

Introduzione
all'ottimizzazione stocastica

F. Bartolucci

Istituto di Scienze Economiche

Università di Urbino

Argomenti

- Cenni preliminari
- Stima Monte Carlo di un valore atteso, non calcolabile analiticamente, di una funzione
- Metodi Markov chain Monte Carlo
- Massimizzazione di un valore atteso, non calcolabile analiticamente, di una funzione
- Applicazione alla stima di massima verosimiglianza dei parametri di un modello statistico
- Un esempio basato sul modello autologistico

Cenni preliminari

- In molte situazioni occorre ottimizzare (minimizzare o massimizzare) una funzione del tipo

$$g(x) = E_x[h(x, Y)] = \int h(x, y) f_x(y) dy,$$

- ★ Y : variabile aleatoria (o vettore aleatorio)
 - ★ x : variabile (scalare o vettore) rispetto alla quale si vuole ottimizzare
 - ★ $f_x(y)$: funzione di probabilità (o densità) di Y che può dipendere da x
 - ★ $E(\cdot)$: valore atteso rispetto a $f_x(y)$.
- Può accadere che l'integrale usato nella definizione di $g(x)$ non sia calcolabile analiticamente.
 - Gli usuali metodi di ottimizzazione (es: Newton-Raphson) non sono quindi applicabili per l'ottimizzazione di $g(x)$.
 - In questi casi si possono utilizzare dei *ottimizzazione stocastica* che sono basati sulla stima di $g(x)$ sulla base di campioni Monte Carlo estratti da $f_x(y)$.

Ambiti di applicazione più comuni

- *Decisioni in condizioni di incertezza:*

- ★ x decisione da adottare (es: quantità di merce da produrre)
- ★ Y possibili scenari che si possono presentare (es: numero di acquirenti di una certa merce)
- ★ $h(x, y)$ funzione di utilità o di perdita (es: profitto per una certa quantità di merce venduta)

- *Stima di Massima Verosimiglianza:*

- ★ $x = \theta$ *parametri* di un modello statistico da stimare (es: modello autologistico per dati spaziali)
- ★ : Y campione che si può osservare (es: dati spaziali)
- ★ : $h(x, y)$ è definita in modo che massimizzare $g(x)$ equivale a massimizzare la verosimiglianza.

Metodi di ottimizzazione stocastica

- Tipicamente, $g(x)$ non è calcolabile quando:
 - ★ la costante di normalizzazione della distribuzione di Y , $c(x)$, non è nota

$$f_x(y) = \frac{k_x(y)}{c(x)}, \quad c(x) = \int k_x(u) du.$$

- ★ anche se $f_x(y)$ è calcolabile analiticamente, il numero delle possibili realizzazioni di Y e la complessità di $h(x, y)$ sono troppo elevate.
- Distinguiamo 4 situazioni a seconda che:
 1. la distribuzione di Y dipende o meno da x ;
 2. è possibile o no estrarre campioni Monte Carlo (con unità indipendenti) dalla distribuzione di Y .

Stima Monte Carlo di $g(x)$ (caso 1)

- Se la distribuzione di Y non dipende da x e campioni Monte Carlo possono essere estratti da $f(y)$,

$$g(x) = \int h(x, y) f(y) dy$$

può essere stimata con:

$$\hat{g}(x) = \frac{1}{n} \sum_{i=1}^n h(\theta, y^{(i)}),$$

★ $y^{(1)}, y^{(2)}, \dots, y^{(n)}$: campione casuale estratto da $f(y)$

- In generale, la stima di $g(x)$ diventa più precisa al crescere della dimensione del campione (n).
- Le derivate di $g(x)$ possono essere approssimate con quelle di $\hat{g}(x)$. Per le derivate prime e seconde si ha:

$$\frac{\partial \hat{g}(x)}{\partial x_j} = \frac{1}{n} \sum_{i=1}^n \frac{\partial g(x, y^{(i)})}{\partial x_j}$$

$$\frac{\partial^2 \hat{g}(x)}{\partial x_j \partial x_k} = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 g(x, y^{(i)})}{\partial x_j \partial x_k}$$

Ottimizzazione stocastica di $g(x)$ (caso 1)

- Anzichè ottimizzare $g(x)$ si ottimizza $\hat{g}(x)$; ciò può essere fatto iterativamente con il metodo Newton-Raphson:

