Computation (chapters 11+13)

Slide 1

- Algorithms
  - Gibbs Sampling, Metropolis, Metropolis-Hastings
  - rejection sampling
  - importance sampling
  - other
- Convergence
  - diagnostics, burn-in
  - PSRF
  - Kolmogorov-Smirnov-statistics
- Dependence of iterations
  - autocorrelation
  - effective number of samples
  - thinning

Convergence diagnostics and burn-in

Slide 2

- The beginning of the MCMC chain not usable before the starting point has been forgotten
- When the chain has *converged*, samples come from the desired distribution
- Convergence is analysed with *convergence diagnostics*
  - comparison of several independent chains
  - comparison of the different parts of single chain
- Samples simulated before convergence thrown away
  - *burn-in*

MCMC samples not independent

- Monte Carlo estimates still valid
- Monte Carlo error estimates slightly more difficult
  - time series analysis
  - thinning
  - batching

Slide 3

- Estimation of the effective number of samples
  - comparison of independent chains
  - time series analysis

Using single chain

- If simulation time is very long, sometimes only one chain is simulated
  - avoiding several burn-ins
- After burn-in, compare e.g. first and last thirds of the chain

Slide 4

Visual checking

- Quick, useful, but more difficult for large number of parameters
  - not enough for accepting convergence
  - maybe enough for rejecting convergence
  - may hint what is the problem
  - human vision system is efficient to detect patterns which maybe difficult to present mathematically

Slide 5

  - more quantities to monitor makes visual inspection harder

Comparison of between and within variances

- m independent chains, each length is n (after removing first half)
  - samples of scalar to be estimated $\psi_{ij} \quad (i = 1, \ldots, n; j = 1, \ldots, m)$
- Gelman *et al.*: *potential scale reduction factor* (PSRF)
  - based on comparison of between and within variances
  - suitable for distributions which can be approximated with normal distribution
  - scalars should be transformed to get more normal distributions
    - eg. logarithm of positive quantities

Slide 6

  - Gelman *et al.* remove first halves and compare second halves

Comparison of between and within variances

- Estimate variance between chains B

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\psi}_{.j} - \bar{\psi}_{..})^2, \text{ where } \bar{\psi}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}, \bar{\psi}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{\psi}_{.j}$$

  - B/n is variance of chain means

- Estimate variance within chains W

Slide 7

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2, \text{ where } s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\psi_{ij} - \bar{\psi}_{.j})^2$$

- Estimate marginal posterior variance $\mathrm{var}(\psi|y)$ as a weighted average of W and B

$$\widehat{\mathrm{var}}^{+}(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

Comparison of between and within variances

- Estimate $\mathrm{var}(\psi|y)$ as a weighted average of W and B

$$\widehat{\mathrm{var}}^{+}(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

  - this overestimates marginal posterior variance if starting points are overdispersed and B larger
  - unbiased in stationary state or if $n \to \infty$

Slide 8

- For finite n, W underestimates marginal posterior variance
  - single chains have not yet visited everywhere, thus less variation
  - if $n \to \infty, \quad \mathrm{E}(W) \to \mathrm{var}(\psi|y)$
- Since $\widehat{\mathrm{var}}^{+}(\psi|y)$ overestimates and W underestimates, compute

$$\hat{R} = \sqrt{\frac{\widehat{\mathrm{var}}^{+}}{W}}$$

Potential scale reduction factor

- PSRF

$$\hat{R} = \sqrt{\frac{\widehat{\mathrm{var}}^{+}}{W}}$$

  - estimates how much scale of distribution of ψ might reduce, if simulation were continued in limit $n \longrightarrow \infty$
  - $R \longrightarrow 1$, when $n \longrightarrow \infty$
  - if R is large, additional sampling might improve result
  - if R not almost 1 for all scalar quantities, continue sampling
  - suitable threshold often 1.1

- Esim8_1.m

Slide 9

Potential scale reduction factor

- Even if R almost 1, does not guarantee convergence
  - if starting points are not overdispersed
  - distribution differs too much from normal
  - chance due to finite n

