Key References:

- Betancourt, M. (2017). *A Conceptual Introduction to Hamiltonian Monte Carlo.* arXiv:1701.02434.
- Chen, T., Fox, E., & Guestrin, C. (2014). *Stochastic Gradient Hamiltonian Monte Carlo.* ICML.

Example: Implement SGHMC with variance reduction and compare it to full-data HMC on a large Bayesian model.

2. Normalizing Flows for Improved Proposal Distributions in MCMC/SMC

Description: Use neural network-based normalizing flows to design flexible proposal distributions for SMC or MCMC.

Challenges: Efficient training of flows under limited particle budgets.

Key References:

*PhD-level course offered at the School of Applied Mathemtics, Getulio Vargas Foundation. Instructor: Luiz Max Carvalho.

- Papamakarios, G., Nalisnick, E., et al. (2021). *Normalizing Flows for Probabilistic Modeling and Inference*. JMLR.
- Naesseth, C. A., Linderman, S. W., et al. (2018). *Variational Sequential Monte Carlo*. ICML.

Example: Implement flow-based SMC for a high-dimensional non-linear state-space model.

3. Tempered Transitions and Annealing for Multimodal Posteriors

Description: Explore advanced simulated annealing techniques or tempered transitions for models with highly multimodal distributions.

Key References:

- Neal, R. M. (1996). *Sampling from multimodal distributions using tempered transitions*. Statistics and Computing.
- Geyer, C. J., & Thompson, E. A. (1995). *Annealing Markov chain Monte Carlo with applications to ancestral inference*. JASA.

Example: Apply tempered transitions to mixture models where label-switching occurs.

4. Adaptive SMC for High-Dimensional Bayesian Inference

Description: Develop adaptive resampling and proposal strategies in SMC to handle high-dimensional problems.

Key References:

- Del Moral, P., Doucet, A., & Jasra, A. (2006). *Sequential Monte Carlo samplers*. Journal of the Royal Statistical Society: Series B.
- Beskos, A., et al. (2016). *Sequential Monte Carlo Methods for Bayesian Model Comparison*. Biometrika.

Example: Apply adaptive SMC to model selection in Gaussian process regression or deep Bayesian neural networks.

5. Efficient EM Variants for Latent Variable Models with Intractable E-steps

Description: Explore stochastic or variational EM algorithms when the E-step is analytically intractable.

Key References:

- Neal, R. M., & Hinton, G. E. (1998). *A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants*. In Learning in Graphical Models.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society: Series B.

Example: Implement stochastic EM for a mixture of Gaussian processes model.

6. Combining Variational Inference with MCMC (Hybrid Methods)

Description: Investigate hybrid methods that use variational inference to initialize or guide MCMC, improving convergence.

Key References:

- Salimans, T., Kingma, D. P., & Welling, M. (2015). *Markov Chain Monte Carlo and Variational Inference: Bridging the Gap.* ICML.
- Zhang, J., et al. (2018). *Advances in Variational Inference.* IEEE Transactions on Pattern Analysis and Machine Intelligence.

Example: Compare VI-initialized MCMC with standard MCMC for Bayesian neural networks.

7. MCMC for Non-Euclidean Spaces (e.g., Manifolds, Phylogenies)

Description: Extend HMC or other MCMC methods to work on non-Euclidean spaces like spheres, simplices, or trees.

Key References:

- Byrne, S., & Girolami, M. (2013). *Geodesic Monte Carlo on Embedded Manifolds.* Scandinavian Journal of Statistics.
- Dinh, V., et al. (2017). *Probabilistic Path Hamiltonian Monte Carlo.* ICML.

Example: Implement Riemannian HMC for Bayesian models with simplex constraints.

8. MCMC Diagnostics and Pathologies in High Dimensions

Description: Study the limitations of MCMC (mixing, convergence) in high-dimensional settings and propose diagnostics or remedies.

Key References:

- Betancourt, M. (2017). *A Conceptual Introduction to Hamiltonian Monte Carlo.* arXiv:1701.02434.
- Roberts, G. O., & Rosenthal, J. S. (2001). *Optimal Scaling for Various Metropolis-Hastings Algorithms.* Statistical Science.

Example: Empirically study how dimension affects autocorrelation in different MCMC variants. In particular, discuss what happens when the variables in question are not standard, for instance when there is a moderately high-dimensional indicator parameter $\boldsymbol{\delta} \in 0, 1^d$ such as when performing variable selection a la George & McCulloch (1993) [1].

[1] George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88(423):881–889.

Implementation and Deliverables

Each project may involve:

- Writing custom samplers in Python, R, or Julia.
- Applying methods to synthetic and real datasets (e.g., UCI datasets, image datasets, or biological models).
- Evaluating performance using effective sample size, log marginal likelihood, convergence diagnostics, and computational efficiency.

Deliverables:

- A detailed report including theoretical background, method implementation, experimental results, and comparative analysis.
- An open-source code repository (e.g., GitHub).