1. Si sceglie opportunamente un valore iniziale di x , $x^{(0)}$
2. si estrae un campione $y^{(1)}, y^{(2)}, \dots, y^{(n)}$ da $f(y)$
3. Al passo t , si aggiorna il valore di x come:

$$x^{(t)} = x^{(t-1)} - \delta [H^{(t-1)}]^{-1} d^{(t-1)},$$

$$d^{(t)} = \left\{ \frac{\partial \hat{g}(x)}{\partial x_j} \Big|_{x^{(t)}} \right\} \text{ vettore delle derivate prime}$$

$$H^{(t)} = \left\{ \frac{\partial^2 \hat{g}(x)}{\partial x_j \partial x_k} \Big|_{x^{(t)}} \right\} \text{ matrice delle derivate seconde}$$

δ scalare che consente di variare opportunamente la lunghezza del passo.

- x viene aggiornato fino a convergenza in $\hat{g}(x)$, cioè fino a quando:

$$\left| \hat{g}(x^{(t)}) - \hat{g}(x^{(t-1)}) \right| < \varepsilon$$

- Il valore a convergenza viene indicato con \hat{x}

Esempio

- Un'impresa deve decidere la quantità di 25 prodotti da produrre nel periodo autunnale:

$$x = (x_1, x_2, \dots, x_{25})'$$

- Le quantità da produrre dipendono dalle quantità che saranno vendute che però non sono note. Queste vengono rappresentate tramite il vettore aleatorio:

$$Y = (Y_1, Y_2, \dots, Y_{25})'$$

in cui Y_i sono indipendenti e hanno distribuzione di Poisson

$$f_i(y_i) = p(Y_i = y_i) = \frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i}.$$

- Il profitto quando la quantità prodotta è x e la richiesta è y è data dalla funzione concava

$$h(x, y) = \sum_i \sum_j a_{ij}(x_i - y_i)(x_j - y_j) + \sum_i b_i(x_i - y_i) + cost$$

- Per cercare la quantità ottima da produrre si massimizza rispetto a x il profitto atteso

$$g(x) = \sum_y h(x, y) \prod_i f(y_i).$$

- Per stimare $g(x)$ si decide di utilizzare un campione di dimensione $n = 1.000$ estratto dalla distribuzione di Y .

```
Y = zeros(1000,25);
for i = 1:1000,
    Y(i,:) = poissrnd(la,1,25);
end
```

- Stima della funzione e delle sue derivate in x :

```
g = 0; d = zeros(25,1); H = A;
for i = 1:1000,
    y = Y(i,:)' ;
    g = g + ((x-y)'*A*(x-y) + b'*(x-y))/1000;
    d = d + (2*A*(x-y) + b)/1000;
end
```

- Newton-Raphson per la massimizzazione della funzione:

```
x = mean(Y)';
'stima di g, d e H
g0 = g-1;
while abs(g-g0)>10^-6,
    x = x - delta*inv(A)*d; g0 = g;
    'stima di g, d e H
end
```

Stima e ottimizzazione stocastica di $g(x)$ (caso 2)

- In alcune situazioni la distribuzione di Y non dipende da x , ma non possono essere estratti e campioni Monte Carlo da $f(y)$. Ciò tipicamente accade quando la costante di normalizzazione di $f(y)$ non è nota.
- In queste situazioni si può ancora utilizzare il metodo precedente, ma bisogna ricorrere a metodi di tipo *Markov chain Monte Carlo* (MCMC) per l'estrazione del campione da $f(y)$.
- I metodi MCMC consistono nel costruire una catena di Markov che abbia come distribuzione stazionaria $f(y)$; le unità del campione, $y^{(1)}, y^{(2)}, \dots, y^{(n)}$ sono estratte sequenzialmente secondo questa catena.
- Metodi MCMC più noti:
 - ★ *Metropolis-Hastings*;
 - ★ *Gibbs*.