Slide 10

Convergence diagnostics based on samples

- Convergence diagnostics based on samples can only indicate if no convergence
  - even if a diagnostic indicates that convergence possible, it is possible that due to combination of starting points, algorithm and chance, none of the chains visit areas with a substantial amount of mass
  - common problems e.g. funnel-shaped and multimodal distributions

Slide 11

Funnel-shaped distributions

- E.g. if in Metropolis-algorithm proposal density is chosen based on the wide part of the funnel, probability to jump to the narrow part of the funnel can be very low
  - if now samples is obtained from the narrow part of the funnel, problem may be undetected by a diagnostic
- possible solutions
  - re-parametrization
  - locally adaptive algorithms such as slice sampling

Slide 12

Multimodal distributions

- In multimodal distribution, the probability that chain travels from one mode to another through the low density region may be very small
  - if the chain does not change the mode, it may appear as the chain has converged, even if the other mode might have substantial amount of mass
- possible solutions
  - prior removes modes

Slide 13

  - more advanced algorithms
    - several coupled chains with different starting points
    - tempering algorithms
    - adaptive population MC algorithms

Convergence diagnostics based on samples

- Diagnostics based on marginals may indicate convergence even if no convergence in joint distribution
  - diagnostics for non-normal multidimensional is difficult
  - diagnostics based on several marginals, has also problem of multiple hypothesis testing

Slide 14

Perfect sampling*

- For some models, there are methods which can indicate convergence perfectly
  - possible to get certainly independent samples
    - only for some restricted models

Slide 15

Convergence diagnostics

- I have used
  - several chains
  - visual inspection
  - potential scale reduction factor
  - Kolmogorov-Smirnov statistic
    - useful also for non-normally distributed
    - later today

Slide 16

Effective number of samples

- Since samples are not independent, the estimate of between variances is overestimate
- Effective number of samples can be estimated as follows

$$n_{\text{eff}} = mn\frac{\widehat{\text{var}}^{+}(\psi|y)}{B}$$

Slide 17

- unreliable estimate if m small
- unreliable estimate if $n_{\text{eff}}/n < 5\%$
- super-efficient sampling, for which $n_{\text{eff}} > mn$, possible, but in practice unlikely
- Gelman et al. carefully report $\min(n_{\text{eff}}, mn)$

How many samples?

- How many independent samples needed?
- What is the effective number of samples?

Slide 18

Thinning

- Use only every k:th MCMC sample
  - if k large enough, remaining samples practically independent

$$k > mn/n_{\text{eff}}$$

  - discards information
  - saves memory and disk space

Slide 19

  - makes sample based inference faster
  - makes estimation of Monte Carlo error easier
    (if k estimated well)

Batching*

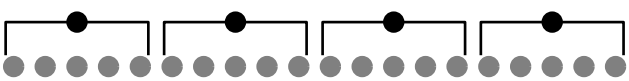
- Batching computes mean (or other summary statistic) from batch of k MCMC-samples
  - if k large enough batch statistics almost independent

$$k > mn/n_{\text{eff}}$$

  - uses data more efficiently than thinning

Slide 20

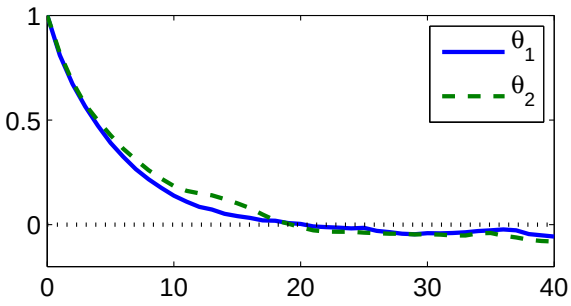
  - makes estimation of Monte Carlo error easier
    (if k estimated well)
  - batch statistics not samples from the desired distribution



Time series analysis

- Autocorrelation
  - describes how much samples correlate on average with samples with certain lag
  - also used to compare efficiencies of algorithms

Slide 21



Time series analysis

- Monte Carlo error estimated using time series analysis
- For expectation of $\bar{\theta}$

$$\mathrm{Var}[\bar{\theta}] = \frac{\sigma_\theta^2}{L/\tau}$$

where τ is summed autocorrelation

  - τ can be interpreted as how many dependent samples correspond to one independent samples
  - uses information more efficiently than thinning
  - simple to compute for only some summary statistics

