Monte Carlo methods

Advanced Topics in Urban Informatics
NYU CUSP 2017

Summary: Stochastic methods in scientific computing

Stochastic Processes in Science Inference: with the advent of computers (1940s), simulations became a valuable alternative to analytical derivation to solve complex scientific problems, and the only way to solve non-tractable problems. Events that occur with a known probability can be simulated, the possible outcomes would be simulated with a frequency corresponding to the probability.

Applications: Instances of the evolution of a complex systems can be simulated, and from this synthetic (simulated) sample solutions can be generalized as they would from a sample observed from a population:

Urban e.g.. *simulate traffic flow to determine the average trip duration instead of measuring many trips to estimate the trip duration*,

or a better scheme would be: *simulate traffic flow and validate your simulation by comparing the average trip duration for a synthetic sample and from a sample observed from the real system, then simulate proposed changes to traffic to validate and evaluate planning options before implementing them.*

Simulations require drawing samples from distributions.

Drawing samples from a distribution can be done directly if the probability PDF *P(X)* can be integrated *analytically* to find a CDF *F(x)* and if this CDF is invertible (*$F^{-1}(u)$ can be calculated analytically)* . The algorithm is:

1. draw a *uniformly distributed* number ***u*** between [0-1]
2. invert the CDF of your distribution evaluated at *u*: ***$x=F^{-1}(u)$*** *is a sample from the desired PDF (i.e. x's are drawn at a frequency P(x))*

If *F(x)* or *$F^{-1}(u)$* cannot be calculated analytically **Rejection Sampling** allows to sample from the desired *P(x)* . The algorithm is:

1. find a function *Q(x)* that is larger than *P(x)* for every x and that has an analytical, integrable, invertible form
2. draw a sample x from *Q(x)* (see above)
3. draw a *uniformly distributed* number ***u*** between [0-Q(x)]
4. only accept x where *$u <= P(x)$*

If your proposal distribution is poorly chosen (much higher than P(x) in some regions) this can be an extremely wasteful process. The higher the problem dimensionality the more this issue becomes a concern. Alternatives include Importance sampling where the integral of the PDF is performed numerically of a sample from *Q(x)* with a correction for every *x* given by the ratio of *P(x)* to *Q(x).*

Summary: MCMC, background concepts

Markovian processes: A process is Markovian if the next state of the system is determined stochastically as a perturbation of the current state of the system, and only the current state of the system, i.e. the system has no memory of earlier states (a *memory-less* process).

A state being a stochastic perturbation of the previous state means that given the conditions of the state at time *t* (e.g. *A(t)* = (position+velocity)) the *next* set of conditions *A(t+1)* (updated position+velocity) will be drawn from a distribution related to the earlier state. For example the *next* velocity can be a sample from a Gaussian distribution with mean equal to the *current* velocity. $A(t+1) \sim \mathcal{N}(A(t), s)$

Bayes theorem: relates observed data to proposed models by allowing to calculate the *posterior of model parameters* for a given prior and observed dataset (see glossary for term definition).

$$P(\theta|D,f) = \frac{P(D|\theta,f)\,P(\theta,f)}{P(D|f)}$$

where *θ* is the set of model parameters (for example slope and intercept for a line model) *D* are the data, *f* is the functional form of a model (a line in this example) and *D* are the observed data.

Posterior(data, model-parameters) = <u>Likelihood(data, model-parameters) * Prior(model-parameters)</u>
Evidence(data)

P(*θ* |*D,f)* is the *posterior,*
P(D | θ ,f) *i*s the *likelihood* of the data given the model
P(θ,f) is the *prior,* our overall believe about the model that however must not come from the data *D* (can come from data from another experiment, from logic, from experience…)
P(D |f) is the *evidence*, and it is independent of the model parameters, so it can be ignored when the goal is comparing sets of model parameters.

Summary: MCMC

Markov Chain Monte Carlo: is a method to sample a parameter space that is based on Bayes theorem. The MCMC samples the ***joint posterior*** of the parameters in the model (up to a constant, the *evidence,* probability of observing your data under any model parameter choice, which is generally not calculable). Thus we can get posterior median, confidence intervals, covariance, etc… The algorithm is:

1. starting at some location in the parameter space propose a new location as a Markovian perturbation of the current location
2. if the proposal posterior is better than the posterior at the current location update your position (and save the new position in the chain)
3. if the proposal posterior is worse than the posterior at the current location update your position with some probability ***α***

The choice of the proposal distribution and rule ***α*** for accepting the new step in the chain have to satisfy the ***ergodic*** condition, that is: given enough time the entire parameter space would be sampled. (***Detailed Balance*** is a sufficient condition for ergodicity)
If the chain is Markovian and the proposal distribution is *ergodic <u>the entire parameter space is sampled, given enough time, with sampling frequency proportional to the posterior distribution</u>*
Different MCMC algorithms: while all MCMC algorithm share the structure above the choice of proposal and the acceptance rule are different for different MCMC algorithms.
Metropolis Hastings MCMC is the first and most common MCMC with acceptance proportional to the ratio of posteriors: ***α~posteriorNew/posteriorCurrent.*** This becomes problematic when the posterior has multiple peaks (may not explore them all) or parameter are highly covariant (may take a very long time to converge)
Convergence: It is crucial to confirm that your chains have converged and your parameter space is properly sampled, but it is also very difficult to do it. Methods include checking for stationarity of the chain means and low auto correlation in the chains. The beginning of the chain is typically removed as the chains require a minimum number of steps to move away from the initial position effectively.

Glossary

- **Stochastic**: random, following any distribution

- **PDF**: probability distribution function *P(x)* describes the *relative* likelihood of sample *x* compared

- **CDF**: cumulative distribution function - the probability that a value drawn from a distribution will be smaller than *x*

  $$F(x)=\int_{-\infty}^{x} P(x)$$

- **Marginalize**: integrate along a dimension

- **Gaussian distribution**: a distribution with PDF $N(\mu,\sigma)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-\mu)^2}{\sigma^2}}$

- **Chi Squared χ^2:** a model fitting method based on the provable fact that (under proper assumption) the function below follows a χ^2 distribution

  $$\sum_{i=1}^{N}\frac{(M-D)^2}{\sigma^2} \sim \chi^2_{DOF}$$

- **Likelihood**: in Bayes theorem its the term indicating the probability of the data under the model for a choice of parameters. More generally it can be thought of the probability of the parameters given the data

- **Posterior**: the probability of data given model calculated by Beyes theorem as likelihood * prior / evidence

- **Evidence**: the probability of the data given a model marginalized over all parameters

- **Prior**: prior, or otherwise obtained, knowledge about the problem which indicates how likeli the model parameter are for any value

- **Markovian process**: a process whose next stage depends stochastically on the current state only

- **Ergodic**: a process that given enough time would visit all location of the space

- **Chain**: an N dimensional sequence of values of each parameter of the N-dim parameter space that is explored by an MCMC