Metropolis-Hastings sampler

- Nell'ambito del metodo *Metropolis-Hastings* si procede nel modo seguente:

1. Si sceglie un punto iniziale per la catena $y^{(0)}$;
2. Al passo i , l'unità $y^{(i)}$ viene scelta proponendo un nuovo valore y^* estratto da un'opportuna distribuzione $q(y|y^{(i-1)})$ chiamata proposal. Questo valore viene accettato con probabilità:

$$\alpha(y_j^*, y^{(i)}) = \frac{f(y^*)q(y^{(i-1)}|y^*)}{f(y^{(i)})q(y^*|y^{(i-1)})} = \frac{k(y^*)q(y^{(i-1)}|y^*)}{k(y^{(i)})q(y^*|y^{(i-1)})};$$

in caso di accettazione si pone $y^{(i)} = y^*$; altrimenti $y^{(i)} = y^{(i-1)}$.

- La probabilità di accettazione è calcolabile anche se non si conosce la costante di normalizzazione.
- Normalmente le prime unità estratte vengono escluse dal campione (burn-in) per garantire il raggiungimento della distribuzione stazionaria.

- Un metodo spesso utilizzato spesso per proporre dei candidati \mathbf{y}^* consiste nel cambiare un solo elemento di $\mathbf{y}^{(i-1)}$, y_j , lasciando inalterate le altre componenti. La proposal è del tipo:

$$q(y_j | \mathbf{y}^{(i-1)});$$

probabilità di accettazione

$$\alpha(\mathbf{y}_j^*, \mathbf{y}^{(i-1)}) = \frac{f(\mathbf{y}_j^* | \mathbf{y}_{-j}^{(i-1)})q(\mathbf{y}_j^{(i-1)} | \mathbf{y}^*)}{f(\mathbf{y}_j^{(i-1)} | \mathbf{y}_{-j}^{(i-1)})q(\mathbf{y}_j^* | \mathbf{y}^{(i-1)})};$$

in cui

$$f(y_j | \mathbf{y}_{-j}) = \frac{f(y_j, \mathbf{y}_{-j})}{f(\mathbf{y}_{-j})}$$

è chiamata *full conditional*.

- Nel caso particolare in cui si utilizza la full conditional come proposal si ha il *Gibbs sampler*

$$q_j(y_j | \mathbf{y}^{(i-1)}) = f(y_j | \mathbf{y}_{-j}^{(i-1)});$$

la probabilità di accettazione α è sempre pari a 1 e ogni unità proposta \mathbf{y}^* viene sempre accettata.

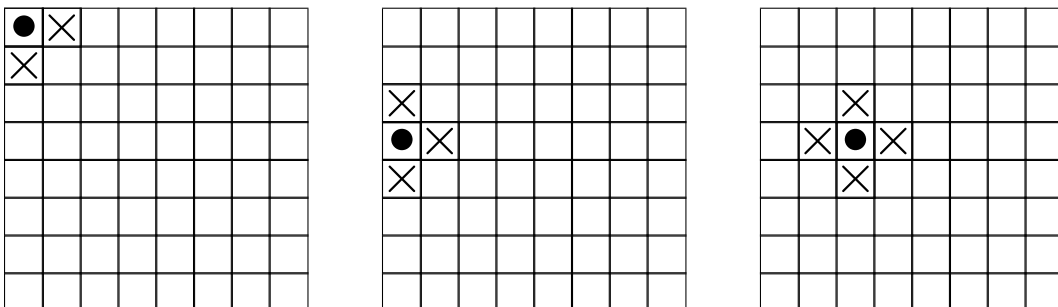
Esempio

- Il *modello autologistico* (Besag, 1974, Huffer and Wu, 1998) è spesso utilizzato per modellare di dati spaziali binari.
- Sia Y_{ij} la variabile casuale associata al sito (r, c) di una griglia regolare di dimensione $R \times C$; questo modello è basato sulle assunzioni:

1.
$$P(Y_{rc} = 1 | \mathbf{y}_{rc}^*) = \frac{\exp(\beta_0 + \mathbf{x}'_{rc}\beta_1 + \gamma\tilde{\mathbf{y}}_{rc})}{1 + \exp(\beta_0 + \mathbf{x}'_{rc}\beta_1 + \gamma\tilde{\mathbf{y}}_{rc})},$$

- ★ \mathbf{x}_{rc} vettore di covariate associate al sito (r, c)
- ★ β_0, β_1, γ parametri del modello
- ★ $\tilde{\mathbf{y}}_{rc}$ somma delle variabili nel neighbourhood.