Slide 22

Time series analysis

- Estimating τ

$$\tau = 1 + 2\sum_{m=1}^{\infty} \gamma(m)$$

where $\gamma(m)$ is empirical autocorrelation

  - empirical autocorrelation is noisy and thus estimate is noisy
  - longer lags have larger noise
  - noise can reduced by using truncated estimate

$$\tau = 1 + 2\sum_{m=1}^{l} \gamma(m)$$

- Since τ estimated from finite number of samples it is over-optimistic
  - if $\tau > L/20$ estimate certainly unreliable

Slide 23

Geyer's adaptive window estimators

- Markov chain properties can be used to improve estimates
  - for stationary, irreducible, recurrent
  - $\Gamma_m = \gamma_{2m} + \gamma_{2m+1}$ is positive, monotonously decreasing and convex wrt. m
- Geyer's adaptive window estimators
  - initial positive sequence estimator (Geyer's IPSE)
  - initial monotone sequence estimator (Geyer's IMSE)
  - initial convex sequence estimator (Geyer's ICSE)

Slide 24

Time series analysis

- Effective number of samples
  - comparing estimates for independent and dependent samples

$$n_{\text{eff}} \approx L/\tau$$

- Since thinning and batching discard information, k should be chosen

$$k > \tau$$

Slide 25

- Geyer's IMSE more accurate than PSRF for n_{eff}
  - accuracy of PSRF limited by the number of chains
  - Esim8_2.m, geyer_imse.m, geyer.pdf

Kolmogorov-Smirnov statistic fro convergence diagnostics

- Non-parametric approach
  - no assumption on the functional form of the distribution
  - only for independent samples (thin first)
- Compare empirical cumulative distributions

$$K(n) = \sup_x |F_{1,n}(x) - F_{2,n}(x)|$$

Slide 26

  - examine $\sqrt{n}K(n)$
  - useful threshold is 95%-quantile, i.e. 1.36
  - for multiple chains threshold by simulation
- Esim8_3.m, ksstat.m

Kolmogorov-Smirnov statistic fro convergence diagnostics

- When using several chains, threshold estimated using simulation of independent random numbers
  - KS compares two chains at time
  - in case of several chains make all pairwise comparisons and choose largest statistic value
    - this can be compared to 95% quantile of distribution obtained by simulating, e.g. 100 times equal sized independent sample sets (see Esim8_3.m)

Slide 27

Random walk

- Markov chain traverses randomly
  - i.e. random walk
  - time to get independent sample is at least the time to traverse from the one end of substantial mass to other end
  - if chain takes small steps, random walk is slower
  - stepsize is constrained by proposal distribution

Slide 28

Random walk

- Due to random walk average number of samples to get independent sample is T
  - Gibbs: $T \simeq (\sigma^{\text{marg}}/\sigma^{\text{cond}})^2$,
    where σ^{marg} is the width of the marginal distribution and σ^{cond} is the average width of the conditional distribution
    - in toy example $T \simeq 4$
    - time-series analysis says $T \sim 3.7$

Slide 29

  - Metropolis: $T \simeq (\sigma^{\text{max}}/\sigma^{\text{prop}})^2/f$,
    where σ^{max} is the largest width of the distribution, σ^{prop} is the width of the proposal distribution and f acceptance probability
    - in toy example $T \simeq 24$
    - time-series analysis says $T \sim 23$
- Above estimate is a lower limit for T
  - gives some hint on the effect of posterior dependencies

Reducing autocorrelation

- Parametrization
- Algorithms which reduce random walk

Slide 30

Gibbs sampling (ch 11.8)

- Reparametrization
  - with independent parameters efficiency of Gibbs sampling is 1
- Auxiliary variables
  - eg. t-distribution as a scale mixture of normals
- Parameter expansion

Slide 31

  - additional parameter which allows longer steps
  - makes model under-identifiable, but interesting quantities still identifiable

Metropolis algorithm (ch. 11.9)

