Giới thiệu MCMC

Trong các chương trước, chúng ta đã thấy quá trình mô phỏng lấy mẫu các mẫu hệ số của mô hình theo 1 phân phối xác xuất. Các mẫu hệ số này phải có số lượng đủ lớn để đảm bảo đại diện cho posterior và phân phối ổn định và chính xác.

Ví dụ với bài toán đồng xu ở chương 2, ta có 1 hệ số $$\theta$$, không gian lấy mẫu ở đây là 1 chiều. Giả sử nếu ta cần có 2 hệ số để xác định được đồng xu, muốn tính posterior, ta phải lấy mẫu 2 hệ số tổ hợp (combinations) bằng cách join 2 phân phối của 2 hệ số này. Lúc này không gian lấy mẫu của chúng ta là 2 chiều.

Giả sử chỉ với 1000 giá trị mẫu cho một hệ số, trên một mô hình với 4 hệ số, không gian mẫu 4-chiều của chúng ta đã có tổ hợp giá trị hệ số là 1000^4. Với số lượng mẫu nhiều, khối lượng tính toán sẽ tăng rất nhanh làm quá tải các hệ thống máy tính.

Thuật toán MCMC cho phép tạo các mẫu ngẫu nhiên từ 1 phân phối xác xuất theo một hàm đã biết. Thao tác này rất phù hợp với phương pháp mô phỏng bayesian đã mô tả ở các chương trước.

Giải thuật MCMC sử dụng autocorrelation, tương quan của các mẫu lấy được sử dụng định lý giới hạn trung tâm chuỗi Markov khi ước tính sai số của các giá trị trung bình.

Đánh giá MCMC

Chain tracing

Lag

ACF & Effective sample size:

* Cỡ mẫu hiệu quả (effective sample size):

Correlated observations

Giả sử ta có 1 bộ mẫu (observations) yi với phân phối mean tại $$\mu$$ và standard deviation $$\sigma$$

$$y = normal(\mu, \sigma)$$

Khi đó mean của phân phối này sẽ được tính bằng công thức:

$$\hat{\mu} = $$

Phương sai (variance) của $$\hat{\mu$$ sẽ là:

$$Var(\hat{\mu}) = \sigma^2/n $$

Điều này hoạt động tốt ở các thái cực. Nếu các mẫu quan sát thực tế của bạn là độc lập, kích thước mẫu hiệu quả của bạn bằng với kích thước mẫu thực tế. Nếu mối tương quan ở độ trễ thứ i giảm cực kỳ chậm, chậm đến mức tổng ở mẫu số phân kỳ, kích thước mẫu hiệu dụng của bạn bằng không.

Bất kỳ chuỗi Markov hợp lý nào đều nằm giữa các thái cực. Quá nhiều hy vọng về tương quan trễ bằng 0, nhưng lý tưởng là các tương quan chết đi đủ nhanh để tổng ở mẫu số không chỉ hội tụ mà còn không phải là một giá trị quá lớn.