2. Dato il neighbourhood, ogni variabile è condizionatamente indipendente da tutte le altre variabili.



- La costante di normalizzazione della distribuzione congiunta di Y non è nota; quindi bisogna usare metodi MCMC per estrarre dei campioni.
- Metropolis-Hastings estraendo una singola unità, scelta casualmente da una Bernoulliana con parametro $p = 1/2$.

```

R = 20; C = 20;
Y = zeros(R+2,C+2); Ys = zeros(R,C,10000);
for i = 1:10000,
    r = unidrnd(R)+1; c = unidrnd(C)+1;
    yp = rand<0.5;
    yn = Y(r-1,c)+Y(r,c-1)+Y(r,c+1)+Y(r+1,c);
    fc = exp(beta0+beta1*X(r,c)+gamma*yn);
    fc = fc/(1+fc);
    al = fc^yp*(1-fc)^(1-yp);
    al = al/(fc^Y(r,c)*(1-fc)^(1-Y(r,c)));
    if rand<al,
        Y(r,c) = yp;
    end
    Ys(:, :, i) = Y(2:R+1,2:C+1);
end

```

- Gibbs sampler:

```
R = 20; C = 20;
Y = zeros(R+2,C+2); Ys = zeros(R,C,10000);
for i = 1:10000,
    r = unidrnd(R)+1; c = unidrnd(C)+1;
    yn = Y(r-1,c)+Y(r,c-1)+Y(r,c+1)+Y(r+1,c);
    fc = exp(beta0+beta1*X(r,c)+gamma*yn);
    fc = fc/(1+fc);
    Y(r,c) = rand<fc;
    Ys(:, :, i) = Y(2:R+1,2:C+1);
end
```

Stima e ottimizzazione di $g(x)$ (casi 3 e 4)

- Se la distribuzione di Y dipende da x , la funzione

$$g(x) = \int h(x, y) f_x(y) dy$$

può ancora essere stimata con

$$\hat{g}(x) = \frac{1}{n} \sum_{i=1}^n h(x, y^{(i)}),$$

con $y^{(1)}, y^{(2)}, \dots, y^{(n)}$ estratto da $f_x(y)$.

- Tuttavia per la per la sua massimizzazione non si può utilizzare la tecnica Newton-Raphson del caso 1. Questa tecnica è inefficiente in quanto:
 - richiede l'estrazione di un campione a ogni passo (tecnica *many samples*);
 - normalmente le derivate di $\hat{g}(x)$ rispetto a x non approssimano adeguatamente quelle di $g(x)$;
 - è difficile verificare la convergenza dell'algoritmo dato che la stima di $g(x)$, per quanto precisa, varia ogni volta che si utilizza un nuovo campione.

Importance sampling

- Un'espressione alternativa per $g(x)$ è:

$$g(x) = E_{\bar{x}}[h(x, Y)w_{x, \bar{x}}(Y)] = \int g(x, y)w_{x, \bar{x}}(y)f_{\bar{x}}(y)dy,$$

dove $w_{x, \bar{x}}(y) = f_x(y)/f_{\bar{x}}(y)$.

- Si conseguenza, $g(x)$ può essere stimato con il metodo *importance sampling* come:

$$\hat{g}(x, \bar{x}) = \frac{1}{n} \sum_{i=1}^n g(x, y_i)w_{x, \bar{x}}(y_i),$$

dove y_1, y_2, \dots, y_n un campione estratto da $f_{\bar{x}}(y)$.

- La stima $\hat{g}(x, \bar{x})$ di $g(x)$ è normalmente tanto più precisa quanto più \bar{x} è vicino a x e quanto più è elevata la dimensione del campione (n).

- Le derivate di $g(x)$ sono approssimabili con le corrispondenti derivate di $g(x, \bar{x})$. Es. per la derivata prima:

$$\frac{\partial \widehat{g}(x)}{\partial x_j} = \frac{\partial \hat{g}(x, \bar{x})}{\partial x_j} = \frac{1}{n} \sum_{i=1}^n \frac{\partial g(x, y_i)}{\partial x_j} w_{x, \bar{x}}(y_i) + \frac{g(x, y_i)}{f_{\bar{x}}(x)} \frac{\partial f_x(y_i)}{\partial x_j}$$

Ottimizzazione di $g(x)$ (casi 3 e 4)