- Reparametrization
- "Optimal" rejection rate
  - optimal scale $c \approx 2.4/\sqrt{d}$
  - efficiency $0.3/d$
  - rejection rate $0.56 - 0.77$ depending on number of dimensions

Slide 32

- Adaptivity
  - possible to use pre-adaptation
  - (continuous adaptation possible, but the it is not any more Markov chain)

Rejection sampling

- Used in
  - adaptive rejection sampling in WinBUGS
  - Ziggurat method
- Proposal distribution forms an envelope over the target distribution
- Esim11_1.m

Slide 33

Rejection sampling

- Works if proposal g is good approximation for q
  - for unidimensional log-concave and almost log-concave can be formed efficiently adaptively
  - average rejection probability describes the efficiency
- For high-dimensional difficult to choose good proposal
  - eg: q and p both normal $\sigma_q = 1.01\sigma_p$
  - if $N = 1000$ then $M \simeq 20000$
  - acceptance probability is $1/M$

Slide 34

Importance sampling

- Used in
  - sequential Monte Carlo, particle filters
  - improving analytic approximations (e.g. normal, variational, etc.)
  - adaptive methods
- Expectation of $f(\theta)$ is estimate

Slide 35

$$\mathrm{E}(f(\theta)) \equiv \frac{\sum_l w_l f(\theta^{(l)})}{\sum_l w_l}, \qquad \text{where} \quad w_l \equiv \frac{q(\theta^{(l)})}{g(\theta^{(l)})}$$

  where g is a proposal distribution
- Esim11_2.m

Importance sampling

- Estimation of reliability difficult if proposal density has low density in areas where interesting density is not low
  - variance of importance weights can be used to estimate effective number of samples
  - variance of weights can be infinite
  - also finite but very large variance, indicate problems

Slide 36

- Importance sampling with re-sampling
  - $p(\theta|y)$ approximated with discrete distribution in sampled points with probabilities relative to weights
  - samples from the desired distribution
  - typical part of eg. particle filters
- Adaptive methods using importance sampling gaining popularity
  - importance sampling does not need to care about Markov chain rules

Slice sampling

- Used in
  - part of Gibbs sampling
  - sampling all parameters using single component version
  - multidimensional version less used
- Locally adaptive

Slide 37

- Not sensitive to algorithmic parameter values
- Esim11_3.m

RJMCMC - Reversible jump MCMC

- Also known as Metropolis-Hastings-Green
- Allows jumps between different parameter spaces
  - the number of parameters can change
  - can be used to integrate over uncertainty in model structure
- Jump probability takes into account the change in the measure

Slide 38

$$r = \frac{p(y|\theta_{k^*}, M_{k^*})p(\theta_{k^*}|M_{k^*})}{p(y|\theta_k, M_k)p(\theta_k|M_k)} \frac{J_{k^*,k}J(u^*|k^*, k, \theta_{k^*})}{J_{k,k^*}J(u|k, k^*, \theta_k)} \left| \frac{\nabla g_{k,k^*}(\theta_k, u)}{\nabla(\theta_k, u)} \right|$$

Hybrid Monte Carlo, Langevin

- Uses gradient information
  - Langevin is a Metropolis-algorithm where proposal distribution is moved to direction of the gradient
  - HMC is a multi-step Langevin
  - use of gradient reduces random walk
  - use of moment reduces random walk
  - good for joint sampling

Slide 39

Simulated tempering

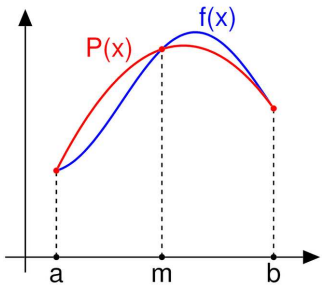
- Jumps also between different temperatures
  - similarity to simulated annealing
  - in higher temperature mode change more likely
    - higher temperature flattens peaks
  - lowest temperature corresponds to target temperature

Slide 40

Slide 41

Quadrature integration*

- In design of experiment exercise you may use
  adaptive Simpson quadrature
- Simpson methods approximates integral with second order polynomial



- Adaptive Simpson methods divides interval iteratively until desired accuracy
  - creating intervals iteratively includes some heuristics to improve computation time for common types of functions