- Per ottimizzare $g(x)$ si procedere nel modo seguente:
 1. Si sceglie opportunamente un valore iniziale x , $x^{(0)}$
 2. si estrae un campione y_1, y_2, \dots, y_n da $f_{\bar{x}}(y)$ con $\bar{x} = x^{(0)}$;
 3. Al passo $t + 1$, si effettua un passo tipo Newton-Raphson:

$$x^{(t+1)} = x^{(t)} - \delta [H_{\bar{x}}^{(t)}]^{-1} d_{\bar{x}}^{(t)},$$

$$d_{\bar{x}}^{(t)} = \left\{ \frac{\partial \hat{g}(x, \bar{x})}{\partial x_k} \Big|_{x^{(t)}} \right\} \text{ vettore delle derivate prime}$$

$$H_{\bar{x}}^{(t)} = \left\{ \frac{\partial^2 \hat{g}(x, \bar{x})}{\partial x_h \partial x_k} \Big|_{x^{(t)}} \right\} \text{ matrice delle derivate seconde}$$

- Il vantaggio rispetto alla tecnica *many samples* è che non è necessario estrarre un nuovo campione a ogni passo.
- Se il valore di x a cui converge l'altorimo, \hat{x} , è troppo lontano da $x^{(0)}$, si ripete il procedimento partendo da $x^{(0)} = \hat{x}$.

Applicazione alla stima di massima verosimiglianza

- Il metodo di massima verosimiglianza è il più utilizzato metodo inferenziale per stimare i parametri θ di un modello statistico $f_{\theta}(y)$.
- Si supponga di aver osservato un certo valore, \bar{y} della variabile casuale Y . Il metodo in questione consiste nello stimare θ con il valore che massimizza la verosimiglianza

$$l(\theta) = f_{\theta}(y).$$

- Quando non si conosce la costante di normalizzazione di $f_{\theta}(y)$ non si possono applicare le usuali tecniche di ottimizzazione e si ricorre all'ottimizzazione stocastica.

Ottimizzazione stocastica di $l(\theta)$

- Invece che di $l(\theta)$ si può procedere alla massimizzazione stocastica del *rapporto di verosimiglianza*

$$r_{\bar{\theta}}(\theta) = \log \frac{l(\theta)}{l(\bar{\theta})} = \log \frac{k_{\theta}(\bar{y})}{k_{\bar{\theta}}(\bar{y})} - \log \frac{c(\theta)}{c(\bar{\theta})}$$

rispetto a θ , con $\bar{\theta}$ fissato opportunamente.

- Infatti, $r_{\bar{\theta}}(\theta)$ può anche essere espresso come

$$r_{\bar{\theta}}(\theta) = \log \frac{l(\theta)}{l(\bar{\theta})} = \log \frac{k_{\theta}(\bar{y})}{k_{\bar{\theta}}(\bar{y})} - \log \int \frac{k_{\theta}(v)}{k_{\bar{\theta}}(v)} f_{\bar{\theta}}(v) dv$$

e quindi può essere stimato come

$$\hat{r}_{\bar{\theta}}(\theta) = \log \frac{k_{\theta}(\bar{y})}{k_{\bar{\theta}}(\bar{y})} - \log \frac{1}{n} \sum_{i=1}^n \frac{k_{\theta}(y^{(i)})}{k_{\bar{\theta}}(y^{(i)})},$$

dove $y^{(1)}, y^{(2)}, \dots, y^{(n)}$ è un campione estratto da $f_{\bar{\theta}}(y)$.

- Le derivate di $r(\theta)$ rispetto a θ possono essere approssimate con le corrispondenti derivate di $\hat{r}_{\bar{\theta}}(\theta)$. In particolare, gli elementi dello *score* (derivata prima) sono

$$\frac{\partial \hat{r}(\theta)}{\partial \theta_j} = \frac{k'_{\theta}(y)}{k_{\theta}(y)} - \frac{\sum_{i=1}^n k_{\theta}(y^{(i)})' / k_{\bar{\theta}}(y^{(i)})}{\sum_{i=1}^n k_{\theta}(y^{(i)}) / k_{\bar{\theta}}(y^{(i)})}$$

- Per massimizzare $r_{\bar{\theta}}(\theta)$ si procedere nel modo seguente:
 1. Si parte da un valore iniziale di θ scelto opportunamente, $\theta^{(0)}$, e si estrae un campione $y^{(1)}, y^{(2)}, \dots, y^{(n)}$ da $f_{\bar{\theta}}(y)$ con $\bar{\theta} = \theta^{(0)}$;
 2. Al passo $t + 1$, si effettua un passo tipo Fisher-scoring:

$$\theta^{(t+1)} = \theta^{(t)} + \delta [F_{\bar{\theta}}^{(t)}]^{-1} s_{\bar{\theta}}^{(t)},$$

$$F_{\bar{\theta}}^{(t)} = - \left\{ \frac{\partial^2 \hat{r}_{\bar{\theta}}(\theta)}{\partial x_j \partial x_k} \Big|_{x^{(t)}} \right\} \text{ stima della matrice di informazione osservata}$$

- Il processo viene ripetuto fino a convergenza in $\hat{r}_{\bar{\theta}}(\theta)$.
- Lo stimatore di massima verosimiglianza ha distribuzione asintotica normale:

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, \hat{F}_{\bar{\theta}}^{-1}).$$

- Gli errori standard da associare alle stime possono essere calcolati come radice quadrata degli elementi sulla diagonale di $\hat{F}_{\bar{\theta}}^{-1}$.

Applicazione al modello autologistico

- La distribuzione congiunta di Y appartiene alla *famiglia esponenziale*:

$$f_{\theta}(y) = \frac{e^{t(y)' \theta}}{c(\theta)}, \quad c(\theta) = \sum_y e^{t(y)' \theta}$$

★ t vettore di statistiche sufficienti

$$t(y) = \left(\sum_r \sum_c y_{rc}, \sum_r \sum_c x_{rc} y_{rc}, \sum_r \sum_c \tilde{y}_{rc} \right)'$$

★ $\theta = (\beta_0 \quad \beta_1 \quad \gamma)'$ vettore dei parametri

- Stima del rapporto di verosimiglianza

$$\hat{r}_{\bar{\theta}}(\theta) = t(y)'(\theta - \bar{\theta}) - \log \frac{1}{n} \sum_{i=1}^n e^{(\theta - \bar{\theta})' t^{(i)}}$$

- Elementi dello score e della matrice di informazione

$$\frac{\partial \hat{r}_{\bar{\theta}}(\theta)}{\partial \theta_j} = t_j(\bar{y}) - \sum_{i=1}^n t_j^{(i)} w^{(i)}, \quad w^{(i)} = \frac{e^{(\theta - \bar{\theta})' t^{(i)}}}{\sum_h e^{(\theta - \bar{\theta})' t^{(h)}}}$$

$$-\frac{\partial^2 \hat{r}_{\bar{\theta}}(\theta)}{\partial \theta_j \partial \theta_k} = \sum_{i=1}^n t_j^{(i)} t_k^{(i)} w^{(i)} - \left(\sum_{i=1}^n t_j^{(i)} w^{(i)} \right) \left(\sum_{i=1}^n t_k^{(i)} w^{(i)} \right)$$

- Stima di $r_{\bar{\theta}}(\theta)$, di $s_{\bar{\theta}}(\theta)$ e $F_{\bar{\theta}}(\theta)$.

' estrazione dei campioni con MCMC

w = exp(T*(th-th0));

r = t'*(th-th0)-log(sum(w)/10000);

w = w/sum(w);

s = t-T'*w;

F = T'*diagv(w,T)-(T'*w)*(T'*w)';

- Newton-Raphson per la massimizzazione di $\hat{r}_{\bar{\theta}}(\theta)$:

'calcolo di r, s e F per th0

r0 = r-1;

while abs(r-r0)>10^-6,

th = th+de*inv(F)*s;

r0 = r;

'calcolo di r, s e F per th0

end

Bibliografia

- Besag, J. (1974), Spatial interactions and statistical analysis of lattice system (with discussion), *Journal of the Royal Statistical Society* **B 35**, pp. 192-236.
- Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1995), Markov Chain Monte Carlo in Practice, ser. Interdisciplinary Statistics. Boca Raton, Florida: Chapman & Hall/CRC, 1995.
- Geyer, C. J. (1992), Practical Markov chain Monte Carlo (with discussion), *Statistical Science* **7**, pp. 473-511.
- Geyer, C. J. and Thompson, E. A. (1992), Constrained Monte Carlo maximum likelihood for dependent data (with discussion), *Journal of the Royal Statistical Society, ser. B* **54**, pp. 657-699.
- Huffer, F. W. and Wu, H. (1998), Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species, *Biometrics* **54**, pp. 70-85.